Development and Application of Quantum Mechanical Tight-Binding Methods for the Exploration of Chemical Space

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

vorgelegt von
Philipp Pracht
aus Wirges
–2021–

Dekan:	Prof. Dr. Walter Witke
Erster Gutachter:	Prof. Dr. Stefan Grimme
Zweiter Gutachter:	Prof. Dr. Thomas Bredow
Tag der Disputation:	5. November 2021
Erscheinungsjahr:	2021

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

Publications

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

1. Pracht, P.; Bohle, F.; Grimme, S. "Automated Exploration of the Low-Energy Chemical Space with Fast Quantum Chemical Methods", *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

2. Pracht, P.; Grimme, S. "Calculation of Absolute Molecular Entropies and Heat Capacities Made Simple", *Chem. Sci.* **2021**, *12*, 6551–6568.

3. Pracht, P.; Grant, D. F.; Grimme, S. "Comprehensive Assessment of GFN Tight-Binding and Composite Density Functional Theory Methods for Calculating Gas-Phase Infrared Spectra", *J. Chem. Theor. Comput.* **2020**, *16*, 7044–7060.

4. Pracht, P.; Wilcken, R.; Udvarhelyi, A.; Rodde, S.; Grimme, S. "High Accuracy Quantum-Chemistry-Based Calculation and Blind Prediction of Macroscopic pKa Values in the Context of the SAMPL6 Challenge", *J. Comput.–Aided Mol. Des.* **2018**, *32*, 1139–1149.

5. Pracht, P.; Grimme, S. "Efficient Quantum-Chemical Calculations of Acid Dissociation Constants from Free Energy Relationships", *J. Phys. Chem. A* **2021**, *125*, 5681–5692.

Further publications:

1. Pracht, P.; Bauer, C. A.; Grimme, S. "Automated and Efficient Quantum Chemical Determination and Energetic Ranking of Molecular Protonation Sites", *J. Comput. Chem.* **2017**, *30*, 2618–2631.

2. Grimme, S.; Bannwarth, C.; Dohm, S.; Hansen, A.; Pisarek, J.; Pracht, P; Seibert, J.; Neese, F. "Fully Automated Quantum-Chemistry-Based Computation of Spin–Spin-Coupled Nuclear Magnetic Resonance Spectra", *Angew. Chem. Int. Ed.* **2017**, *56*, 14763–14769.

3. Pracht, P.; Caldeweyher, E.; Ehlert, S.; Grimme, S. "A Robust Non-Self-Consistent Tight-Binding Quantum Chemistry Method for Large Molecules", *ChemRxiv* **2019**, preprint, DOI: 10.26434/chemrxiv.8326202.

4. Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, S.; Spicher, S.; Grimme, S. "Extended Tight-Binding Quantum Chemistry Methods", *WIREs Comput. Mol. Sci.* **2021**, *11*, e1493.

5. Bursch, M.; Hansen, A.; Pracht, P.; Kohn, J. T.; Grimme, S. "Theoretical Study on Conformational Energies of Transition Metal Complexes", *Phys. Chem. Chem. Phys.* **2021**, *23*, 287–299.

6. 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.

7. Karunaratne, E.; Hill, D.; Pracht, P.; Gascónn, J.A.; Grimme, S.; Grant, D.F. "High-Throughput Non-targeted Chemical Structure Identification Using Gas-Phase Infrared Spectra", *Anal. Chem.* **2021**, *93*, 10688–10696.

Presentations:

1. Poster: "Automated and efficient quantum chemical determination and energetic ranking of molecular protonation sites", *53rd Symposium on Theoretical Chemistry*, **2017**, Basel, Switzerland.

2. Talk: "Recent Development and Application of the extended Tight-Binding Model", *Max-Planck Institute für Kohlenforschung*, **2018**, Mülheim, Germany.

3. Poster: "Fully Automated Quantum-Chemistry-Based Computation of Spin-Spin-Coupled Nuclear Magnetic Resonance Spectra", *58th Sanibel Symposium*, **2018**, St. Simons Island, GA, USA.

4. Poster: "Fully Automated Quantum-Chemistry-Based Computation of Spin-Spin-Coupled Nuclear Magnetic Resonance Spectra", *54th Symposium on Theoretical Chemistry*, **2018**, Halle (Saale), Germany.

5. Poster: "Fully Automated Quantum-Chemistry-Based Computation of Spin-Spin-Coupled Nuclear Magnetic Resonance Spectra", *COSMO-RS-Symposium*, **2018**, Cologne, Germany.

6. Talk: "Fully Automated Quantum-Chemistry-Based Computation of Spin-Spin-Coupled Nuclear Magnetic Resonance Spectra", *Bonn International Graduate School of Chemistry – Summer School*, **2018**, Bonn, Germany.

7. Poster: "Fully Automated Quantum-Chemistry-Based Computation of Spin-Spin-Coupled Nuclear Magnetic Resonance Spectra", *Modern Wavefunction Methods in Electronic Structure Theory Summer School*, **2018**, Gelsenkirchen, Germany.

8. Poster: "Extended Tight-Binding Methods for the Calculation of Geometries, Frequencies and Non-Covalent Interactions", *18th International Conference on Density-Functional Theory and its Applications*, **2019**, Alicante, Spain.

9. Poster: "Meta-Dynamics Based Conformational Sampling with Semiempirical Tight-Binding Methods", *258th ACS National Meeting*, **2019**, San Diego, CA, USA.

10. Talk: "Quantum Mechanical Based Exploration of Chemical Space", *Merck KGaA – Computational Chemistry Exchange*, **2021**, Bonn/Darmstadt, Germany.

Abstract

This thesis centrally focuses on the systematic exploration of the so-called "chemical space" using fast quantum chemical methods. In computational chemistry any calculation requires a detailed knowledge about a molecules spatial three-dimensional structure which defines the potential energy surface (PES). The prediction of molecular properties is therein directly linked to the populated (local) minima on the PES for a given connectivity. Because finding these minima requires a systematic sampling of the PES, the use of conventional quantum mechanical (QM) methods is often prohibitively expensive and other methodologies have to be pursued. The reason for this can easily be seen for typical molecules up to a size of roughly 200 atoms, where thousands to millions of energy evaluations are required to thoroughly explore the PES. One of the few suitable schemes for this are extended tight-binding (xTB) methods, which are derived from *ab initio* density functional theory (DFT) and introduce semiempirical approximations to accelerate calculations.

The combination of semiempirical quantum mechanical (SQM) calculations at the xTB levels with automatized sampling workflows and sophisticated sorting procedures led to the development of a computer code called CREST, to which major parts of this thesis are dedicated. As implied by the name CREST, an abbreviation for conformer-rotamer ensemble sampling tool, this program was initially introduced as a procedure to identify molecular conformers, but now adapts procedures for the screening of other representatives of the low-energy chemical space such as protonation sites and tautomers. The automated sampling procedures in CREST profit from the use of the so-called GFNn–xTB methods, purpose specific xTB schemes that are parametrized for all elements up to radon ($Z \leq 86$). It is shown that CREST and GFNn–xTB can be employed to a wide variety of chemical systems, including drug like molecules and polypeptides, organometallic compounds, transition state conformers and small molecular clusters. Due to a high robustness of the calculations and low computational times, the program is a sophisticated foundation in multilevel approaches for the calculation of molecular properties.

In an extension to the basic capabilities of CREST, a connection to a fundamental thermostatistical property, the entropy, is established. For molecules, the entropy describes a temperature dependent energy measure for the internal molecular degrees of freedom (DOF). It is commonly associated with a state of disorder and is usually obtained from partition functions for molecular motions (vibrations) in a rigid-rotor harmonic-oscillator (RRHO) approximation. In the respective chapter of this thesis, the often missing conformational dependence in typical QM calculations of the entropy is investigated, which can be obtained from partition functions for the conformer ensemble. The problem herein is that a detailed knowledge of the (full) con-

formational space is required for these contributions to the entropy and so far no generally applicable procedures existed for their calculation. A revised workflow of the CREST conformational sampling procedure is presented that provides an automated and numerically stable algorithm for the treatment of conformational entropies of flexible molecules. From thermodynamic expressions closely related to the entropy also conformational molecular heat capacities are obtained. Both quantities are benchmarked in comparison with experimental data and the computational robustness of the procedure is tested for large, flexible molecules up to roughly 100 atoms. Furthermore, the significance of the conformational terms is exemplified for some prototypical chemical reactions.

The last part of this thesis is devoted to the applications of low cost DFT, GFNn–xTB and CREST for calculation of gas-phase infrared (IR) spectra and acid dissociation constants (pK_a). Vibrational spectroscopy such as IR spectroscopy is used to characterize molecules and can identify unknown compounds when supplemented with other experiments or theoretical calculations. Theoretical IR spectra are obtained in a harmonic approximation from second derivatives of the energy and first derivatives of the molecular dipole moment with respect to nuclear positions, respectively, providing the vibrational frequencies and IR intensities. In comparison with over seven thousand experimental gas-phase IR spectra the performance of GFNn–xTB and the composite DFT method B3LYP-3c is evaluated. It is found that B3LYP-3c as a representative of DFT provides excellent, almost quantitative predictions of IR spectra. GFNn–xTB also shows reasonable accuracy and much better performance than force field or competitor SQM methods. Furthermore, an empirical correction of vibrational frequencies based on modification of atomic masses is introduced and conformational effects are studied by the use of CREST.

Acid dissociation constants are obtained from the eponymous acid dissociation reaction of molecules in solution and the associated Gibbs free energies. These energies are calculated using QM total energies from DFT or SQM, solvation free energies from implicit solvation models and free energy contributions from GFNn–xTB vibrational frequencies. By fitting empirical parameters of free energy relationship to experimental reference values a generally applicable and efficient composite protocol for the calculation of pK_a values is formulated. It is found that rather independently of the underlying DFT method errors below one pK_a unit can be achieved for flexible drug like molecules, but a strong conformational dependence is observed. CREST is herein used to identify (de-) protonation sites and to sample conformers. pK_a values calculated entirely at the GFNn–xTB level typically do not reach this accuracy and require corrections for heterolytic dissociation free energies. However, due to the low computational cost and high generalizability, they are still useful for pK_a pre-screening applications.

In summary, this thesis provides a broadly applicable framework for computational studies of conformational effects and other representatives of the low-energy chemical space. The CREST program is already being used by several computational chemistry groups, but due to sophisticated automatization and robustness of calculations is also aimed at the general chemistry community.

Zusammenfassung

Im Mittelpunkt dieser Arbeit steht die systematische Erforschung des sogenannten chemischen Raums mit schnellen quantenchemischen Methoden. Jede Berechnung in der computergestützten Chemie erfordert eine detaillierte Kenntnis der räumlichen dreidimensionalen Struktur eines Moleküls, welche die potentielle Energiehyperfläche (PES) definiert. Die Vorhersage von molekularen Eigenschaften ist darin direkt mit den populierten (lokalen) Minima auf der PES verbunden, unter Berücksichtigung einer gegebenen Konnektivität. Da das Auffinden dieser Minima eine systematische Erfassung der PES erfordert, ist die Verwendung konventioneller quantenmechanischer (QM) Methoden oft unerschwinglich und andere Vorgehensweisen müssen angestrebt werden. Der Grund dafür lässt sich leicht an typischen Molekülen bis zu einer Größe von etwa 200 Atomen erkennen, wo Tausende oder sogar Millionen von Energieberechnungen erforderlich sind, um eine vollständige Untersuchung der PES durchzuführen. Eine geeignete Möglichkeit hierfür sind erweiterte Tight-Binding (xTB) Methoden, welche aus der quantenmechanischen Dichtefunktionaltheorie (DFT) abgeleitet sind und semiempirische Näherungen zur Beschleunigung der Berechnungen einführen.

Die Kombination von semiempirischen quantenmechanischen (SQM) Berechnungen auf den xTB Levels mit automatisierten Sampling-Workflows und ausgereiften Sortierverfahren führte zur Entwicklung eines Computercodes namens CREST, welchem wesentliche Teile dieser Arbeit gewidmet sind. Wie der Name CREST, eine Abkürzung für Conformer-Rotamer-Ensemble-Sampling-Tool, bereits andeutet, ist dieses Programm zunächst als Verfahren zur Identifizierung von Molekülkonformeren eingeführt worden, adaptiert nun aber auch Prozeduren für die Untersuchung von anderen Vertretern des niederenergetischen chemischen Raums wie Protonierungsstellen und Tautomere. Die automatisierten Sampling-Verfahren in CREST profitieren von der Verwendung der sogenannten GFNn–xTB Methoden, zweckgebundenen xTB Schemata, die für alle Elemente bis hin zu Radon ($Z \leq 86$) parametrisiert sind. Es wird gezeigt, dass CREST und GFNn–xTB auf eine Vielzahl chemischer Systeme angewendet werden können, darunter wirkstoffartige Moleküle und Polypeptide, metallorganische Verbindungen, Übergangszustand-Konformere und kleine Molekülcluster. Durch eine hohe Reliabilität der Berechnungen und geringe Rechenzeiten ist das Programm eine anspruchsvolle Grundlage in mehrstufigen Ansätzen zur Berechnung molekularer Eigenschaften.

In einer Erweiterung der grundlegenden Fähigkeiten von CREST wird eine Verbindung zu einer fundamentalen thermostatistischen Eigenschaft, der Entropie, hergestellt. Für Moleküle beschreibt die Entropie ein temperaturabhängiges Energiemaß der internen molekularen Freiheitsgerade und wird üblicherweise aus Zustandsfunktionen für molekulare Bewegungen (Schwing-

ungen) in einer Starrer-Rotator und Harmonischer-Oszillator (RRHO)-Näherung gewonnen. Im entsprechenden Kapitel dieser Arbeit wird die in typischen quantenmechanischen Berechnungen fehlende Konformationsabhängigkeit der Entropie untersucht, welche aus Zustandsfunktionen für das Konformerensemble gewonnen werden kann. Das Problem hierbei ist, dass für diese fehlenden konformationellen Beiträge zur Entropie eine detaillierte Kenntnis des (vollständigen) Konformationsraums erforderlich ist und bisher keine allgemein anwendbaren Verfahren zu dessen Berechnung existieren. Ein überarbeiteter Workflow des CREST Konformations-Sampling-Verfahrens wird vorgestellt, der einen automatisierten und numerisch stabilen Algorithmus für die Behandlung von Konformationsentropien flexibler Moleküle bietet. Aus thermodynamischen Ausdrücken, die eng mit der Entropie verwandt sind, werden ebenfalls konformationelle molekulare Wärmekapazitäten erhalten. Beide Größen werden im Rahmen einer Vergleichsstudie mit experimentellen Daten verglichen und die rechnerische Robustheit des Verfahrens wird für große, flexible Moleküle bis zu etwa 100 Atomen getestet. Weiterhin wird die Bedeutung der Konformationsterme für einige prototypische chemische Reaktionen exemplarisch dargestellt.

Der letzte Teil dieser Arbeit ist den Anwendungen von kostengünstiger DFT, GFNn–xTB und CREST für Vorhersagen von Infrarotspektren (IR) in der Gasphase und Säuredissoziationskonstanten (pK_a) gewidmet. Schwingungsspektroskopie wie die IR-Spektroskopie wird zur Charakterisierung von Molekülen verwendet und kann in Kombination mit anderen Experimenten oder theoretischen Berechnungen unbekannte Verbindungen identifizieren. Theoretische Infrarotspektren werden aus zweiten Ableitungen der Energie und den ersten Ableitungen des molekularen Dipolmoments in Bezug auf die Kernpositionen gewonnen und liefern die Schwingungsfrequenzen bzw. IR-Intensitäten. Im direkten Vergleich mit über siebentausend experimentellen Gasphasen-Infrarotspektren wird die Performanz von GFNn–xTB und der Komposit-DFT-Methode B3LYP-3c bewertet. Dabei wird gezeigt, dass B3LYP-3c als Vertreter der DFT hervorragende, nahezu quantitative Vorhersagen von IR-Spektren liefert. GFNn–xTB zeigt ebenfalls eine hinreichende Genauigkeit, welche weitaus bessere Vorhersagen liefert als Kraftfeld- oder konkurrierende SQM-Methoden. Des Weiteren wird eine empirische Korrektur der Schwingungsfrequenzen basierend auf der Modifikation der Atommassen eingeführt und Konformationseffekte werden durch den Einsatz von CREST untersucht. Säuredissoziationskonstanten von Molekülen in Lösung werden aus freien Gibbs-Dissoziationsenergien gewonnen. Diese Energien werden berechnet unter Verwendung von elektronischen Gesamtenergien aus DFT oder SQM, freien Solvatationsenergien aus impliziten Solvatationsmodellen und freien Energiebeiträgen aus GFNn–xTB Frequenzberrechnungen. Durch die Anpassung empirischer Parameter der sog. freien Energiebeziehung an experimentelle Referenzwerte wird ein allgemein anwendbares und effizientes Gesamtprotokoll für die Berechnung von pK_a Werten formuliert. Es zeigt sich, dass relativ unabhängig von der zugrundeliegenden DFT-Methode Fehler geringer als eine pK_a Einheit für flexible arzneimittelähnliche Moleküle erreicht werden können, jedoch eine starke Konformationsabhängigkeit vorliegt. CREST wird hierbei zur Identifizierung von (De-)Protonierungsstellen und zur Bestimmung von Konformeren verwendet. pK_a Werte, die

ausschließlich auf dem GFNn–xTB Level berechnet werden, erreichen typischerweise nicht diese Genauigkeit und erfordern Korrekturen für heterolytische Dissoziationsenergien. Aufgrund des geringen Rechenaufwands und der hohen Zuverlässigkeit sind sie aber dennoch für pK_a Voruntersuchungen nützlich.

Zusammenfassend bietet diese Arbeit einen umfassend einsetzbaren Ansatz für computergestützte Studien von Konformationseffekten und anderen Vertretern des niederenergetischen chemischen Raums. Das CREST Programm wird bereits von mehreren Arbeitsgruppen der computergestützten Chemie verwendet, richtet sich aber aufgrund der ausgefeilten Automatisierung und Robustheit der Berechnungen auch an die allgemeine Chemikergemeinschaft.

Contents

Part I.

Introduction and Theoretical Background

1. Introduction

The continuous growth of computing capacities and advances of quantum mechanical (QM) methods have made theoretical and computational chemistry a cornerstone of modern chemical research.[1–5] Computational simulations mainly serve the purpose to interpret and verify experimental findings, but also enable the virtual study of chemical systems beyond technical or resource limitations in the laboratory. This provides the opportunity to combine both experimental and theoretical methodologies and establish iterative research processes (*cf.* Fig. 1.1). In the past, theoretical investigations have been especially beneficial for research fields of drug,[6–8] catalyst[9–11] and materials design,[12,13] as well as spectroscopy.[14–16]

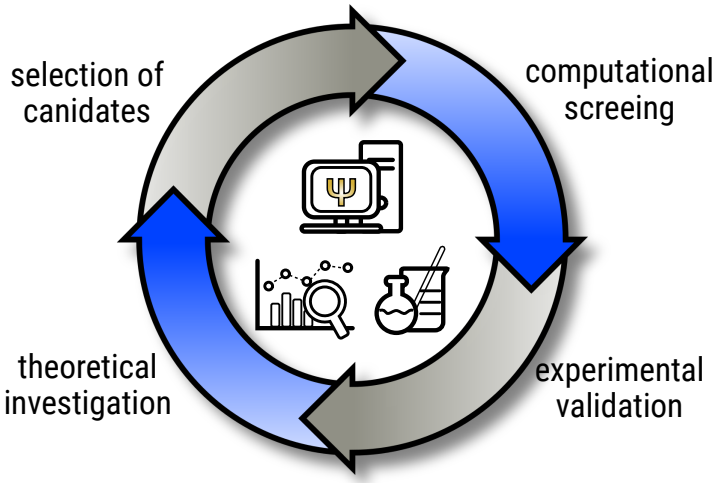

Figure 1.1.: "Back-feed" interplay of theoretical/computational chemistry and experiments in context of drug and materials design. Calculations can either be used to efficiently screen promising candidate structures or to verify and explain experimental findings.

A key aspect here is the knowledge about the spatial molecular structure, which can be obtained from measurements or computational simulations and defines chemical and physical properties.[4] For a given molecular geometry, calculations can be conducted by classical force field (FF) methods or by a variety of QM electronic structure methods, mainly belonging to either wave function theory (WFT),[17] Kohn–Sham density functional theory (KS–DFT, DFT),[18,19] or semiempirical quantum mechanics (SQM).[20,21]

Efficient algorithms and powerful processing units (CPUs, GPUs) are nowadays able to perform QM calculations on standard desktop computers.[22–25] However, Moor's law of technological advance,[26] which has been in act for over half a century, is beginning to falter and it is expected that silicon-based computing capacities reach their physical limitations by the end of the

decade.[27,28] New developments will likely focus on the underlying architecture such as quantum computing, and more efficient computational algorithms, also in the context of computational chemistry.[29–32] With respect to the latter, in an article by Houk and Liu titled "Holy Grails for Computational Organic Chemistry and Biochemistry"[32] several objectives of the research in theoretical chemistry were formulated for the next decades. One of such "holy grails" is said to be the *conquest of the combinatorial conundrum*, *i.e.*, the development of methodologies for accurately sampling the conformational space of molecules and its link to molecular properties and thermostatistical quantities, especially the entropy. A significant portion of this thesis is devoted to exactly this problem. More specifically, the CREST program (abbreviated from Conformer-Rotamer Ensemble Sampling Tool) will be introduced, which combines fast and robust SQM calculations with efficient sampling algorithms and sorting procedures for the exploration of the low-energy chemical space.[33]

There is no clear-cut definition of chemical space. In the context of chemoinformatics, the term typically refers to the entirety of all known molecules collected in databases.[34,35] However, in the scope of this thesis and theoretical chemistry in general, it is more sensible to relate the concept of chemical space to the potential energy surface (PES) of a molecule. The intuitive understanding provided by this is that the molecular chemical space consists out of all relevant low-energy structures (minima on the PES) with respect to a similar composition or topology of the molecule. From a simplified point of view, this topology refers to all PES minima associated with a single two-dimensional Lewis structure of a molecule. An appropriate labeling would hence be the above mentioned *low-energy chemical space*. Some examples for these molecules are shown in Fig. 1.2. In a broader sense, the low-energy chemical space includes closely related iso-

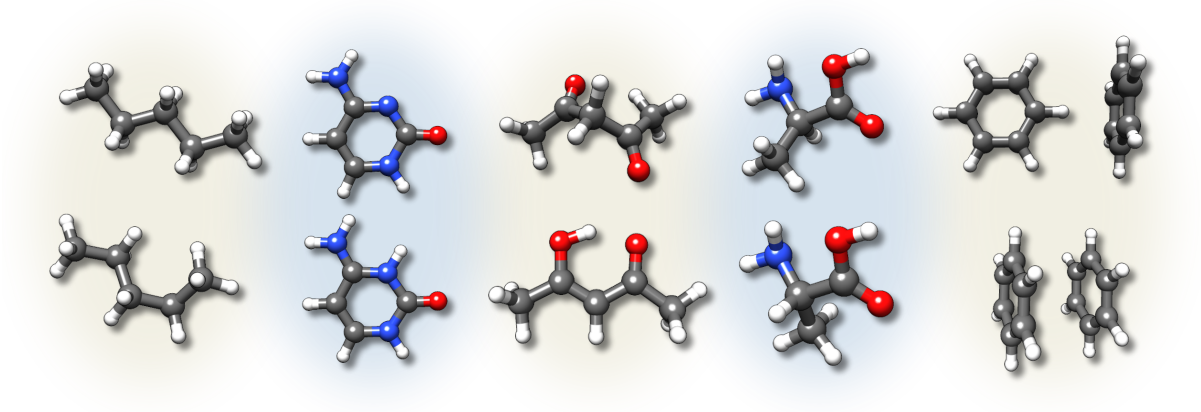

Figure 1.2.: Examples for the low-energy chemical space. From left to right: conformers of *n*-pentane, neutral and protonated cytosine, tautomers of acetylacetone, (*R*)- and (*S*)-alanine isomers, T-shaped and π–π-stacked benzene dimer.

mers and chemical "derivatives" of a molecule. Most important isomers therein are conformers, *i.e.*, spatial isomers of a molecule showing *identical* covalent connectivity and topology. Other stereoisomers also have the same connectivity but differ with regards to the orientation or chi-

rality of some atoms and are not interconvertible by rotation around covalent bonds. Chemically different isomers can be relevant if they involve only comparatively small energetical changes between the respective PES, for example upon protonation or tautomerization. A special case of the low-energy chemical space are non-covalently bound aggregates or small molecular clusters. Similar to conformers, the aggregates show the same covalent connectivity but relative orientations of molecules within the system may differ. While a major part of the combinatorial problem is the knowledge of the molecular conformation, it can be further complicated by chemical differences due to the surrounding molecular environment leading to, *e.g.*, different protonation or tautomerization states of the molecule. Since each of these states also have conformers themselves, the knowledge of low-energy chemical space is not just a conundrum but rather a combinatorial *nightmare*.

Essentially, the relevant question connecting the low-energy chemical space with theoretical predictions and the combinatorial conundrum is "What structure(s) represent the system for the calculation of a property best?" In principle all of these structures can be screened (*explored*) by computational simulations without the need of costly and time-consuming experimental studies. However, due to the enormous size of the low-energy chemical space a few requirements for the underlying theoretical methods need to be formulated:

1. The underlying method should be general, *i.e.*, most elements of the periodic table must be treatable at the same theoretical level.

2. The method should be robust for the sake of automatizing calculations whilst providing reasonable accuracy.

3. The method must be computationally cheap enough to handle hundreds to thousands of consecutive energy and gradient (derivatives of the energy) evaluations.

The problem therein is that these points often are mutually exclusive. QM treatments are general but often lack robustness and face convergence issues, *e.g.*, for transition metal containing molecules. Furthermore, typical QM methods are not nearly fast enough to handle large amounts of calculations. FFs, on the other hand, are highly robust and have low computational cost but often specialized and hence only applicable to a limited number of systems. The only practicable methods that are currently in use in computational chemistry and build a necessary compromise between these three requirements are SQM methods. Combining the robustness of parametrized methods (FFs) with the flexibility of quantum chemistry, SQM represent a "best of two worlds" approach to electronic structure computations. Sadly, many SQM methods often face similar problems of the generalizability as FFs due to inchoate parametrizations. This problem has recently been (largely) solved by introduction of the so-called GFNn–xTB methods (n= $\{0, 1, 2\}$) .[36–39] These methods are an extension to the established density functional Tight-Binding (DFTB) schemes[20] and were constructed for the specific purposes of Geometry optimizations, calculation of Frequencies and description of Non-covalent interactions (hence the acronym GFNn–xTB). Importantly, these methods are parametrized for the major part of

the periodic table (all elements up to radon, Z $\leq$ 86). Coupled with their robust and fast performance, the GFNn–xTB methods build a great opportunity for the application in automated screening procedures and exploration of the low-energy chemical space. In a more general perspective, they may be used in combination with CREST for initial stages of a computational project where thousands of structures have to be evaluated and automatization is highly beneficial. This may provide the basis for "bottom-up" screening procedures[22,40] that result in accurate molecular property predictions calculated with high level electronic structure methods (*cf.* Fig. 1.3).

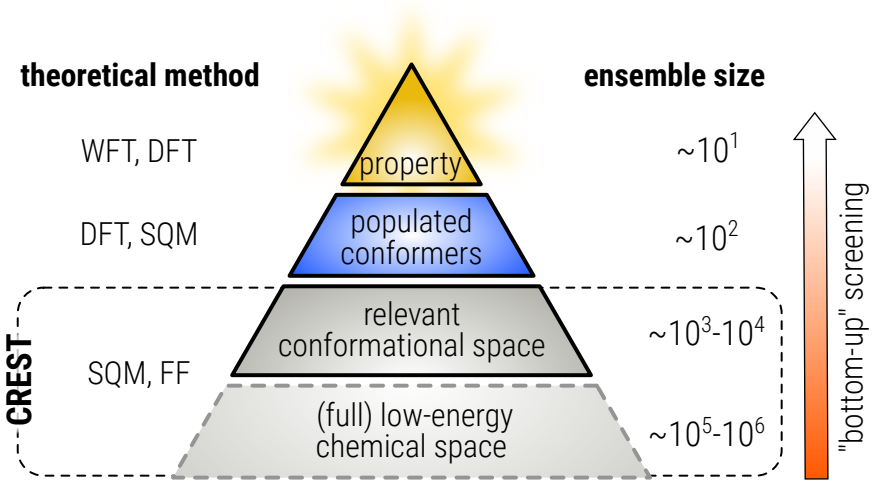

Figure 1.3.: Multilevel approach to theoretical modelling of molecular properties. The huge low-energy chemical space (conformational space) requires fast and robust methods for energy evaluations and geometry optimizations. For typical drug sized systems initial ensembles can include thousands to millions of structures which have to be screened by efficient procedures.

In the following section (Chapter 2), a more detailed overview of different methods in computational chemistry is given, with a focus on electronic structure methods. The second part of the chapter discusses how molecular properties are derived from quantum chemical calculations and how the chemical space might influence predictions of the latter. A connection is also made to free energy computations and the corresponding statistical thermodynamics.

Part II (Chapter 3) introduces the CREST program. The respective chapter reviews multiple aspects of the low-energy chemical space exploration with a focus on molecular conformations. Furthermore, the concept of atomic root-mean-square deviation (RMSD) based metadynamics simulations (MTD) is re-introduced.[41] Examples for the conformational sampling of challenging systems are shown, including macrocyclic molecules, organometallic compounds, and a large peptide. Additionally, special sampling types are discussed for conformers in the transition state, non-covalently bound aggregates and protonation sites. This part serves as the foundation for all further chapters of this thesis.

Part III is an extension to the CREST program and Chapter 3, discussing how the conformational space is linked to physical quantities, namely the molecular entropy and heat capacity.

These are in general obtained from thermostatistical expressions for the ro-vibrational frequencies of the molecule, but often neglect important anharmonic and conformational contributions due to an employed rigid-rotor harmonic-oscillator (RRHO) approximation[42] and single structures as starting points. Corresponding mathematical formulations to account for the conformational entropy are known for a long time[43] but no generally applicable computational procedure was available so far. An algorithm was developed and implemented in CREST that enables the accurate and numerically stable calculation of the conformational entropy (and heat capacity) contributions. The approach was tested in comparison with experimental absolute molecular entropies, where exceptionally low errors much below chemical accuracy were obtained. Furthermore, some prototypical applications of the CREST entropy procedure are shown to point out potential fields of usage in computational studies. This chapter provides a practical connection between the combinatorial conundrum and the thermostatistical description of drug sized molecules, which is an important development towards the above mentioned "holy grail" defined by Houk and Liu.[32]

Finally, Part IV is devoted to applications of the GFNn–xTB methods in combination with CREST and KS–DFT. Chapter 5 treats the calculation of gas-phase infrared (IR) spectra from vibrational harmonic frequency calculations. In this context, the performance of GFN1–, GFN2–xTB, GFN–FF, and the newly introduced B3LYP-3c composite functional were evaluated in comparison with more than seven thousand experimental IR spectra. Furthermore, a new atomic mass scaling approach for improvement of harmonic frequencies is introduced and the influence of conformational sampling is illustratively investigated for medium flexible systems. The goal of this project was to provide robust and fast IR spectra computations for the use in automated compound identification workflows.[44] Chapters 6 and 7 present the computation of macroscopic pK_a values in solution from free energy relationships (FER). First attempts at this topic were made in collaboration with the Novartis AG in context of the SAMPL6 blind challenge.[45,46] By combining high level KS–DFT calculations with GFN1–xTB and automated screening procedures (pre-dating CREST) for conformers, protonation sites and tautomers, the most accurate blind predictions of pK_a values in the context of the challenge were contributed. Recently, the respective computational workflow was refined and assessed for a larger number of structures. The corresponding results are presented in Chapter 7. This includes the introduction of higher order FERs for the conversion of free energies to pK_a values and an empirical heterolytic dissociation energy correction for the GFN2–xTB level. Excellent performance of KS–DFT and GFN2–xTB for pK_a calculations are observed, often with errors below one pK_a unit for typical drug molecules.

Finally, the findings and accomplishments of this thesis are summarized in Part V.

2. Theoretical Background

2.1. Methodological Overview

Theoretical chemistry allows investigations of atoms, molecules, or solids by the means of computational simulations based on physical principles. This is distinct from chemoinformatics, in which chemical systems are *not* explicitly modelled, but rather represented by collections of low-dimensional data and heuristic rules.[47] Restricting ourselves to the molecular case in context of this thesis, the computational simulations typically operate on an atomistic level. In other words, the relevant motions (mechanics) of nuclei in a Cartesian or internal coordinate space and the electrons around the nuclei are studied, which requires a set of rules or physical equations.[4] The central quantity herein is the energy of the system as a function of the nuclear coordinates. Generally, a distinction is made between classical *molecular mechanics* (MM) based on the Newtonian equations of motion and quantum mechanical (QM) methodologies as governed by the Schrödinger equation.

In case of computational chemistry the so-called Born–Oppenheimer approximation[48] has become a cornerstone for theoretical investigations. Within this approximation, the electrons are treated as quantum mechanical particles that move around the much heavier nuclei, which themselves are assumed to be moving according to classical mechanics. The respective QM methods are commonly referred to as *electronic structure methods*. Another often employed term is *ab initio* or *first principles* methods. These expressions are used somewhat inconsistently in the literature and a better distinction would be "empirical" or "non-empirical".[4,49] Methods discussed in the following are suitable for treating different chemical systems, mostly depending on the methods degree of empiricism and the size (number of atoms) of the investigated system. A schematic overview is provided in Fig. 2.1.

Methods that are conceptionally closest to exactly solving the Schrödinger equation are referred to as *wave function theory* (WFT).[17] Electrons are fully interacting in WFT, leading to archetypal many-body problems and high computational costs. While WFT methods generally provide the best accuracy,[17] their high computational cost typically limits the application to systems not much larger than a few atoms. In the simplest form of single determinant WFT, *Hartree–Fock* (HF) theory[50,51] is an *ab initio* method in which a single electron "experiences" only the average field of all other electrons. HF is therefore referred to as a *mean-field* method.[4,17] Another mean-field method (*not* derived from WFT) of comparable cost but (potentially) much better accuracy is *density functional theory* (DFT), or more specifically *Kohn–Sham* (KS) DFT.[52] While being mathematically similar to HF, DFT treats the otherwise missing

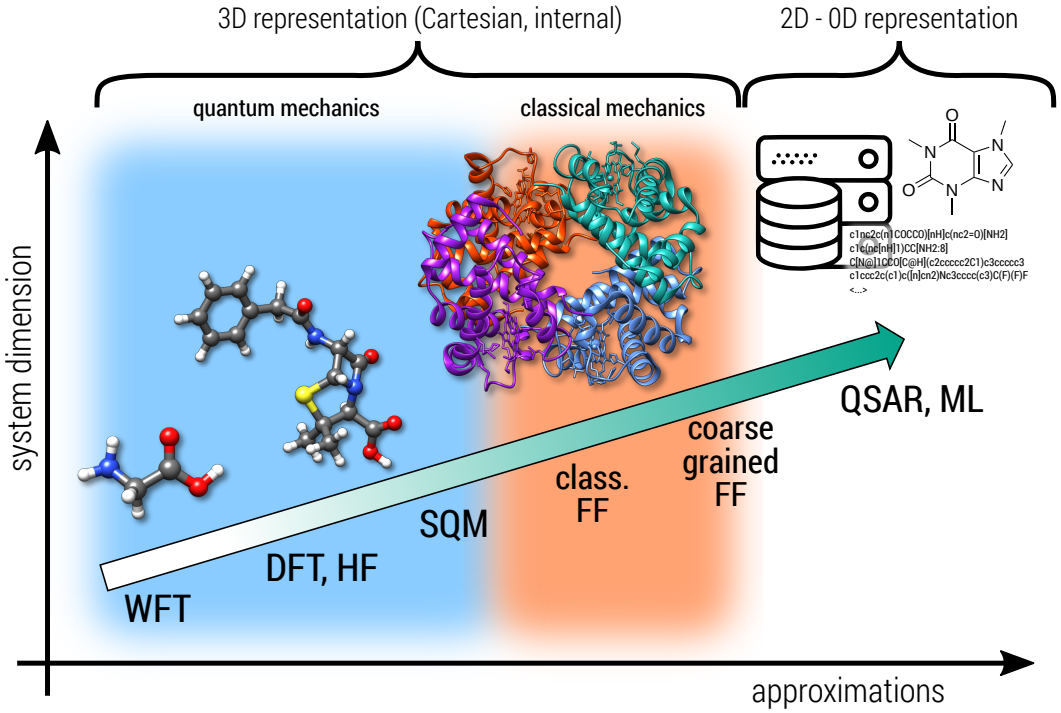

Figure 2.1.: Overview of methods in computational chemistry. The shown molecules are glycine, penicillin, hemoglobin, and caffeine.

electronic exchange-correlation (XC) by an approximated potential. KS–DFT is the *de facto* standard electronic structure method in computational chemistry and molecular physics,[4,18,19,53] with two of the ten globally most cited scientific papers stemming from this field of research[54,55] and a Nobel prize awarded to W. Kohn in 1998.[56] The mean-field HF and DFT methods show much lower computational cost than WFT and systems of a few tens to hundred atoms can be treated.

If a large number of energy evaluations is required, or larger systems such as (small) proteins shall be treated, faster methodologies are still necessary. Semiempirical quantum mechanical (SQM) methods are an empirical approximation to *ab initio* schemes. Formally, they are also electronic structure methods but introduce empirical potentials and approximations to various terms in order to significantly accelerate calculations. Starting points for the construction of SQM methods are either HF theory,[21,57] for example leading to various NDDO/MNDO[58–63] and PMx methods,[64–67] or KS–DFT which is approximated by density functional tight-binding (DFTB) theory.[20,36,68–70] Whilst being significantly faster than most QM methods, SQM methods often face a lack of generalizability due to the parametrization of the empirical approximations. Excellent reviews of SQM methods can be found in Refs. 20,21. Even more approximations than for SQM methods are required for *classical force fields* (FF).[71–74] Here, no description of QM electronic effects are necessary and atoms are modelled as classical particles. All interatomic forces in FF must be parametrized with a (large) number of parameters, often defined for pairs of elements.[75–77] Hence, only a few general FFs are available that are able to treat many differ-

ent elements in the periodic table.[78–80] Larger proteins, also including shells of explicit solvent molecules, can only efficiently be treated at FF level. Even higher empiricism, *e.g.*, by grouping several atoms into chemically related fragments and describing the fragments movement/interaction as a whole, leads to *coarse graining* FFs.[81,82] Ultimately, a connection is made here to chemoinformatic methodologies as most commonly represented by *machine learning* (ML) or *quantitative structure–activity relationship* (QSAR) models.[47] If a screening of millions of datapoints is required, the latter models are often applied. However, the objective of such methods is *not* the computation of an energy but rather directly the knowledge-based prediction of properties.[47] Often, different methods are combined in multilevel approaches[22,83] or in hybrid schemes such as QM/MM.[84,85] Herein, robust methods should be "backwards compatible", *i.e.*, low-cost methods such as FFs and SQM should provide good results also for small systems.

In Section 2.2 an overview of electronic structure methods including tight-binding based SQM is given and in Section 2.3 it is discussed how molecular properties are derived from theoretical calculations, thus providing some background knowledge required for later chapters. No further discussions will be provided for simulation techniques such as *molecular dynamics* (MD) or geometry optimization procedures since these are rather technically involved and extensive reviews can be found in Refs. 4,86–89.

2.2. Electronic Structure Methods

The subject of study in electronic structure methods are the negatively charged electrons in presence of positively charged nuclei within atoms or molecules. Omitting relativistic effects, the time-dependent Schrödinger equation

$$i\hbar\frac{\partial}{\partial t}\Psi_K(t) = \hat{H}\Psi_K(t) \tag{2.1}$$

herein describes the quantum mechanical connection between the (time-dependent) molecular wave function $\Psi_K(t)$ in the wave function state K and the Hamiltonian $\hat{H}$, *i.e.*, the operator corresponding to the total energy of the molecular system. The Hamiltonian is composed of individual operators for the kinetic energies and Coulomb interactions between the electrons e and nuclei n according to

$$\hat{H} = \hat{T}_n + \hat{T}_e + \hat{V}_{nn} + \hat{V}_{ne} + \hat{V}_{ee} \ . \tag{2.2}$$

Here, it is common practice to apply the Born–Oppenheimer approximation,[48,90] where the electrons are treated as quantum mechanical particles, while the significantly heavier nuclei are described as classical particles. Operators solely depending on the nuclei ($\hat{T}_n$, $\hat{V}_{nn}$) can be removed from Eq. 2.2 and treated separately, allowing to formulate the electronic Hamiltonian

for N electrons and M nuclei as

$$\hat{H}_e = \hat{T}_e + \hat{V}_{ne} + \hat{V}_{ee} \tag{2.3a}$$

$$= -\frac{1}{2}\sum_i^N \hat{\nabla}_i^2 - \sum_i^N \sum_A^M \frac{Z_A}{|\mathbf{r}_i - \mathbf{R}_A|} + \sum_{i>j}^N \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} \ . \tag{2.3b}$$

In Eq. 2.3, $\mathbf{r}$ and $\mathbf{R}$ are the spatial vectors of the electrons and the nuclei respectively, and Z are the nuclear charges, all defined in atomic units for simplicity.[4,90] Furthermore, since this $\hat{H}_e$ is time-independent, the time-dependent component of $\Psi_K(t)$ can be neglected for stationary states, leading to the time-independent electronic Schrödinger equation

$$\hat{H}_e \Psi_K = E_K \Psi_K \ . \tag{2.4}$$

E_K is the electronic energy corresponding to the wave function Ψ_K for the state K and denotes the main target quantity in quantum chemistry. In mathematical terms, E_K and Ψ_K are the eigenvalues and eigenvectors of the electronic Hamiltonian, respectively. They are obtained by integration of the entire space of variables λ, which in Dirac's bra and ket notation[90,91] is given by

$$\int_{-\infty}^{\infty} \Psi_K^* \hat{H} \Psi_K \mathrm{d}\lambda \equiv \left\langle \Psi_K \middle| \hat{H} \middle| \Psi_K \right\rangle \tag{2.5a}$$

$$\equiv H_{KK} = E_K \left\langle \Psi_K \mid \Psi_K \right\rangle \ . \tag{2.5b}$$

Assuming the wave functions Ψ_K to be orthonormal, the overlap integral $\langle \Psi_K | \Psi_K \rangle$ equals unity, *i.e.*,

$$\langle \Psi_K | \Psi_L \rangle = \begin{cases} 0, & \text{if } K \neq L \\ 1, & \text{otherwise} \end{cases} \tag{2.6a}$$

$$= \delta_{KL} \ , \tag{2.6b}$$

where δ_{KL} is called the Kronecker delta.

The main problem in this formalism is that no exact eigenfunctions of $\hat{H}_e$ are known *a priori* for many electron systems. However, a wide variety of methods exist in quantum chemistry to solve the time-independent electronic Schrödinger equation approximately, some of which will be discussed in the following.

2.2.1. Hartree–Fock Theory and Roothaan–Hall Equations

A possible strategy for solving the time-independent electronic Schrödinger equation is the *variational principle*

$$\tilde{E}_K = \frac{\left\langle \tilde{\Phi}_K \middle| \hat{H} \middle| \tilde{\Phi}_K \right\rangle}{\left\langle \tilde{\Phi}_K \middle| \tilde{\Phi}_K \right\rangle} \geq \left\langle \Psi_K \middle| \hat{H} \middle| \Psi_K \right\rangle = E_K \ , \tag{2.7}$$

which states that the energy $\tilde{E}_K$ for any trial wave function $\tilde{\Phi}_K$ will be higher or equal to the energy for the exact wave function Ψ_K.[90] Specifically, in Hartree–Fock (HF) theory[50,51] a sufficiently accurate trial wave function $\tilde{\Phi}_K$ can be represented by a single *Slater determinant* for the ground state, *i.e.*, for the general case of N electrons in N spin-orbitals

$$\Psi_K \approx \tilde{\Phi}_0 \equiv \Phi_0(\underline{1}, \underline{2}, \cdots, \underline{\mathrm{N}}) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(\underline{1}) & \phi_2(\underline{1}) & \cdots & \phi_N(\underline{1}) \\ \phi_1(\underline{2}) & \phi_2(\underline{2}) & \cdots & \phi_N(\underline{2}) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\underline{\mathrm{N}}) & \phi_2(\underline{\mathrm{N}}) & \cdots & \phi_N(\underline{\mathrm{N}}) \end{vmatrix} . \tag{2.8}$$

Herein, the spin-orbitals $\phi_i(\underline{k})$ denote the i^{th} one-electron wave function for the k^{th} electron. Each spin-orbital refers to a molecular orbital (MO) composed from a spatial orbital ψ_i and a spin part σ_i according to $\phi_i(\underline{k}) = \sigma_i \psi_i(\mathbf{r}_k)$, *i.e.*, they depend on the spin and spatial (Cartesian or internal) coordinates of the electron. The reason for assuming a Slater determinant rather than, *e.g.* a much simpler product of spin-orbitals (Hartree product), is that the wave function has to satisfy the Pauli principle.[90,92] Wave functions for fermionic particles such as electrons must be anti-symmetric upon permutation of two (electronic) coordinates (*i.e.*, $\Psi(\underline{1}, \underline{2}) = -\Psi(\underline{2}, \underline{1})$). The insertion of $\tilde{\Phi}_0$ into Eq. 2.4 yields the HF energy of the N electron wave function, which may be written as

$$E_{HF} = \sum_i^N \left\langle \phi_i \middle| \hat{h}_i \middle| \phi_i \right\rangle + \frac{1}{2} \sum_{i,j}^N \left(\left\langle \phi_j \middle| \hat{J}_i \middle| \phi_j \right\rangle - \left\langle \phi_j \middle| \hat{K}_i \middle| \phi_j \right\rangle \right) \tag{2.9a}$$

$$= \sum_i^N h_{ii} + \frac{1}{2} \sum_{i,j}^N \left[(ii|jj) - (ij|ji) \right] \tag{2.9b}$$

$$= \sum_i^N \epsilon_i - \frac{1}{2} \sum_{i,j}^N \left(J_{ij} - K_{ij} \right) \ . \tag{2.9c}$$

Here $\hat{h}$ is the operator containing all single-electron operators of the Hamiltonian, *i.e.*, $\hat{T}_e$ and $\hat{V}_{ne}$. The Coulomb ($\hat{J}$) and exchange ($\hat{K}$) operators are derived from the two-electron operator V_{ee} (*cf.* Eq. 2.3) leading to integrals of the type

$$(ij|kl) \equiv \iint \phi_i^*(\underline{1}) \phi_j(\underline{1}) \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} \phi_k^*(\underline{2}) \phi_l(\underline{2}) \mathrm{d}\mathbf{r}_1 \mathrm{d}\mathbf{r}_2 \ . \tag{2.10}$$

For the same electron ($i = j$) the J and K terms exactly cancel each other. If this is not the case as, *e.g.* in KS–DFT (*vide supra*), the so-called *self-interaction error* (SIE) arises.

The total energy of the system within the Born-Oppenheimer approximation is obtained by adding the nuclear repulsion energy V_{nn}, *i.e.*, $E = E_{HF} + V_{nn}$. The third line in Eq. 2.9 formulates the HF energy in terms of MO energies ϵ_i, which are derived in the following.
Since the orbitals ϕ_i are unknown initially, a set of MOs has to be determined that minimizes the energy. To do this while containing orthonormality (*cf.* Eq. 2.6), the variational orbital optimization is carried out by the means of Lagrange multipliers[90]

$$\mathcal{L} = E_{HF} \sum_{i,j}^{N} \lambda_{ij} \left(\langle \phi_i | \phi_j \rangle - \delta_{ij} \right) \tag{2.11a}$$

$$\partial\mathcal{L} = \partial E_{HF} \sum_{i,j}^{N} \lambda_{ij} \left(\langle \partial\phi_i | \phi_j \rangle - \left(\langle \phi_i | \partial\phi_j \rangle \right) \right) \ , \tag{2.11b}$$

which leads to

$$\hat{f}_i \phi_i = \sum_j \lambda_{ij} \phi_j \tag{2.12}$$

and finally upon further simplification by unitary transformation of the Lagrange multiplier matrix λ to

$$\hat{f}_i \tilde{\phi}_i = \epsilon_i \tilde{\phi}_i \ . \tag{2.13}$$

This (pseudo-)eigenvalue problem is referred to as *canonical* Hartree–Fock equations employing the special set of canonical MOs $\tilde{\phi}_i$. Herein, $\hat{f}_i$ is the so-called Fock operator

$$\hat{f}_i(\mathbf{r}_1) = \hat{h}_i(\mathbf{r}_1) + \sum_{j}^{N} \left(\hat{J}_{ij}(\mathbf{r}_1) - \hat{K}_{ij}(\mathbf{r}_1) \right) \ , \tag{2.14}$$

which acting on ϕ_i yields the energy ϵ_i of the respective MO, *i.e.*, the eigenvalue in the mean field of all other orbitals. As before, $\hat{h}_i$ is the operator describing the kinetic energy and electron-nuclei interaction of a single electron, while $\hat{J}_{ij}$ and $\hat{K}_{ij}$ are the operators for the Coulomb electron-electron repulsion and (Pauli) exchange, respectively. For $\hat{J}$ and $\hat{K}$ the summation runs over all N MOs, which is due to the approximation of using a single Slater determinant. However, contrary to the Coulomb interaction, the exchange interaction is only non-zero for electrons of the same spin. As a further consequence of this, the interaction between the J_{ii} and K_{ii} electron-electron interactions exactly cancel each other, and each electron experiences only an average contribution of all other $N - 1$ electrons. HF is therefore referred to as a *mean-field* theory.
Generally, the exact form of the MOs is unknown and can only be determined if all other orbitals are known. This requires solving the Hartree–Fock equations in a iterative, self-consistent

manner. The corresponding algorithm is referred to as *self-consistent* field (SCF) procedure,[4,90] which is not further discussed here.

A convenient procedure for defining a set of unknown MOs from a set of known functions was presented by Roothaan and Hall.[93,94] The basis of their approach is the expansion of the molecular orbitals as a linear combination of atomic orbitals (LCAO). The spatial MO ψ_i are herein expanded in a basis of M atomic orbitals (AOs) χ_μ according to

$$\psi_i(\mathbf{r}_1) = \sum_{\mu}^{M} C_{\mu i}\chi_\mu(\mathbf{r}_1) \ . \tag{2.15}$$

Inserting the expansion into Eq. 2.13 gives

$$\hat{f}_i \sum_{\mu}^{M} C_{\mu i}\chi_\mu = \epsilon_i \sum_{\mu}^{M} C_{\mu i}\chi_\mu \ , \tag{2.16}$$

which in matrix notation takes the form of the Roothaan–Hall eigenvalue equation

$$\mathbf{FC} = \mathbf{SC}\epsilon \ . \tag{2.17}$$

In this equation, the Fock matrix elements are given by $F_{\mu\nu} = \left\langle \chi_\mu | \hat{f} | \chi_\nu \right\rangle$ and the overlap elements are $S_{\mu\nu} = \langle \chi_\mu | \chi_\nu \rangle$. Since the basis of AO orbitals is fixed, the only unknown are the LCAO coefficients $\mathbf{C}$ which must be obtained from variational optimization of Eq. 2.17. Here, it is convenient to define a *density matrix* $\mathbf{P}$ with the elements

$$P_{\mu\nu} = 2\sum_{i}^{N/2} C_{\mu i}C_{\nu i} \tag{2.18}$$

and reformulate the Fock matrix elements in terms of a one-electron part $h_{\mu\nu}$ and a two-electron part $G_{\mu\nu}$ according to

$$F_{\mu\nu} = \underbrace{\left\langle \mu | \hat{h} | \nu \right\rangle}_{h_{\mu\nu}} + \underbrace{\sum_{\lambda\sigma} P_{\lambda\sigma}\left[(\mu\nu|\lambda\sigma) - \frac{1}{2}(\mu\lambda|\sigma\nu)\right]}_{G_{\mu\nu}} \ . \tag{2.19}$$

The AOs in Eq. 2.19 are referred to only by their subscript as a short notation. Since the Fock matrix depends on the expansion coefficients $\mathbf{C}$ (*via* $\mathbf{P}$) and the eigenvalue problem in Eq. 2.17 does so too, they are solved in an iterative SCF procedure.

An important point of the Roothaan–Hall formalism is that MOs and hence the total wave function is determined in a finite basis of AOs. Employing this so-called *basis set* expansion is common practice in quantum chemistry and led to the development of pre-compiled sets of AOs for each element.[4,95] Here, the quality of the final wave function is defined by the number

and functional form of the employed AOs to describe an atom in the molecule. While different functionals are employed for the AO shapes (*e.g.* Slater-type orbitals,[96] plane waves[97]), the most commonly employed basis sets for molecular calculations consist of linear combinations of Gaussian-type orbitals (GTO).[95,98,99] The decisive feature herein is the number of AOs combined to represent the physically "correct" orbital or shells of the atoms, also referred to as *cardinal number* ζ. A number of linearly combined AOs that provides the best possible description of the N-electron wave function, *i.e.*, approaching an infinite number of linear combinations, is called the complete basis set limit (CBS). Any other basis set will, to a certain degree, suffer from the so-called *basis set incompleteness error* (BSIE) and *basis set superposition error* (BSSE) as a result of the insufficient LCAO expansion.[4] Since integrals in the Roothaan–Hall formalism are evaluated in the AO basis, the cardinal number is one of the determining factors with respect to the computational cost. Triple-, or quadruple-ζ basis sets are most common in practice but occasionally minimal basis sets[98] find their use for cost efficient methods. The evaluation of the two-electron integrals in $G_{\mu\nu}$ is the most expensive part in solving SCF equations in the Roothaan–Hall formalism. Hence, the computational cost of the mean-field procedure increases as $\mathcal{O}(N^4)$, with N being the number of AOs.

2.2.2. Electron Correlation Methods

Electron correlation methods do not play a vital role in this thesis and only a brief overview is given in the following. While HF typically yields already about 99 % of a systems total energy, the last missing percent of the energy often is crucial for the final accuracy.[17] This is because the total energy is a huge quantity so that even small variations can lead to deviations exceeding the so-called *chemical accuracy*, *i.e.*, energy differences smaller then $1\,\mathrm{kcal\,mol^{-1}}$. The part of the energy not included in HF is defined from the difference between the total and mean-field energy

$$E_{corr} = E_{tot} - E_{HF} \tag{2.20}$$

and is referred to as *correlation* energy. It is important, *e.g.* for the correct description of covalent bonds and non-covalent van-der-Waals interactions.[49,100] Accordingly, methods that include (parts of) the correlation energy are called electron correlation, post-Hartree–Fock, or simply wave function theory (WFT) methods (which technically also includes HF).[17] They have in common that electrons are fully interacting instead of being described in a mean-field manner. Typically, this is described by including not only the ground state Slater determinant into the calculation, but also determinants with electrons excited to virtual orbitals (*cf.* Fig. 2.2).

The three most commonly applied electron correlation methods are configuration interaction (CI),[102,103] coupled cluster (CC),[104,105] and Rayleigh–Schrödinger perturbation theory (PT).[90] In the Møller-Plesset (MP) variant of the latter,[106] the perturbation is expanded as a power series around a zeroth order Hamiltonian $\hat{H}_0$ and the perturbation $\lambda\hat{H}'$, leading to different orders MPn, depending on where the expansion is truncated. The correlation energy appears

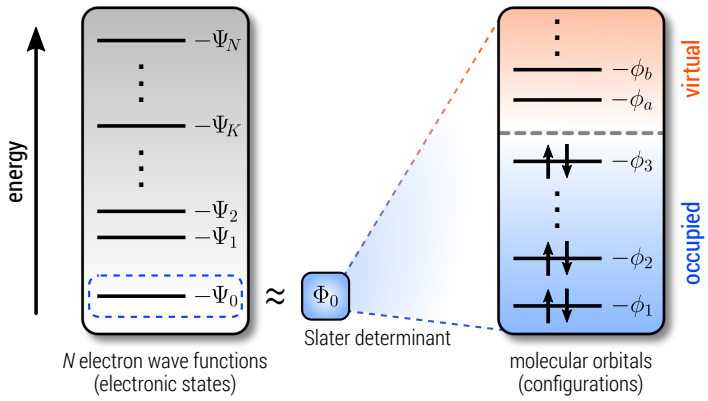

Figure 2.2.: The relation between electronic ground state wave function Ψ_0 and its approximation by a Slater determinant Φ_0. Wave functions labeled Ψ_{1-N} contain also configurations with electrons excited to virtual orbitals. Figure adapted from Ref. 101.

first at second order (MP2, zeroth plus first order MP-PT is equivalent to the HF energy),

$$E_{corr}^{MP2} = \sum_{j>i}^{occ.} \sum_{b>a}^{virt.} \frac{[(ij|ab) - (ia|jb)]^2}{\epsilon_i + \epsilon_j - \epsilon_a - \epsilon_b} \,. \tag{2.21}$$

The summation in MP2 runs over the occupied orbitals i, j as in HF, but also over the virtual orbitals a, b. This provides a significant portion of the correlation energy but introduces other methodological shortcomings (*e.g.* for metals) that are not discussed here further. Formally, the computational cost of the two-electron integrals in Eq. 2.21 is $\mathcal{O}(N^4)$ (as in HF), but since an AO to MO transformation is required the total cost of MP2 scales as $\mathcal{O}(N^5)$.[4]

The perturbation-free CI and CC approaches to solving the electronic Schrödinger equation (Eq. 2.4) are conceptionallay similar. Here the Slater determinant is expanded from a HF calculation to include electronic excitations into virtual space. This is obtained *via* an excitation operator $\hat{T}_\lambda$, with λ being the order of excitation (singles, doubles, triples, ...), applied to the ground state Slater determinant Φ_0.[4,17] The resulting CI and CC wave functions

$$\Psi_{CI} = (1 + \hat{\mathbf{T}})\Phi_0 = (1 + \sum_{\lambda}^{\infty} \hat{T}_\lambda)\Phi_0 \tag{2.22a}$$

$$\Psi_{CC} = \exp(\hat{\mathbf{T}})\Phi_0 = \sum_{k=0}^{\infty} \frac{1}{k!} \hat{\mathbf{T}}^k \Phi_0 \tag{2.22b}$$

differ in the way the excitation operator is applied to Φ_0 (linear for CI and exponential for CC) and optimized according to the excitation amplitudes. Since expansions up to N-fold excitations λ can not practically be calculated, they are truncated. One advantage of CC over CI is that higher excitations are implicitly included to a certain degree upon truncation of the CC expansion. However, the main advantage of CC is the so-called *size consistency* of the

calculations.[17] Both, CI and CC typically have significantly higher computational cost than HF. For example, the so-called "gold standard" of quantum chemistry, CCSD(T),[4,105] that is CC with single and double excitations and triple excitations from MP–PT, already scales as $\mathcal{O}(N^7)$. Modern local implementations such as DLPNO-CCSD(T) reduce the computational cost while largely maintaining the accuracy and often serve as reference level in benchmark studies.[107–109]

2.2.3. Kohn–Sham Density Functional Theory

Density functional theory originated due to a theorem by Hohenberg and Kohn[110] which postulates the energy of a system in the ground state as a functional of its electron density $\rho(\mathbf{r})$. This is possible since $\rho(\mathbf{r})$ in principle provides all information necessary to formulate a Hamiltonian.[4,53,111] As a function of electronic coordinates, the electron density can be integrated to yield the total number of electrons N in the system. Furthermore, $\rho(\mathbf{r})$ has cusps only at the positions of the nuclei and its gradient depends on the respective nuclear charge. Within this framework, an electron density dependent energy may be formulated as

$$E\left[\rho(\mathbf{r})\right] = V_{ne}\left[\rho\right] + T\left[\rho\right] + V_{ee}\left[\rho\right] \tag{2.23a}$$

$$= \int \rho(\mathbf{r}) v(\mathbf{r}) \mathrm{d}^3 r + T\left[\rho\right] + V_{ee}\left[\rho\right] \ , \tag{2.23b}$$

where $V_{ne}\left[\rho\right]$ describes the interaction between electrons and the nuclei. Additionally, all other external fields imposed on the molecule are included in this term, which is then expressed and integrated as the external potential $v(\mathbf{r})$. The terms $V_{ee}\left[\rho\right]$ and $T\left[\rho\right]$ are the electron–electron interaction energy and kinetic energy associated with a given electron density, respectively. This theory depends only on three spatial variables for $\mathbf{r}$ instead of $3N$ for the electrons as in WFT and therefore it is sometimes referred to as orbital-free density functional theory (OF–DFT).[112] Early attempts at OF–DFT by Thomas, Fermi and Dirac[113–115] actually pre-date the work by Hohenberg and Kohn and also wave mechanics,[4] but even with modern approaches the accuracy is found to be insufficient for general use in computational chemistry.[116,117] The reason for this is that the exact forms of the functionals $T\left[\rho\right]$ and $V_{ee}\left[\rho\right]$ inf Eq. 2.23 are unknown and have to be approximated.

A breakthrough was achieved by Kohn and Sham[52] who pointed out that the exact electron density of an interacting system may be replaced by a fictitious determinantal wave function of non-interacting electrons in a mean field potential. For the latter, the (approximate) density can be constructed from a set of auxiliary orbitals ψ_i for the N electrons (or N_{MO} MOs with the occupation number n_i) and is given by their summed probability densities

$$\rho\left(\mathbf{r}\right) = \sum_{i}^{N} \left|\psi_i(\mathbf{r})\right|^2 \tag{2.24a}$$

$$= \sum_{i}^{N_{MO}} n_i \int \psi_i^*(\mathbf{r}) \psi_i(\mathbf{r}) \mathrm{d}\mathbf{r} \ . \tag{2.24b}$$

Formally, the electron density is given as the sum of individual densities of electrons with α and β spin, *i.e.*, $\rho(\mathbf{r}) = \rho_\alpha(\mathbf{r}) + \rho_\beta(\mathbf{r})$, but for simplicity this is always implied in the following. In analogy to HF theory, the formulation in terms of ψ_i provides an exact kinetic energy T_S and allows to reformulate Eq. 2.23 as the Kohn–Sham density functional theory (KS–DFT, or just DFT in the following) energy expression

$$E_{KS}[\rho] = T_S[\rho] + V_{ne}[\rho] + J[\rho] + E_{XC}[\rho] \tag{2.25a}$$

$$= -\frac{1}{2}\sum_i^N \langle\psi_i|\nabla^2|\psi_i\rangle + V_{ne}[\rho] + J[\rho] + E_{XC}[\rho] \ . \tag{2.25b}$$

Here, the density dependent exchange-correlation energy functional $E_{XC}[\rho]$ is introduced, which in a general (semi-)local form[118,119] is given by

$$E_{XC}[\rho] = (T[\rho] - T_S[\rho]) + (V_{ee}[\rho] - J[\rho]) \tag{2.26a}$$

$$= E_X[\rho] + E_C[\rho] \tag{2.26b}$$

$$= \int \varepsilon_{XC}[\nabla^a\rho(\mathbf{r})]\,\rho(\mathbf{r})\,\mathrm{d}\mathbf{r} \ . \tag{2.26c}$$

The exchange-correlation functional is often further decomposed into a exchange and correlation part (*cf.* Eq. 2.26b) and depends on the energy density ε_{XC}, which is a functional of the different electron density derivatives $\nabla^a\rho(\mathbf{r})$ and is often also separated into $\varepsilon_{XC} = \varepsilon_X + \varepsilon_C$. Employing the exchange-correlation potential $v_{XC}[\rho]$, *i.e.*, the functional derivative $\partial E_{XC}[\rho]/\partial\rho$, and minimizing $E^{KS}[\rho]$ (Eq. 2.25) with respect to the orbitals ψ_i gives rise to the Kohn–Sham equations[52]

$$\hat{f}_i^{KS}[\rho]\psi_i = \epsilon_i\psi_i \ , \tag{2.27}$$

with the Kohn–Sham operator

$$\hat{f}_i^{KS}[\rho] = \hat{h}_i[\rho] + \sum_j^N \hat{J}_{ij}[\rho] + v_{XC}[\rho] \ . \tag{2.28}$$

Eq. 2.27 has an obvious similarity to the Fock operator (*cf.* Eq. 2.14) and hence can be solved employing the same methodologies as in HF (*i.e.*, the Roothaan–Hall formalism). Both, KS–DFT and HF, depend on $3N$ variables (instead of just 3 in OF–DFT), have similar computational cost and are much less complicated than, *e.g.*, advanced CI and CC models. If the exact functional $v_{XC}[\rho]$ would be known, the fictitious system of non-interacting particles provides the same density and energy as the fully interacting N electron system and $E_{XC}[\rho]$ yields the exact exchange and correlation energies.[18,52,120] Hence, the advantage of KS–DFT over HF theory, which is missing the correlation energy, becomes clear. However, since no exact exchange-correlation functional is known for an arbitrary density, the main goal in developing KS–DFT methods is to find practical approximations for $v_{XC}[\rho]$. These are either empirically

derived and/or fitted to fulfill a number of theoretical constraints,[121,122] which gave rise to an enormous amount of density functional approximations (DFA).[18,19,109,119,123] An often employed metaphor for the classification of DFAs is the "Jacob's ladder" picture as popularized by Perdew and Schmidt.[118] Here, DFAs are ranked according to their sophistication, from rather crude approximations up to the "heaven" of chemical accuracy. A brief overview of DFA rungs is given in the following.

Rungs of Density Functional Approximations

In the Jacob's ladder picture, DFAs are primarily classified according to the (local) density $\rho(\mathbf{r})$ and its respective derivatives $\nabla^a\rho(\mathbf{r})$ considered for the exchange and correlation terms in Eq. 2.26. Typically, higher rung DFAs show increasingly higher accuracy but also at higher computational cost.

The least sophisticated DFAs are the so-called local spin density approximations (LSDA or LDA). LSDA was initially derived from models for the uniform electron gas (UEG),[115,124] *i.e.*, constant density ($\nabla\rho(\mathbf{r}) = 0$) and consequently the exchange-correlation only depends on $\rho(\mathbf{r})$. This leads to a comparatively simple formulation for the exchange energy

$$E_X^{LSDA}[\rho] = -C_X \int \left(\rho_\alpha(\mathbf{r})^{\frac{4}{3}} + \rho_\beta(\mathbf{r})^{\frac{4}{3}}\right) \mathrm{d}\mathbf{r} \ , \tag{2.29}$$

where $C_X = \frac{3}{4}\left(\frac{3}{\pi}\right)^{\frac{1}{3}}$ and $\rho_{\alpha/\beta}$ are the α and β spin densities. The corresponding correlation part can be found in the literature.[4,125,126]

The second rung corresponds to functionals of the generalized gradient approximation (GGA) type, which utilize information contained in the density derivative $\nabla\rho(\mathbf{r})$. Since the latter describes variations around $\rho(\mathbf{r})$ but still depends on a single $\mathbf{r}$, GGAs are sometimes referred to as being "semi-local". In a very general form, GGAs are based on LSDA exchange and correlation, but employ an enhancement factor $F_{XC}[\rho, \nabla\rho]$, yielding

$$E_{XC}^{GGA}[\rho] = \int \varepsilon_{XC}^{LSDA}[\rho(\mathbf{r})]\, F_{XC}^{GGA}[\rho(\mathbf{r}), \nabla\rho(\mathbf{r})]\, \mathrm{d}\mathbf{r} \ . \tag{2.30}$$

Many different enhancement factors for the exchange and correlation have been proposed, with some of the most prominent DFAs being the B88 and PBE exchange[127,128] and the LYP correlation[54] functionals.

Improvements upon the GGA rung lead to the meta-generalized gradient approximation (meta-GGA) rung. Here, enhancement factors $F_{XC}^{meta-GGA}$ depend not only on $\nabla\rho$, but also higher derivatives $\nabla^2\rho$ or the kinetic energy density

$$\tau(\mathbf{r}) = \frac{1}{2}\sum_i^N |\nabla\psi_i(\mathbf{r})|^2 \ . \tag{2.31}$$

The latter is typically employed in meta-GGAs since less numerical noise is introduced into

the calculations than by employing $\nabla^2\rho$.[119] The most theoretically evolved semi-local DFAs are of the meta-GGA rung.[122,129,130] Computational costs formally scale as $\mathcal{O}(N^4)$ due to the semi-local two-electron integral evaluation, but can be accelerated to $\mathcal{O}(N^3)$ by appropriate approximations such as the *resolution of identity* (RI) approximation.[131–133]

The fourth rung of DFAs is referred to as hybrid DFT or (rarely also) hyper-GGA methods.[4] The main idea here is to find a connection between the exact (non-local) exchange known from HF theory (Fock exchange) and the exchange contained in the exchange-correlation functional $E_{XC}[\rho]$.[134] By invoking the adiabatic connection[135] and assuming a linear correspondence[120] between the Fock and (meta-)GGA exchange as the end points the empirical formula

$$E_{XC}^{hybrid}[\rho] = a_X E_X^{HF} + (1 - a_X)E_X^{DFT} + E_C^{DFT} \tag{2.32}$$

is obtained, where E_X^{HF} is the Fock exchange obtained from a HF calculation and a_X is typically treated as fitting parameter.[119,134] Naturally, employing E_X^{HF} requires an additional Fock energy evaluation which makes hybrid DFT more costly than GGA or meta-GGA calculations. However, hybrid DFAs still are mean-field approaches for which a single KS operator can be formulated, *i.e.*, the Fock part is evaluated using the KS orbitals. In other words, the Fock exchange is evaluated non-locally by integrating $\rho(\mathbf{r}, \mathbf{r}')$ over two spatial parts $\mathbf{r}$ and $\mathbf{r}'$. The respective computational performance hence still scales as $\mathcal{O}(N^4)$. All (semi-)local DFAs suffer from the above mentioned self-interaction error (SIE). The SIE describes the artificial Coulomb interaction of a single electron with itself, which should be canceled by $J[\rho]$ and the exact $E_{XC}[\rho]$. Since this is not the case for (semi-)local DFT, the improvements of hybrid functionals compared to lower rung DFAs can be at least partially attributed to the compensation of SIE due to Fock exchange. Note that the SIE is also linked to the delocalization of the electrons and a wrong asymptotic behavior of the exchange-correlation potential. This gave rise to the class of range-separated hybrid DFAs, in which the Coulomb operator for the Fock exchange is separated in a short- and long-range part.[136–141] This special case of fourth rung DFAs will not be discussed further here. Hybrid functionals, either with globally employed or range-separated Fock exchange, outperform most other (semi-)local DFAs and routinely provide better property predictions for ground and excited states.[18,109,123,134,142]

The final rung on the DFA Jacob's ladder are the so-called double-hybrid density functional (DHDF) approximations.[143–145] Similar to the exchange in hybrid DFAs, part of the correlation energy in DHDFs is replaced by an MP2-like energy to include virtual orbitals, which leads to the general expression

$$E_{XC}^{DHDF}[\rho] = a_X E_X^{HF} + (1 - a_X)E_X^{DFT} + b_C E_C^{DFT} + (1 - b_C)E_C^{MP2} \ . \tag{2.33}$$

Consequently, DFAs of the fifth rung are no mean-field methods. Many variants of DHDFs exist that differ mainly in the parameters a_X and b_C. Furthermore, some DHDFs replace the perturbative correlation energy term by other (modified) WFT or random-phase approximation (RPA) components,[146–150] although including higher order Møller-Plesset PT or CCSD(T) terms

was found to no provide any significant improvements over the MP2 term.[151] While results of DHDFs are usually superior to other DFAs this comes at the cost of a worse $\mathcal{O}(N^5)$ scaling due to the MP2 correlation part.

Empirically Corrected KS–DFT

Besides the above mentioned SIE, other theoretical or practical shortcomings are observed for KS–DFT. Some well understood errors result from the use of incomplete basis sets and consequently are also present for HF (or other WFT). If basis set with a too low count of basis functions are used, a large BSSE and BSIE can severely limit the achievable accuracy. While the BSSE can efficiently be compensated by empirical corrections such as the geometrical Counter–Poise (gCP) scheme,[152–154] the BSIE may be redeemed by the use of more complete basis sets, corrected partially *via* empirical potentials,[155,156] or reduced *via* the parameter fit in the DFA parametrization.[157,158]

A severe problem in KS–DFT are missing London dispersion[159,160] interactions, which are crucial for the description intra- and intermolecular non-covalent interactions.[49,100,161] For example, (semi-)local DFAs yield wrong asymptotic interaction energies (exponentially at short and medium range, instead of $-1/R^6$).[162] Dispersion interactions are a purely non-classical electron correlation effect that is not captured by the mean-field approach and would require the evaluation of non-local functionals.[49] For brevity this is not further discussed here, but it was generally found that dispersion interactions can be accounted for in KS–DFT by special correction schemes. In the last decade a large amount of dispersion corrections were introduced[163–169] of which the so-called DFT–D schemes[170–174] are among the most popular.[100] In the often employed DFT–D3 (or just D3) variant,[172,173] a pair-wise dispersion correction term is introduced form pre-calculated dispersion coefficient C_n^{AB} together with a DFA dependent parameter s_n and a distance dependent damping function $f_{damp}^{(n)}(R)$ ensuring the correct asymptotic behavior at short range. The resulting correction term

$$E_{disp}^{\mathrm{D3},AB} = -\sum_{AB}\sum_{n=6,8} s_n \frac{C_n^{AB}}{R_{AB}^{(n)}} f_{damp}^{(n)}(R) \tag{2.34}$$

can be added directly to the KS energy and only depends on the molecular structure and atomic coordination numbers (CN) used to calculate the $C_n^{AB}(CN_A, CN_B)$ dispersion coefficients. Different variants have been proposed for the damping function[175] (D3(0),[172] D3(BJ)[173]), where the most commonly applied Becke–Johnson (BJ) damping function[164,165] is given by

$$f_{damp,BJ}^{(n)}(R) = \frac{R^n}{R^n + (a_1\sqrt{C_8^{AB}/C_6^{AB}} + a_2)^n}\ , \tag{2.35}$$

with the two empirical parameters a_1 and a_2. Besides this two-body dispersion correction, three-body contributions are often considered for larger systems *via* an Axilrod–Teller–Muto (ATM) term[176,177] but will not be further discussed here. The recently introduced DFT-D4 model

represents the next generation of DFT-D dispersion corrections.[174,178,179] Compared to the D3 models, in DFT–D4 atomic charge dependent reference polarizabilities $\alpha(i\omega)$ and modified electronegativity dependent CNs are used to calculate the C_n^{AB} dispersion coefficients. Adding dispersion corrections improves most KS–DFT results significantly[100,109,179] and their inclusion should always be considered.[4,49]

In the last years the so-called "3c" composite methods have become popular.[157,158,180–182] These methods are derived from established DFAs (or HF) but incorporate pre-defined empirical corrections to simplify the input handling in quantum chemistry codes and speed-up calculations. To achieve the latter, one of the three eponymous corrections for all the 3c methods are tailored small (minimal) to medium sized basis sets. Other employed modifications are the use of DFT–D dispersion corrections, gCP, an empirical short-range bond (SRB) correction,[180] or re-fits of the DFAs exchange-correlation functionals. The 3c composite methods have proven to yield good geometries, accurate conformational energies[182,183] and reliable thermochemistry,[109,184–186] which makes them a robust choice for DFT based computational studies.

2.2.4. Tight-Binding Methods

Tight-binding (TB) methods are a semiempirical approximation to Hartree–Fock or KS–DFT and in the latter context also are referred to as *density functional tight-binding* (DFTB).[68–70,187] As for all SQM methods, the foundation are *first principles* energy expressions but approximations for the integrals are introduced and other simplifications, such as the use of minimal basis sets, are employed.[20,21]

The starting point for the TB energy, or more correctly for the *extended* TB energy (*vide infra)*, is a non-local correlation DFA[169] of the general form

$$E_{tot} = V_{nn} + \sum_{i}^{N_{MO}} n_i \int \psi_i^*(\mathbf{r}) \left[\hat{T}_e + \hat{V}_{ne} + v_{XC}^{LDA}[\rho] + v_C^{NL}[\rho(\mathbf{r}), \rho(\mathbf{r}')]\right] \psi_i(\mathbf{r}) \mathrm{d}\mathbf{r} \ , \tag{2.36}$$

where the different operators and the density are defined as in the previous section. The component v_C^{NL} is the operator for the non-local electron-electron Coulomb energy with the correlation kernel Φ_C^{NL} and is given by

$$v_C^{NL}[\rho(\mathbf{r}), \rho(\mathbf{r}')] = \frac{1}{2} \int \left(\frac{1}{|\mathbf{r} - \mathbf{r}'|} + \Phi_C^{NL}(\mathbf{r}, \mathbf{r}') \right) \rho(\mathbf{r}') \mathrm{d}\mathbf{r}' \ . \tag{2.37}$$

At this point in TB the density ρ is reformulated in terms of a reference density ρ_0 and density difference $\Delta\rho$, *i.e.*, $\rho = \rho_0 + \Delta\rho$. The reference density is typically constructed from spherical atomic reference densities $\rho_o = \sum_A \rho_o^A$ and Eq. 2.36 may be expressed as

$$E_{tot} = E_0^H + \delta E^H + E_{XC}^{LDA}[\rho] + E_C^{NL}[\rho, \rho'] \ . \tag{2.38}$$

Herein, the energies at reference density E_0^H and at the density fluctuations δE^H are given by

$$E_0^H = V_{nn} + \sum_i^{N_{MO}} n_{0,i} \int \psi_i^*(\mathbf{r}) \left[\hat{T}_e + \hat{V}_{ne} + \frac{1}{2} \int \frac{1}{|\mathbf{r} - \mathbf{r}'|} \rho_0(\mathbf{r}') \mathrm{d}\mathbf{r}' \right] \psi_i(\mathbf{r}) \mathrm{d}\mathbf{r} \tag{2.39}$$

$$\delta E^H = \sum_i^{N_{MO}} \Delta n_i \int \psi_i^*(\mathbf{r}) \left[\hat{T}_e + \hat{V}_0 + \frac{1}{2} \int \frac{1}{|\mathbf{r} - \mathbf{r}'|} \Delta\rho(\mathbf{r}') \mathrm{d}\mathbf{r}' \right] \psi_i(\mathbf{r}) \mathrm{d}\mathbf{r} \ , \tag{2.40}$$

with the reference potential

$$\hat{V}_0 = \sum_A^{nuclei} \left(\int \frac{1}{|\mathbf{r} - \mathbf{r}'|} \rho_0^A(\mathbf{r}') \mathrm{d}\mathbf{r}' - \frac{Z_A}{|\mathbf{r} - \mathbf{R}_A|} \right) \ . \tag{2.41}$$

Eqs. 2.38–2.41 provide the basis for deriving all the energy terms in DFTB. To do so, as the second central approximation the total energy for TB is Taylor expanded around the density variations $\Delta\rho$ according to

$$E[\rho] = \sum_{k=0}^{\infty} \frac{1}{k!} E^{(k)}[\rho_0, (\delta\rho)^k] \tag{2.42a}$$

$$= E^{(0)}[\rho_0] + E^{(1)}[\rho_0, \delta\rho] + E^{(2)}[\rho_0, (\delta\rho)^2] + E^{(3)}[\rho_0, (\delta\rho)^3] + \dots \ , \tag{2.42b}$$

which is truncated depending on the sophistication of the respective TB SQM. The most advanced TB schemes truncate at third order.[38,39,69,70] In doing so, different empirical approximations can be applied for the individual order terms. A brief overview is given in the following.

At zeroth order only E_0^H terms depending on the neutral atomic reference densities ρ_0^A remain, which may be written as

$$E^{(0)}[\rho_0] = \sum_A^{nuclei} E_A[\rho_0^A] + \frac{1}{2} \sum_{A,B}^{nuclei} \left(E_{rep}[\rho_0^A, \rho_0^B] + E_{disp}[\rho_0^A, \rho_0^B] \right) \tag{2.43a}$$

$$= \sum_A^{nuclei} E_A[\rho_0^A] + \left(E_{rep}^{(0)} + E_{disp}^{(0)} \right) \ . \tag{2.43b}$$

$E_A[\rho_0^A]$ herein denote non-interacting atomic energies that can be pre-computed for the reference density ρ_0. The atomic pair-wise terms $E_{rep}^{(0)}$ and $E_{disp}^{(0)}$ result from the exchange-correlation kernels and correspond to a pair-wise repulsion and non-local correlation (dispersion) energy, respectively. In an empirical context, the latter can be seen as equivalent to the well-known Buckingham/Lennard–Jones potentials employed to describe non-covalent interactions in classical FFs. For most DFTB methods the entire zeroth order term is expressed as an empirical repulsion potential $E^{(0)}[\rho_0] = \frac{1}{2} \sum_{A,B} V_{AB}^{rep}$, employing fitted element pair-wise parameters.

The first order term is derived from δE^H, now experiencing density fluctuations via $\Delta n_i = \delta n_i$ but no interatomic Coulomb interactions, and long-range exchange and correlation part. This

is given by

$$E^{(1)}[\rho_0, \delta\rho] = \delta E^H + \frac{\partial}{\partial\rho}\left(E_{XC}^{LDA}[\rho_0] + E_C^{NL}[\rho_0, \rho_0']\right)\delta\rho \tag{2.44a}$$

$$\approx E_{EHT}^{(1)} + E_{disp}^{(1)} \ , \tag{2.44b}$$

where the first order non-local contribution $E_{disp}^{(1)}$ contains the potential for the fluctuations $\delta\rho$ experiencing the correlation with ρ_0. The important extended Hückel theory (EHT) term[188,189] $E_{EHT}^{(1)}$ is primarily responsible for the covalent bonding in TB. It is usually formulated in terms of AOs in a minimal basis set

$$E_{EHT}^{(1)} = \sum_i^{N_{MO}} \left\langle \psi_i | \hat{H}^{EHT} | \psi_i \right\rangle = \sum_i \sum_{\mu\nu}^{N_{AO}} n_i C_{\mu i} C_{\nu i} H_{\mu\nu}^{EHT} \tag{2.45a}$$

$$= \sum_{\mu\nu} P_{\mu\nu} H_{\nu\mu}^{EHT} \ , \tag{2.45b}$$

where the density matrix elements are defined from the reference density and density fluctuations $P_{\mu\nu} = P_{\mu\nu}^0 + \delta P_{\mu\nu}$. This is the essential QM component for all TB schemes, although variations in the construction of the EHT Hamiltonian exist. In first order DFTB no self-consistent treatment is necessary.[37,190,191]

The second order energy is, as the first order energy, derived from δE^H, but now explicitly depends on the interaction of density fluctuations $\delta\rho$ between the atoms. The resulting expression

$$E^{(2)}[\rho_0, (\delta\rho)^2] = \delta E^H + \frac{\partial^2}{\partial\rho\partial\rho'}\left(E_{XC}^{LDA}[\rho_0] + E_C^{NL}[\rho_0, \rho_0']\right)\delta\rho\delta\rho' \tag{2.46a}$$

$$\approx E_{ES+XC}^{(2)} + E_{disp}^{(2)} \tag{2.46b}$$

$$\approx \frac{1}{2}\sum_{AB}^{nuclei} q_A q_B \gamma_{AB} \tag{2.46c}$$

contains the respective short-range Coulomb electrostatic and (semi-)local exchange-correlation energies, as well as second order non-local correlation. The former, as seen in the third line of Eq. 2.46, are often approximated by (some form of) damped Coulomb term that depends on atomic monopoles $q_{A/B}$, *e.g.* obtained from a Mulliken population analysis. For the function γ_{AB} in most DFTB schemes a Mataga–Nishimoto–Ohno–Klopman damping function[192–195] is employed.

The highest order term employed in the most sophisticated TB schemes, *i.e.*, the third order

energy,[36,69] depends only on the derivatives of local and non-local XC energies

$$E^{(3)}[\rho_0,(\delta\rho)^3] = \frac{\partial^3}{\partial\rho\partial\rho'\partial\rho''}\left(E_{XC}^{LDA}[\rho_0] + E_C^{NL}[\rho_0,\rho_0']\right)\delta\rho\delta\rho'\delta\rho'' \tag{2.47a}$$

$$\approx E_{XC}^{(3} + E_{disp}^{(3)} \tag{2.47b}$$

$$\approx \frac{1}{3}\sum_{AB}(q_A)^2 q_B \Gamma_{AB}\ . \tag{2.47c}$$

The empirical approximation to this term (Eq. 2.47c) depends cubically on the atomic charges and on Γ_{AB}, an empirical charge deviate term. Due to the latter, the on-site third order term can be interpreted as a charge dependent correction to the second order term that improves the description of highly charged systems.

By replacing the different order terms with empirical potentials significant savings (about 2–3 orders of magnitude) can be achieved in the computational cost of TB methods compared to KS–DFT. A major part of this computational speed-up is accounted for by evaluation of only one-electron integrals. The expensive two-electron terms are herein neglected or implicitly compensated by the empirical approximations to higher order Coulomb and XC terms. Most TB schemes are, as KS–DFT or HF, solved in a self-consistent procedure. This can be explained because the charges q enter the TB Hamiltonian in Eqs. 2.46 and 2.47 and are obtained from a population analysis, *i.e.*, the charges themselves depend on the overlap integral $S_{\mu\nu}$ and density matrix $P_{\mu\nu}$. Since the central part of the self-consistency in DFTB are the charges, it is often referred to as *self-consistent charge* (SCC), rather than SCF procedure.

While the performance of "conventional" DFTB methods is promising, a large flaw is the general availability. Many of the semiempirical terms are constructed with element pair-wise parameters that are pre-computed by first principles methods.[196–200] Parameterizations are often available only for a couple of element pairs and focus on a description of chemical interaction energies, while other important features are taken little into account. Compared to *e.g.* KS–DFT, DFTB methods lack the generalizability and are limited in their applicability to explore the chemical space for a wide range of chemical systems. This led to the development of the extended tight-binding methods that are discussed below.

Extended Tight-Binding Methods

Extended tight-binding (xTB) methods by construction employ only atomistic and a few global fitting parameters which reduces the total amount of empirical data compared to other DFTB significantly. They were introduced as part of the sTDA-xTB method[201] for computation of electronic spectra and later further developed in context of the GFNn–xTB (n= $\{0,1,2\}$) schemes, which were quickly adopted by the computational chemistry community.[36–39] Herein, the acronym GFN denotes the special purposes guiding the design of these methods, *i.e.*, they were constructed to yield good description of Geometries, (vibrational) Frequencies and Non-covalent interactions. The eponymous *extensions* mainly refer to a broad parametrization (all

elements up to radon, $Z = 86$), the employed Gaussian AO minimal basis set, empirically augmented EHT expressions and improvements of the underlying theory. An overview of the total energy terms in the three GFNn–xTB variants is given in Fig. 2.3 and will be briefly discussed below. A detailed composition of the individual energy terms is omitted here for brevity but can be found in Appendix A1.

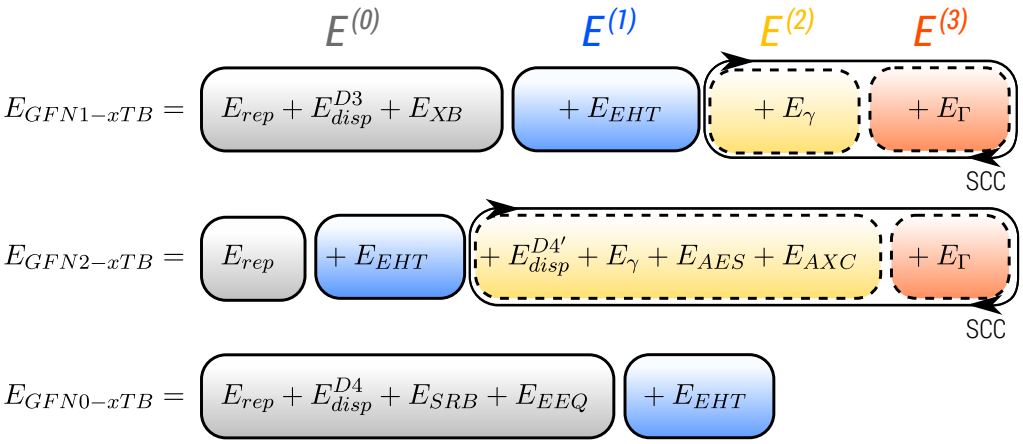

Figure 2.3.: Overview of the GFNn–xTB energy expressions. The zeroth to third order TB energy terms are color-coded for better distinction. Self-consistent parts of the energy at GFN1– and GFN2–xTB level are marked by a SCC sling. All energies additionally include a term G_{Fermi} (see Eq. 2.49), which is not shown here.

The first generation xTB method, GFN1–xTB,[36,38] is conceptionally most similar to the older DFTB3-D3 method. At zeroth order, Coulomb interactions and the Pauli repulsion are described by an empirical repulsion term E_{rep}, a correction for halogen bonds E_{XB}, and the well known D3 dispersion correction. For the first to third order terms, similar approximations as in DFTB are employed. Covalent bonds are described by the Hückel term E_{EHT}. The third order term E_Γ is approximated similar to Eq. 2.47c, but evaluated only for the diagonal elements. At second order, E_γ is described as in Eq. 2.46c but mostly in a shell-wise manner in GFNn–xTB and provides an approximation for the isotropic monopole Coulomb interactions.

Advances mainly in the second order term led to development the of GFN2–xTB method.[36,39] Here, instead of only isotropic monopole interactions, anisotropic dipole and quadrupole terms are included for the electrostatic E_{AES} and exchange-correlation E_{AXC} energies in a so-called cumulative atomic multipole moment (CAMM) scheme.[202,203] Furthermore, a self-consistent implementation of the D4 dispersion model $E_{disp}^{D4'}$ is added at second order which uses self-consistent charges of the SCC procedure. While the first and third order TB terms are mostly similar to GFN1–xTB, at zeroth order the empirical halogen bond correction E_{XC} was removed for GFN2–xTB because halogen bonds are sufficiently described by the anisotropic terms. GFN2–xTB currently is the most sophisticated available TB method and can be considered as a major improvement over previous variants.

The latest xTB variant is GFN0–xTB,[36,37] in which no second and third order terms are

present. Covalent bonds are still described at first order by an extended Hückel term. This is improved by a classical short-range bond correction E_{SRB}, the regular repulsion E_{rep} and an non-self-consistent D4 dispersion E_{disp}^{D4} energy. Furthermore, atomic charges and the electrostatic energy are not obtained from QM Mulliken populations, but from a classical charge equilibrium model (EEQ). Since this does not require an SCC procedure, GFN0–xTB requires only a single diagonalization of the Hamiltonian matrix, which makes it the fastest, but most empirical GFNn–xTB method.

Very recently also a non-electronic variant of the xTB methods, called GFN–FF, has been proposed.[80,204] It can be formally seen as a zeroth-order-only TB scheme and is classified as a *general* FF method, *i.e.*, it is available for the majority of elements in the periodic table. GFN–FF is applicable to similar problems as its xTB predecessors but not discussed here further.

Some important ingredients are common among all GFNn–xTB methods. Most prominently, all of them employ predefined minimal basis sets of the STO–mG type,[205] where several Gaussian-type atomic orbitals are used to approximate Slater-type orbitals (STO). The majority of computational savings are therein achieved by only treating valence electrons. Another prominent feature of the GFN approaches is that charges, for example in the second and third order terms, are distributed over the atomic shells l, *i.e.*, $q_A = \sum_{l \in A} q_l$, and atomic shell dependent parameters are employed. This provides more flexibility in the methods parametrization. With regards to the latter, only element specific parameters are employed, off-diagonal elements of the Hamiltonian are derived from

$$H_{\mu\nu}^{EHT} \propto S_{\mu\nu} \frac{1}{2} \left(H_{\mu\mu} + H_{\nu\nu} \right) \ . \tag{2.48}$$

The final common GFNn–xTB ingredient to be mentioned here is an electronic entropy term[206]

$$G_{\text{Fermi}} = k_B T_{el} \sum_i \sum_{\sigma=\alpha,\beta} \left[n_{i\sigma} \ln \left(n_{i\sigma} \right) + \left(1 - n_{i\sigma} \right) \ln \left(1 - n_{i\sigma} \right) \right] \tag{2.49}$$

that is used to augment the total energy. In Eq. 2.49, k_B is the Boltzmann constant, $n_{i\sigma}$ is a fractional spin-MO occupation number, and T_{el} is an electronic temperature with the default value of 300 K. The fractional occupation $n_{i\sigma}$ depends on the energy ϵ_i of the spatial MO ψ_i and the Fermi level ϵ^σ within the respective spin orbital space ($\sigma \in \{\alpha, \beta\}$), according to

$$n_{i\sigma} = \left(\exp \left[(\epsilon_i - \epsilon_F^\sigma) / k_B T_{el} \right] + 1 \right)^{-1} \ . \tag{2.50}$$

While the GFNn–xTB variants are formally spin-restricted, the fractional occupation can be used to mimic open-shell systems and static correlation. Generally, G_{Fermi} serves as an enhancement for the SCC convergence which can be modified (*via* T_{el} as an adjustable parameter) for critical cases.

The leading computational cost component is the diagonalization of the Hamiltonian (a real symmetric matrix), hence resulting in a formal $\mathcal{O}(N^3)$ scaling of the xTB methods. Cost pre-

factors differ slightly between the GFN schemes, with GFN0–xTB being roughly two to twenty times faster than GFN1– and GFN2–xTB due to the missing SCC procedure.[37] All GFNn–xTB methods provide good robustness across the periodic table and due to their parametrization are (almost) as generally applicable as KS–DFT. This makes them suitable for the quantum mechanical exploration of chemical space.

2.3. Calculation of Molecular Properties

While the different quantum chemical methods give insight about the electronic structure on a very fundamental and theoretical level, the extended objective in computational chemistry is the calculation and prediction of molecular properties. Luckily, any observable property can be derived from the wave function and its total energy. For intelligibility it is reasonable to think of most molecular properties either as derivatives of the total energy or eigenvalues of a hermitian operator for that property. For example, some properties derived from the energy are harmonic vibrational frequencies

$$\nu \propto \frac{\partial^2 E}{\partial \mathbf{R}^2} = \frac{\partial^2 \langle \Psi | \hat{H} | \Psi \rangle}{\partial \mathbf{R}^2} \ , \tag{2.51}$$

where $\mathbf{R}$ may be the nuclear coordinates, infrared and Raman intensities, or nuclear magnetic resonance (NMR) parameters. The importance of energy (and wave function) derivatives for various kinds of spectroscopy[16] is obvious already from these examples.

In essence, what is interesting about the electronic structure with regards to observable properties is how it reacts to changes of external (*i.e.*, measurement) conditions or *perturbations*. More specifically, molecular properties are usually calculated from derivatives of the energy *via* perturbation theory, or, for higher order properties, from the so-called propagator and response methods.[4,207] In the perturbation ansatz, the Hamiltonian is expanded around a unperturbed Hamiltonian employing different order perturbation operators $\hat{P}_n$ and the perturbation strength λ, *e.g.* up to second order as

$$\hat{H}_\lambda = \hat{H}_0 + \lambda \hat{P}_1 + \lambda^2 \hat{P}_2 \ . \tag{2.52}$$

For molecular properties X_λ obtained from first order perturbations, a theoretical formulation was provided by Güttinger,[208] Hellmann[209] and Feynmann[210] who theorized that

$$X_\lambda = \frac{\partial E_\lambda}{\partial \lambda} = \frac{\partial}{\partial \lambda} \langle \Psi(\lambda) | \hat{H}_\lambda | \Psi(\lambda) \rangle \tag{2.53a}$$

$$= \langle \Psi(\lambda) | \frac{\partial \hat{H}_\lambda}{\partial \lambda} | \Psi(\lambda) \rangle \ . \tag{2.53b}$$

In this aptly named Hellmann–Feynman theorem, a property X_λ with regards to some continuous parameter λ is the derivative of the eigenvalue E_λ. The latter is obtained as expectation value of the (perturbed) Hamiltonian $\hat{H}_\lambda$ and a wave function $\Psi(\lambda)$. The theorem states that for

molecular properties derived from first order perturbations only the derivative of the Hamiltonian with regards to λ is required. Importantly, it is valid even for variationally optimized wave functions (*cf.* Sec. 2.2.1) and often employed to calculate simple first order properties such as the dipole moment. Higher order properties usually require also the wave function response,[17,207] *i.e.*, the derivatives $\frac{\partial \Psi}{\partial \lambda}$. However, this is not discussed here further for brevity.

The above discussion relates mostly to molecular properties of a single molecule as governed by the Born–Oppenheimer approximation. Herein quantum effects of the nuclei and their movement are neglected. If nuclei were to behave as quantum mechanical particles, they would *tunnel* to the global minimum geometry[4] and the expectation value of this hypothetical wave function would provide the property accordingly. In fact, this would prove the simultaneous existence of several minima on the PES that contribute to the observable. However, the only significant borderline case for quantum mechanically-behaving nuclei is the hydrogen atom.[4,211,212] For all heavier atoms, the Born–Oppenheimer approximation is a central constituent and nuclei will in good approximation behave as classical particles. Hence, the above mentioned coexisting PES minima need to be identified to obtain a prediction of the property. At the timescale of experimental measurements and finite temperature, the respective observable will be an average $\langle X \rangle$ over the occurring nuclear movements. Assuming the time period of the measurement to be (infinitely) large compared to the time required for movement of nuclei, the averaged property is obtained from

$$\langle X \rangle = \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^\tau X(t) \mathrm{d}t \ . \tag{2.54}$$

This is impracticable because no such long simulation of a physical system can be conducted. A more practical solution is obtained by invoking the so-called *ergodic hypothesis*, which allows the replacement of a large time average with an average over a representative collection of M microstates, characterized by their energies E_i with

$$\langle X \rangle \approx \langle X \rangle_M = \frac{1}{M} \sum_1^M X(E_i) \ . \tag{2.55}$$

Herein, "representative" means that the sample size M is large enough to be proportional to a probability distribution at infinite timescale.[4,213] In a discrete form, the probability averaged property $\langle X \rangle$ is hence expressed as

$$\langle X \rangle = \sum_i^M p_i(E_i) \, X_i \ , \tag{2.56}$$

where the Boltzmann population p_i is introduced. Essentially, p_i is the probability of "finding"

a molecule in the state E_i at a given temperature, which is provided by

$$p_i = \frac{g_i e^{-E_i\beta}}{\sum g_i e^{-E_i\beta}} \ , \tag{2.57}$$

where $\beta = \frac{1}{kT}$, with the temperature T and k being the Boltzmann constant. This equation is of central importance and will appear several times in this thesis (Eqs. 3.2, 4.9, 6.1). Note that here the "level" formulation of the Boltzmann population is used, *i.e.*, the state E_i can be degenerate by a factor of g_i. The denominator in Eq. 2.57 is also referred to as the partition function

$$Q = \sum_i^M g_i e^{-E_i\beta} \ , \tag{2.58}$$

which is the central quantity in statistical mechanics[213,214] and acts as an normalization factor for the populations. A goal at this point is to relate a QM calculation of an isolated molecule to macroscopic observables of samples containing 10^{20} or more particles. The framework of this is provided by statistical thermodynamics as discussed in the following. A central aspect herein is to represent the single molecule as an conformational ensemble, *i.e.*, the summation in Eqs. 2.56–2.58 is made over M representative conformers of the molecule. Definitions and further discussion of conformers are provided in Part II.

2.3.1. Statistical Thermodynamics and Free Energies

Statistical thermodynamics provides a connection between the partition function Q for an ensemble of particles and external macroscopic parameters such as the temperature T and the systems volume V. This allows the construction of the three fundamental quantities Gibbs free energy G, enthalpy H, and entropy S.

$$G = H - TS = kTV\left(\frac{\partial \ln Q}{\partial V}\right)_T - kT\ln Q \tag{2.59}$$

$$H = kT^2\left(\frac{\partial \ln Q}{\partial T}\right)_V + kTV\left(\frac{\partial \ln Q}{\partial V}\right)_T \tag{2.60}$$

$$S = kT\left(\frac{\partial \ln Q}{\partial T}\right)_V + k\ln Q \tag{2.61}$$

For any meaningful computational study at finite temperature, these are the desired quantities, and not for example the total energy calculated from solving the electronic Schrödinger equation.

By again referring to the ergodic hypothesis, the partition function of many particles can be replaced by the partition function of a single isolated molecule, given that *all* possible energetic states of the molecule are known.[213–215] In other words, all electronic, translational, rotational, vibrational, and conformational (as a result of discretization in the harmonic approximation) degrees of freedom (DOF) must be known, which leads of a product of the individual partition

functions

$$Q_{tot} = Q_{elec}\, Q_{trans}\, Q_{rot}\, Q_{vib}\, Q_{conf} \ . \tag{2.62}$$

Since the enthalpy and entropy contributions depend on the logarithm of Q, the respective quantities be separated according to their DOFs

$$S_{tot} = S_{elec} + S_{trans} + S_{rot} + S_{vib} + S_{conf} \tag{2.63a}$$

$$H_{tot} = H_{elec} + H_{trans} + H_{rot} + H_{vib} + H_{conf} \ . \tag{2.63b}$$

This separation allows the efficient treatment of each individual contribution and approximations are introduced. For example, in case of the electronic partition function often just the ground state is considered,[4] with all other contributions (rot., vib., conf.) computed at this surface, and the vibrational contributions are calculated from harmonic frequencies (*cf.* Eq. 2.51). Conformational terms arise from the vibrational contributions due to the harmonic approximation and are important for flexible molecules. However, in practice the separation of electronic, vibrational and conformational terms can be problematic due to breakdown of the Born–Oppenheimer approximation. With a focus on the conformational contribution Part III is devoted to this topic. There it is shown that the discrete total energies of (optimized) molecular conformations may be treated as energetic states similar to the electronic partition function. For brevity, the individual terms typically employed for the enthalpy and entropy are not shown here but can be found in Appendix A1.

The free energy for a single molecule describes its energy content at finite temperature, *i.e.*, it contains thermodynamic information at standard conditions. Following an established thermostatistical protocol,[22,42,216]

$$G = E_{tot} + G^T_{trv} + \delta G^T_{solv} \ , \tag{2.64}$$

where E_{tot} is the total energy for the molecule obtained from QM, SQM, or FF calculation using the Born–Oppenheimer approximation, G^T_{trv} is the thermodynamic contribution from translational, rotational and vibrational DOFs and δG^T_{solv} are contributions to the solvation free energy. G^T_{trv} is calculated from the respective translational, rotational and vibrational entropy and enthalpy contributions, most commonly obtained in a *rigid-rotor harmonic-oscillator* (RRHO) approximation. Furthermore, G^T_{trv} includes a correction for the so-called *zero-point vibrational energy* (ZPVE), that is the residual energy arising from quantum mechanical motions even at zero temperature. In good approximation, the ZPVE correction is obtained by a sum over the vibrational (RRHO) modes ν_i of the molecule

$$\Delta\text{ZPVE} = \frac{1}{2} \sum_{i}^{modes} \nu_i \ . \tag{2.65}$$

The contributions to the solvation free energy δG^T_{solv} are only required if the molecule is not modelled in the gas-phase, but in solution. Typically, this term comes from implicit solvation models and includes continuum electrostatic, surface (cavity), and volume work terms. These are not further discussed here and comprehensive reviews can be found in the literature (see Refs. 217,218).

As mentioned above, the averaged molecular property is obtained from a population average according to Eq. 2.56. For calculations of a single isolated molecule, the respective Boltzmann

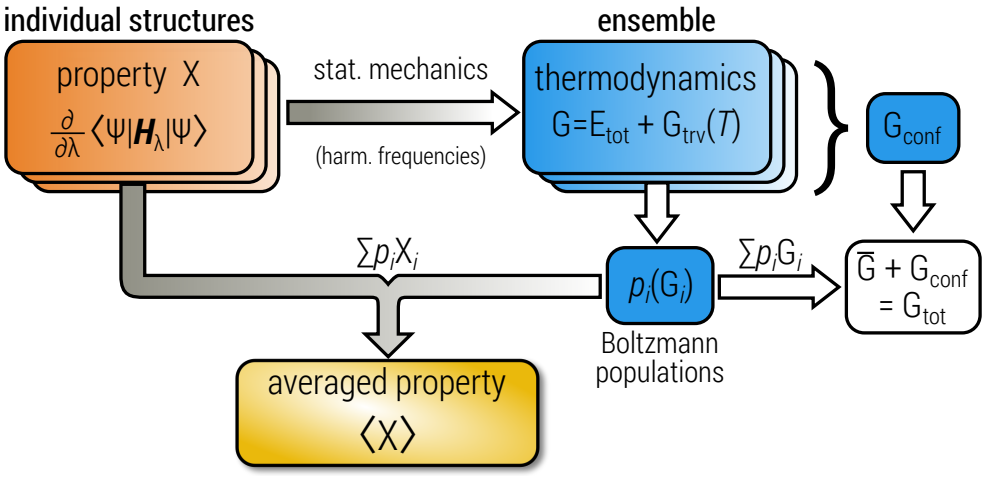

Figure 2.4.: Schematic representation of how to arrive at an averaged property $\langle X\rangle$. Correct Boltzmann populations under consideration of thermodynamic information are calculated from the individual free energies G_i of all conformers in the ensemble. The conformational partition function furthermore provides G_{conf}, which together with the population average $\overline{G} = \sum p_i G_i$ gives an absolute free energy G_{tot}.

populations are obtained from the conformational ensemble in order to satisfy ergodicity. Or, in other words, the conformational ensemble should represent the entire phase space of the molecule necessary to calculate the property average. Thermodynamic information is herein included by calculating the populations from the free energies according to Eq. 2.64 instead of total energies. However, the property itself and the vibrational modes required for G^T_{trv} are calculated directly from the total energy derivatives. A schematic representation can be seen in Fig. 2.4. For the purpose of calculating reaction free energies ΔG, *e.g.* in a supramolecular approach or thermodynamic cycles,[42,216] it is also advisable to calculate an absolute free energy

$$G_{tot} = \sum_i p_i\, G_i + G_{conf} \ , \tag{2.66}$$

where G_{conf} is the conformational contribution of the conformer ensemble (see Part III). The conformational term is often neglected but it can be significant (several kcal mol^{-1}) for flexible molecules.[216] Some thermodynamic properties, such as acid dissociation constants (pK_a) or phase partition coefficients (log K) are proportional to the reaction free energies and require an accurate description of G_{tot},[22] which translates back to accurate electronic structure methods.

Part II.

Quantum Mechanical Exploration of the Low-Energy Chemical Space

Part II (Chapter 3) is devoted to the CREST program and its applications. As outlined in the introduction, some requirements have to be met for theoretical methods used in the exploration of the low-energy chemical space. Due to limitations in modern days computing capacities, essentially only SQM methods provide the necessary compromise between accuracy, robustness, and computational cost suitable for this task. CREST provides an interface for respective calculations at the GFNn–xTB SQM level.

A key aspect in the construction of a program such as CREST is the automatization of calculations and efficient handling of data. Therefore, in Sections 3.1 and 3.2 a review-like overview is provided of automated workflows in quantum chemistry. The CREST algorithms are outlined in detail in Section 3.3. This includes definitions of conformers and rotamers, and threshold based identification thereof using differences of the total energy, atomic root-mean-square-deviations (RMSDs) and rotational constants (B_e). As a simulation technique for structure generation the RMSD-based metadynamics (MTD) simulations are re-introduced from Ref. 41. Herein, a history-dependent potential is used to alter molecular dynamics (MD) simulations and to accelerate the conformational sampling while maintaining the underlying physical plausibility and mechanics. Section 3.3 furthermore discusses the quality of conformational energies at the GFNn–xTB level compared to other SQM and DFT methods based on data from the well-known GMTKN55 database.[109] An initial connection is also made to the conformational entropy, which is discussed in more detail in Part III.

The remaining parts of Chapter 3 treat prototypical applications of CREST and GFNn–xTB. Challenging conformational searches are shown exemplary for the (S)-citronellal molecule, the Ac-Ala$_{19}$-LysH$^+$ polypeptide, several macrocyclic molecules and two organometallic systems (*trans*-Cu$^{\mathrm{II}}$(L-Valine)$_2$ and [Pt(COMe)$_2$(2-py)$_3$COH]). The same workflow is applied also to non-covalently bound aggregates such as the 1-napthol dimer and water hexamer. An unique feature of the CREST workflows is the constrained conformational sampling which was tested for the tyrosine molecule on a graphene surface cut-out and, in context of the Curtin–Hammett principle,[219] for transition state conformations of a S$_N$2 methyl group transfer. Applications of CREST to other representatives of the low-energy chemical space, namely (de-)protomers and tautomers were re-implemented from earlier GFNn–xTB based studies.[220,221] Examples are shown here also for organic and inorganic molecules.

Overall, the workflows and their applications shown in this chapter are characterized by excellent accuracy and broad field of usage provided by the state-of-the-art GFNn–xTB SQM methods. The computational robustness and efficient algorithms even enable the use of CREST on standard desktop computers.

3. Automated Exploration of the Low-Energy Chemical Space with Fast Quantum Chemical Methods

Philipp Pracht,* Fabian Bohle,* and Stefan Grimme*

This article is the front cover article for the April 2020 issue of Phys. Chem. Chem. Phys. and part of the "2020 PCCP HOT Articles" collection

Received 20th of December 2019, Published online 12th of February 2020

Reprinted (adapted) with permission from†
Pracht, P.; Bohle, B.; Grimme, S. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

DOI 10.1039/c9cp06869d

Own manuscript contribution

- Writing major parts of the CREST code
- Performing and supervising the computations
- Interpretation of the computed data
- Writing the manuscript

*Mulliken Center for Theoretical Chemistry, Institut für Physikalische und Theoretische Chemie, Rheinische Friedrich-Wilhelms-Universität Bonn, Beringstraße 4, 53115 Bonn, Germany
†Reproduced with permission from the PCCP Owner Societies.

Abstract

We propose and discuss an efficient scheme for the *in silico* sampling for parts of the molecular chemical space by semiempirical tight-binding methods combined with a metadynamics driven search algorithm. The focus of this work is set on the generation of proper thermodynamic ensembles at a quantum chemical level for conformers, but similar procedures for protonation states, tautomerism and non-covalent complex geometries are also discussed. The conformational ensembles consisting of all significantly populated minimum energy structures normally form the basis of further, mostly DFT computational work, such as the calculation of spectra or macroscopic properties. By using basic quantum chemical methods, electronic effects or possible bond breaking/formation are accounted for and a very reasonable initial energetic ranking of the candidate structures is obtained. Due to the huge computational speedup gained by the fast low-cost quantum chemical methods, overall short computation times even for systems with hundreds of atoms (typically drug-sized molecules) are achieved. Furthermore, specialized applications, such as sampling with implicit solvation models or constrained conformational sampling for transition-states, metal-, surface-, or non-covalently bound complexes are discussed, opening many possible applications in modern computational chemistry and drug discovery. The procedures have been implemented in a freely available computer code called CREST, that makes use of the fast and reliable GFN*n*–xTB methods.

3.1. Introduction

Over the past decades computational methods became a valuable tool in many modern fields of chemistry, and some kind of quantum chemical (QC) calculation can be found in almost every new publication. The big popularity of computational chemistry is also founded on recent advances in density functional theory (DFT) methods, which nowadays can routinely provide gas- or condensed-phase structures and energies for roughly a few hundred atoms.[3,32] For many interesting applications in biochemistry or supramolecular chemistry, however, those calculations are still too expensive. Classical force fields (FFs) are often employed as alternatives for long molecular dynamics simulations[222] but their limitations are manifold and they are not suited for general use, *e.g.*, for metallic systems. Additionally chemoinformatic procedures are used in the drug discovery with increasing popularity.[223–226]

Among the most important application of low-cost atomistic methods is the large scale structural sampling of molecular geometries, *i.e.*, the generation of an ensemble of low-energy structures, generally referred to as conformers. The knowledge about a molecules' conformations is highly important since all its properties are rooted in a thermodynamic ensemble average of the properties of its conformers that are accessible at finite temperature.[227,228] This gave rise to a large framework of computational approaches to generate and screen three-dimensional molecular structures where many challenges have to be faced in the generation process, such as the correct distinction between different conformers or the handling of macrocyclic systems. Hence,

a huge number of conformer generators based on different algorithmic approaches is available today.[229–238] One of the most common types of conformer generators are knowledge-based algorithms, which chemoinformatically try to reproduce structures from reference data (often taken from experimental crystal structures), or generate structures based on heuristic rules.[228] These approaches, sometimes also referred to as systematic methods, have the advantage of very short computation times, but are generally lacking a physically motivated methodology and thus are often not generally applicable. An example for this is the treatment of macrocyclic molecules, where special heuristics are required.[239–242] By contrast a general workflow based on quantum chemical calculations has no need for specialized rules and should recover any structural information of the molecule by an analysis of the potential energy surface (PES). Furthermore such general algorithms can also be exploited in order to find different arrangements of non-covalent bound aggregates or to generate conformations under geometrical constraints.

A related problem to conformational sampling is the prediction of molecular protonation and deprotonation sites. While there are many approaches to obtain these sites based on knowledge, *i.e.*, databases, the computational QC approach is just as simple.[243–245] Here the different structures in the chemical space have to be (automatically) generated and can then be ranked by their total energy (*i.e.*, proton affinities).[220,246,247] These so-called protomers are linked to some other important properties such as the pKa value.[248–252] Furthermore a sequence of protonation and deprotonation at different positions can be employed to obtain all prototropic tautomers of a molecule. By employing QC models that can intrinsically form and break bonds, such a tautomer screening procedure has the advantage that even complicated rearrangements can be recovered, which would otherwise require complicated rules in chemoinformatic treatments.[253,254]

In this article we present a new program called CREST (which is abbreviated from Conformer-Rotamer Ensemble Sampling Tool) and describe the underlying algorithms and typical applications. CREST employs a new scheme for the generation of conformational ensembles based on the direct sampling at a semiempirical QC (SQM) level, rather than using a knowledge based algorithm. While such an approach naturally can not compete with chemoinformatics driven procedures in terms of computation time, it has the clear advantage of providing reasonable conformational energies for basically any chemical species. All procedures introduced in this article are generally applicable and could in principle be employed at any quantum chemical level. However, semiempirical methods have an excellent cost-to-accuracy ratio, and are hence the preferred level of theory for such schemes which otherwise would be require supercomputer-resources.

In the first section we provide a very brief overview on automated quantum chemical procedures for the exploration of the chemical space in various flavors. We then discuss the quantum chemical perspective on conformer ensembles and the employed protocol, with a focus on the distinction between different conformations. Briefly some technical settings of the algorithm are explained and the performance of the used quantum chemical low-cost method (GFNn–xTB)[37–39] for conformational energies is discussed. Several examples are shown for standard and special applications of the CREST program, either in comparison with experimental ob-

servables or high-level theoretical reference data.

3.2. Automatized Quantum Chemical Procedures in the Literature

Although there are many chemoinformatic schemes for the exploration of chemical space, there exist only a few automatized procedures driven by quantum chemical calculations. The reason for this is the enormous amount of required calculations, leading to very high computational cost. However, the obvious inherent advantage of QC schemes is the possibility to generate and predict results in an *ab initio* fashion. They also yield much more chemical insight than purely heuristics guided results. There are several noteworthy efforts by our[40,220,221,255–257] and other groups[83,258–267] for efficient quantum chemistry driven algorithms to calculate various properties. One very large field of these automatized applications is the exploration of reaction mechanisms by so-called reaction networks, which recently gained popularity, *e.g.*, due to work by the groups of Maeda[258,259,268–273] and Reiher.[260,274–277] These reaction networks function by a throughout exploration of the normally reactive part of the PES under some pre-defined (*i.e.*, heuristics-guided) criteria, such as energy cut-offs or bias potentials.[258,269] QC methods are primarily applied in order to perform structure optimizations and transition-state (TS) searches.[259,271,274,275] Graph-based heuristic descriptors, from which verticies (*i.e.*, intermediate points on the different PES) and edges (*i.e.*, reactions) are created, are used to generate the eponymous *networks*, where a connected subgraph represents a single PES of the reaction. New intermediate structures are generated from a single starting point (referred to as the zeroth-generation structure) by identifying reactive sites and placing these in close proximity during a structure optimization. Here it is important to consider the relative orientation of the fragments. In the geometry optimization either the reactants are recovered or a new species is formed, which is detected and yields an approximate reaction path. The exploration of reaction paths can also be refined by the inclusion of conformational sampling.[275] Recently the concept of chemical reaction networks was extended with the application of a kinetic analysis by using semiempirical QC methods.[278,279] This is done for the reduction of noise within the possible reaction pathways by removal of kinetically unfavorable structures. The motivation here is to reduce the computational cost as far as possible, while still maintaining a reasonable degree of accuracy, which is in fact also one of the main motivations of our work. All the different tasks and the building of the network itself can be automatized and parallelized in an efficient computer code, which makes the procedure feasible even at the underlying DFT or wave function theory (WFT) level. The automatized exploration of reaction mechanisms is a huge field in computational chemistry and more comprehensive reviews can be found in the literature (see Refs. 259,265,278,280,281).

Another approach to automated QC is to provide the infrastructure for computational workflows. Representative programs here are for example the PyADF[83] and QMflow[267] frameworks as pioneered by Visscher *et al.* or the atomistic simulation environment ASE.[282] In general these type of programs allow the setup of multiple quantum chemical calculations in a script-

like manner, coupled to some semi-automated analysis of the results. However, many large computational chemistry program packages today also come with their own scripting environments such as the PLAMS driver distributed with ADF2019.[283,284] Although the setup of these automation is certainly less exhausting than conventional scripting, it can still be tedious and requires a large amount of user input. If, however, the setup was done once it can be re-used for multiple calculations.

The research in our own group concerning automated processes so far mainly focused on the generation of spectral data, such as UV-Vis/circular dichroism (CD),[285] nuclear magnetic resonance (NMR)[40], and mass spectra.[255–257] With the introduction of the tight-binding method GFN–xTB[38] (short for Geometries, Frequencies, and Noncovalent interactions – extended Tight-Binding) the efficient automation of screening processes at a semiempirical QC level of theory became feasible and was applied to conformers, protomers, and tautomers including all elements up to Radon. The first SQM based conformer generator was published in scope of the automated calculation of NMR spectra,[40] while the screening of protonation sites and tautomers/isomers had been published in Refs. 220 and 221 accordingly. Since then, the procedures have been refined and implemented into a single computer code (*i.e.*, CREST), which is the subject of this work.

3.3. The Automatized Conformational Search Algorithm

We introduce a new tool called CREST (short for Conformer-Rotamer Ensemble Sampling Tool), for the automatized exploration of the low-energy chemical structure space normally not consisting of any covalent bond break/formation. As its name implies, the main application of CREST is the generation of conformer ensembles with an algorithm called iMTD-GC,[41] but other related applications, such as the screening of different non-covalently bound aggregates, or the screening for different protonation sites are also implemented. An overview of the different procedures and their general workflows is shown in Fig. 3.1.

3.3.1. Identification of Conformer Ensembles

Stereoisomers of a molecule that differ only in their conformation but have the same covalent topology are referred to as conformers and can be characterized by a distinct potential energy minimum. By rotation around covalent chemical bonds (or other complicated inversion-type processes) that interchange nuclei belonging to the same group of nuclides, as for example the interchange of H nuclei at a methyl group, so called rotamers arise. Rotamers have degenerate potential energy minima and thus are indistinguishable by any nuclear spin-independent quantum mechanical observable computed at the respective minima (see Fig. 3.2). In the following the Born–Oppenheimer-approximated equilibrium conformer including all its rotamers is referred to as "conformation". A set of different conformers and their rotamers within a certain energy window around the same global covalent potential energy minimum is referred to as the conformer/rotamer ensemble (CRE).

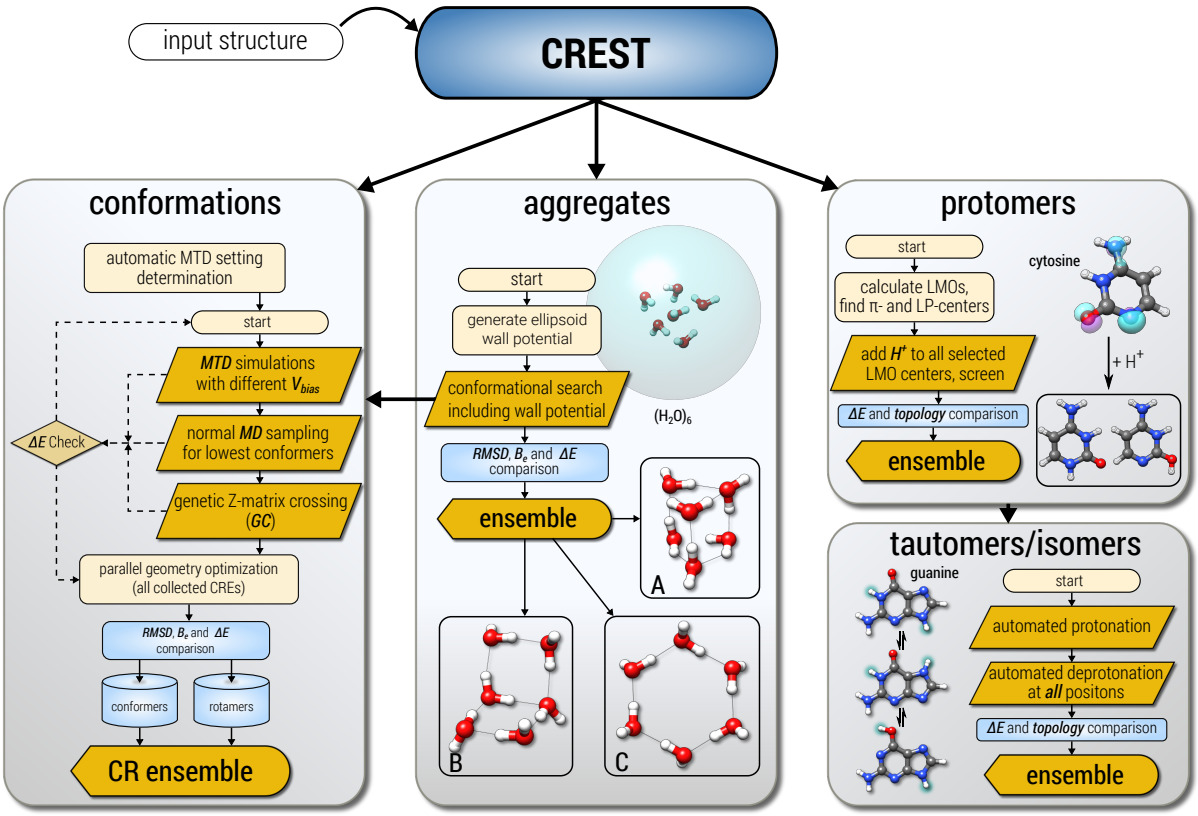

Figure 3.1.: Overview of the automatized quantum chemical sampling procedures that have been implemented in the CREST code.

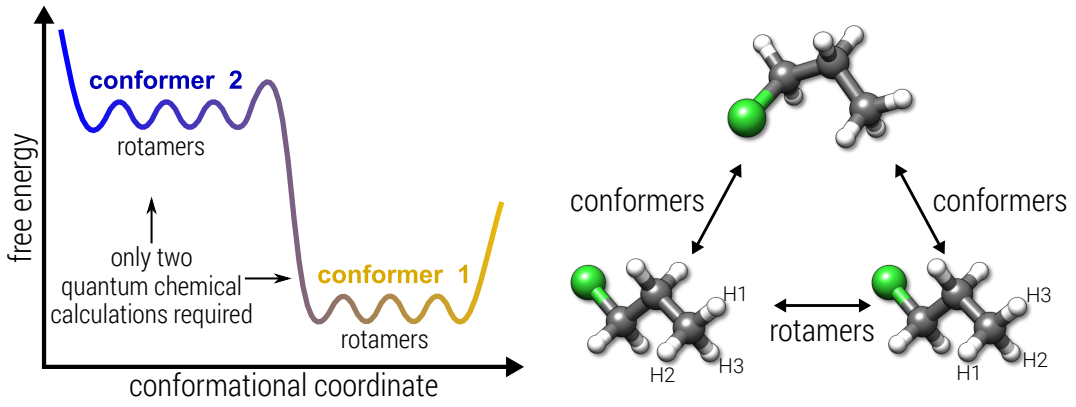

Figure 3.2.: Relation of conformers and rotamers using the example of 1-chloropropane and their schematic representation on the potential energy surface.

For the calculation of properties it is often necessary to include different molecular conformations by averaging the individually obtained Boltzmann weighted property of each constituent in the ensemble. Some examples where this ensemble average is highly relevant are nuclear magnetic resonance (NMR) spectra,[40,286–288] circular dichroism (CD) spectra,[285,289–291] or pK_a values.[221,292] In order to avoid double counting, which leads to incorrect Boltzmann averages and subsequently falsely averaged properties, the precise distinction between identical struc-

tures, conformers, and rotamers is mandatory. This distinction is possible on the basis of three dimensional structures and (free) energies of the isomers. The energy is employed as a criterion, since each conformer is characterized by its own minimum on the PES. For purely structure based comparisons the root-mean-square deviation (RMSD) of atomic Cartesian coordinates and the difference between rotational constants (B_e) of two molecules can be used. However, structural parameters do not include any information whether the structure is an equilibrium geometry or some higher-energetic intermediate.[293] Therefore, structural information must always be combined with the energetics for correct identification of different conformations. In practice, however, many knowledge-based conformer generators still only employ structural criteria (two or three dimensional) for the distinction of conformers,[229,238,294] which can be useful, *e.g.*, for large scale screening of databases.[228,295]

The distinction between identical isomers, conformers, and rotamers on the basis of the energy, the atomic Cartesian RMSD and the molecular rotational constant is schematically outlined in Fig. 3.3. For practical reasons one has to work with predefined thresholds in order to eliminate

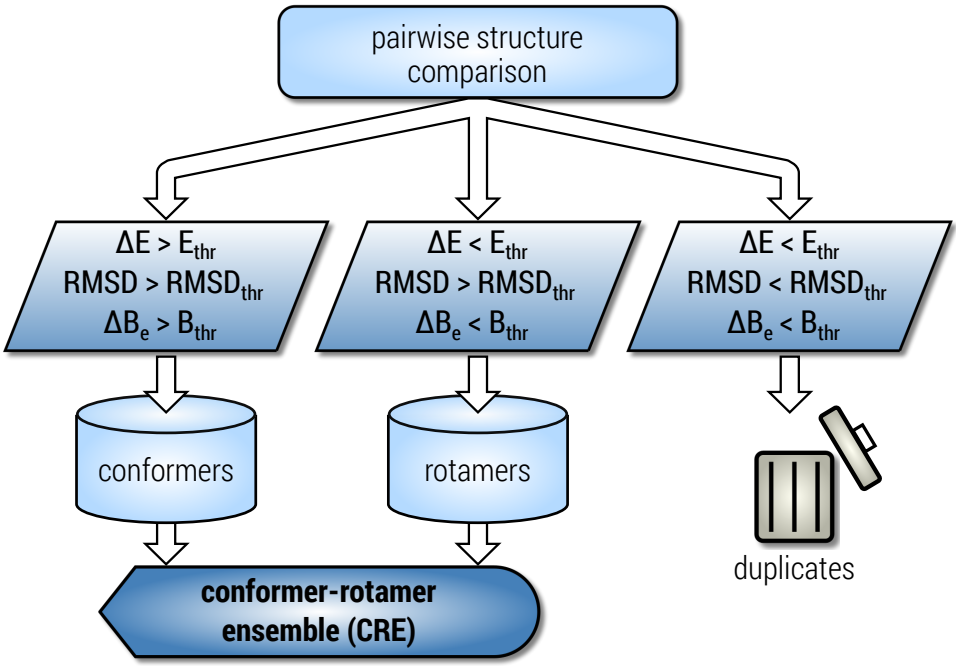

Figure 3.3.: Schematic representation of the sorting criteria to distinguish between identical structures, conformers and rotamers. E_{thr}, RMSD_{thr}, and B_{thr} are the respective predefined thresholds for the energy, atomic RMSD between the considered pair, and rotational constant.

the effect of numerical noise. Conformers are those structures that either have different PES minima ($\Delta E > E_{thr}$), or, if the energy difference is small, a high RMSD and unequal B_e. Two structures with similar energetics can be rotamers if their atomic coordinates differ, but at the same time the rotational constants are equal. Only if two structures have the same energy and matching structural criteria ($\mathrm{RMSD} \approx 0$ and $\Delta B_e \approx 0$), they can be discarded as duplicates.

The final CRE for further practical calculations typically consists of all unique conformers and rotamers within a certain energy window. The choice for this window depends on the accuracy of the used QC and the type of application (*i.e.*, sensitivity of the target property to details of the conformational ensemble).

The quality of an ensemble is related to its completeness which can be assessed by a maximized entropy S_{CR} according to the standard thermodynamic expressions

$$S_{CR} = -R \sum_{i=1}^{CRE} p_i \ln p_i \quad . \tag{3.1}$$

where R is the molar gas constant and the sum runs over all populations p_i of all species with energy E_i at temperature T and the Boltzmann constant k, given as

$$p_i = \frac{e^{-E_i/kT}}{\sum_j^{CRE} e^{-E_j/kT}} \ . \tag{3.2}$$

The ensemble entropy S_{CR} is also linked to the ensemble free energy (at T=298 K) $G_{CR} = -TS_{CR}$, which is minimized for a complete CRE. This completeness criterion of a maximized S_{CR} only holds if the global minimum conformation is included in the ensemble and breaks down otherwise. Therefore, finding the lowest energy conformation of a molecule is the one of the defining tasks a conformer generator must be able to perform robustly. A maximized ensemble entropy was used for determining technical parameter sets that are employed in the CREST program and is discussed in Ref. 41.

Practically, it is difficult to assign a quality and/or completeness to an ensemble, without knowing the "true" conformations. The practical approach for identifying the "true" low-energy conformations are experimental measurements. Crystal structures determined from X-ray are the most common source of experimental geometries. However, interpretation of CREs with respect to crystal structures can be highly problematic, since conformations in the solid can differ significantly from structures in the gas-phase or in solution, *e.g.*, due to packing effects.[296,297] Furthermore, crystal structures intrinsically yield only one or a few conformers instead of entire ensembles. Other experimental techniques for structure elucidation, such as solution NMR (*e.g.*, using variable temperature NMR, NOESY, residual dipolar and ^{3}J scalar couplings), microwave spectroscopy, and gas electron diffraction, are less common. In general experimental conditions will always have an influence on the composition of the ensemble. Even under ideal experimental conditions identifying the global minimum might not be possible, *e.g.*, due to kinetic trapping. For these reasons theoretical calculations provide a valuable alternative to experiment for obtaining CREs under idealized environment. In the literature one can find a variety of benchmark studies[295,298–301] where computer generated conformers are compared to experimentally observed ones in order to assess the performance of different conformer generators. Although a comparison like this gives some insight into the ensembles, it has to be evaluated with caution in respect to the ensemble completeness and the different "measurement" con-

ditions. Therefore in the following sections, rather than benchmarking on crystal structures, we will compare selected conformer ensembles in the gas or liquid phase with spectroscopically evaluated structures.

3.3.2. Algorithmic Details

Generating conformations, *e.g.*, by rotation around dihedral angles is impractical for larger and flexible molecules. In addition this approach requires the manual *a priori* definition of the conformational coordinates (*i.e.*, the angles). To remedy this, we recently proposed a metadynamics (MTD) simulation based screening procedure that can be routinely used for the generation of molecular conformations in the gas-phase or in implicit solvation.[41] A history-dependent biasing potential is applied, where the collective variables (CVs) for the metadynamics are previous structures on the PES, expressed as atomic RMSD between them, which is calculated according to a quaternion algorithm.[302] Although being a well-known concept of MD simulations[303,304] and being used before in the general context of conformational changes,[305–309] it is, to our knowledge, the first combination of MTD simulations with atomic RMSDs in order to generate conformers. The biasing contribution is given in the form of a Gaussian potential by

$$V_{bias} = \sum_{i}^{n} k_i \exp(-\alpha_i \Delta_i^2) \ , \tag{3.3}$$

where the RMSDs enter as collective variables Δ_i, n is the number of reference structures, k_i is the pushing strength and the parameter α_i determines the potentials' shape. From this energy expression atomic forces are derived that enter as additional forces in the MTD simulations, which is also sometimes referred to as *guiding forces*.[306] Since the addition of each bias Gaussian potential drives the structure further away from previous geometries this allows otherwise unlikely high-barrier crossings where all atoms collectively explore huge regions of the PES. A schematic representation of a 1-dimensional PES that is filled by additive bias potentials over time is given in Fig. 3.4. For more realistic examples see Sections 3.4 and 3.6.1.

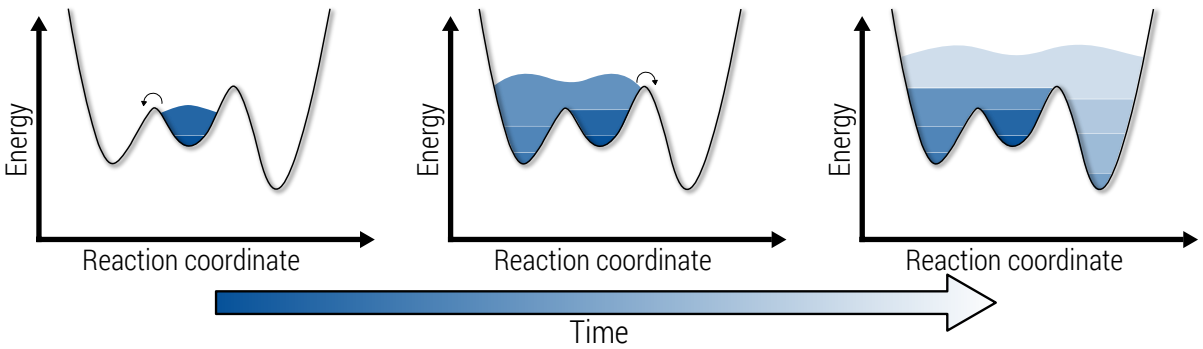

Figure 3.4.: Schematic one dimensional PES that is "filled" by several bias potentials over time, which allows larger barrier heights to be overcome.

While calculations in this manuscript were conducted entirely at the semiempirical tight-

binding level, it has to be noted that the application of a RMSD-based bias potential is a general approach that, in principle, works at all levels of theory, *i.e.*, at the FF, DFT or even WFT level. Furthermore, the employed CVs do not have to be atomic RMSDs, but could also be some other kind of alignment factor between structures, as long as forces can be obtained from the partial derivatives $\frac{\partial \mathrm{CV}}{\partial r}$.

In the literature one can also find other MD based approaches with various bias potentials under the general keyword *accelerated molecular dynamics* (aMD).[310–318] The basic idea of those aMD approaches is to smooth the PES and to fill its minima by reshaping the potential. This is done in order to decrease inter-conformational energy barriers and enable the simulation to explore larger regions of the PES. The aMD approach was already successfully applied in the generation of molecular conformations, *e.g.*, for macrocycles.[319] The fundamental difference of aMD to the RMSD based MTD approach is the missing directionallity of the PES exploration. In aMD the shape of the energy surface is in general retained after addition of the bias potential, *i.e.*, during the simulation it is possible to arrive at the same minimum on the PES again. With the RMSD based approach however, previous minima on the PES are "occupied" by the V_{bias}, leading to history dependent forces and thus to an implied directionallity of the simulation. Since the potential in aMD is modified by a single bias potential and retains the general shape of the energy surface, very long simulation times can still be required in order to sample the entire conformational space. This time can be expected to be much shorter with a history dependent guiding force. It must be stressed however, that the target quantities of the two approaches are slightly different: In aMD, the desired quantity usually is the canonical ensemble average of some observable on the unmodified PES, which can simply be obtained by back-correcting the observable average on the biased energy surface.[310,318,320] In our MTD based approach the targets are the "true" quantum chemical energy minima, as defined in section 3.3.1. Therefore, the latter requires a separate geometry optimization of the generated MTD structure snapshots.

For the automatized generation of conformers we developed a composite approach consisting of MTD sampling, regular MD sampling and a procedure that is related to genetic structure crossing algorithms (GC).[230,321,322] Hence, the procedure was termed iMTD-GC, where the lowercase i indicates an iterative strategy within the algorithm. As mentioned above, the approach heavily relies on the semiempirical GFNn–xTB methods (n = 0–2),[37–39] which offer the possibility for fast and robust calculations. The general workflow is outlined in Fig. 3.5.

First, the maximum MTD length is determined, which mainly depends on the molecular size and flexibility of the system. Then, technical general settings are evaluated, to check if the MTDs will run stable. The main step of each iMTD-GC conformational search is the MTD sampling. Since the automatization is the key step and different molecules require adjusted pairs of k_i and α_i to produce the best results, a set of twelve MTDs is performed with different settings for the V_{bias} parameters. Here, α_i typically has values between 0.1 and 1.3 Bohr^{-2}, which can be seen as the "range" of the bias. The constant k_i is scaled by the number of atoms N, where k_i/N has magnitudes of 0.75 to 3.00 mEh. Within the MTD simulation a new structure is added to the V_{bias} potential every 1.0 ps, which constantly drives the molecule

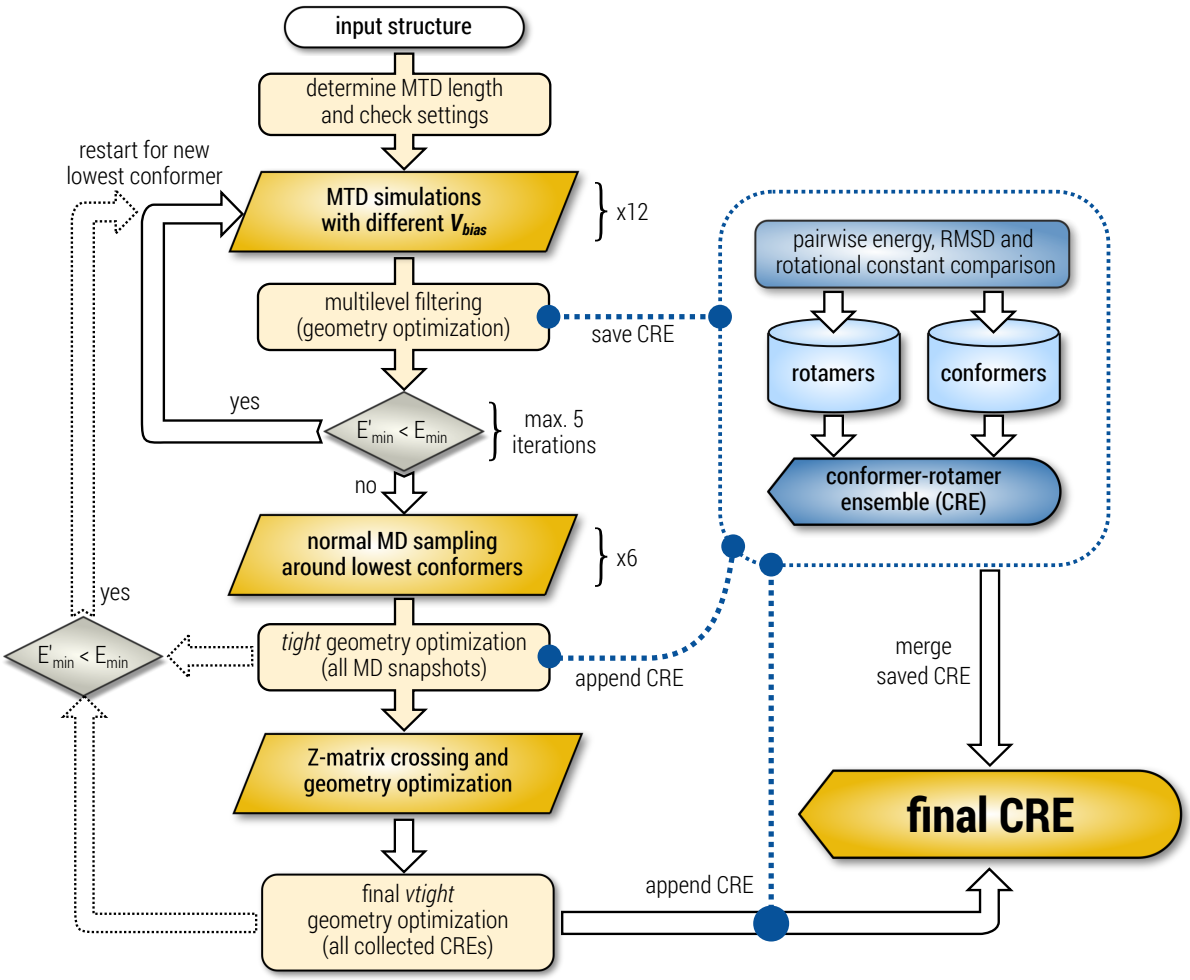

Figure 3.5.: Outline of the iMTD-GC workflow.

into new conformations as time progresses. However, since the PES is constantly modified by the bias potential, the conformers form the MTD trajectory can not directly compared to each other and have to be re-optimized without the biasing potential. This is done in a two-step filtering procedure, first with very crude and then with tight convergence criteria. Afterwards the re-optimized structure snapshots from the trajectory are sorted according to the procedure outlined in Fig. 3.3, which yields an initial CRE. If a new conformer is found that is lower in energy than the input structure, the entire procedure is restarted on this conformer, otherwise the workflow is continued. By default, the MTD iteration is restarted at least once, but not more than five times. All intermediate CREs are saved to be compared at a later stage. In the second step two unbiased MD simulations (*i.e.*, at two different temperatures 400 K and 500 K) are run on the three lowest conformers. This is done to get conformations with low-energy barrier crossings, opposed to high barrier conformational changes that can be obtained by the MTD simulations. The low-energy barrier crossings include simple torsional motions, such as group rotations, which are needed to complete the CRE regarding the rotamers. All structures

from the MD simulations are sorted again and included in the intermediate CRE. In the final step the genetic structure crossing (GC) is performed with automatically generated Z-matrices, as was described in previous publications.[40,41] Together with the regular MD simulations this approach helps to further complete the CRE and is particularly useful for flexible systems, *e.g.*, with many alkyl chains. If in the MD or GC step a new lowest energy conformer is found, the entire procedure is restarted. However, unlike the MTD iterations, these iterations do not have a maximum number of cycles and will only terminate if the lowest conformer does not change any more. The advantage of these restarts is mainly observed for larger molecules whose global minimum structure can be way off the initial input geometry. All collected CREs are then optimized once more with very tight energy convergence criteria and the final CRE is created. Various energy thresholds and other MTD settings are employed within the workflow, which will be discussed in detail below.

3.3.3. Conformations at Low-Cost QM Level

Conformations are generated at the GFN*n*–xTB level within the iMTD-GC workflow, as implemented in the CREST program. For the reliable generation of conformers at a semiempirical level there are two main questions that have to be answered: First, "how trustworthy are SQM conformations (ΔE and geometries) compared to higher level theoretical methods such as density functional theory (DFT)? " And secondly, "can experimentally observed low-energy conformations be reproduced at a low-cost level of theory?"
Concerning the first question a huge amount of literature exists in which various theoretical methods are benchmarked on conformational energies and geometries. It seems to be consensus in the literature that geometries are often quite well reproduced by SQM methods.[323,324] This is particularly true for methods of the GFN*n*–xTB family, which, as their name conveys, are parameterized to yield reasonable structures.[38,39,324] However, the calculation of accurate conformational energies at a semiempirical level remains difficult since small energy differences have to be described quite accurately. In Fig. 3.6 the mean average deviations (MADs) for some SQM methods are shown as evaluated on subsets of the GMTKN55 database[109] that investigate conformational energies and the MALT205 set containing the energies for 205 conformers of maltose.[325] The deviation of conformational energies from highly accurate WFT reference values is very dependent on the type of molecular system. *E.g.*, for simple alkane isomers in the ACONF set the MADs of all depicted semiempirical methods are well below 1 kcal mol^{-1}, while for the conformers of maltose (MALT205 set) all methods show deviations >3 kcal mol^{-1}. Likewise, different semiempirical methods do not describe all systems equally well. For example the Hartree-Fock derived method PM6-D3H4X[66,326,327] has a MAD for the SCONF set that is more than twice the MAD of the tight-binding based methods, while for the melatonin conformers in MCONF it is the best performing semiempirical method. In Fig. 3.7 the mean average deviations (MADs) from Fig. 3.6 are averaged for different levels of theory, which provides a more general overview for the average performance of conformational energies. Their accurate description requires a balanced description of covalent and non-covalent energy contributions.

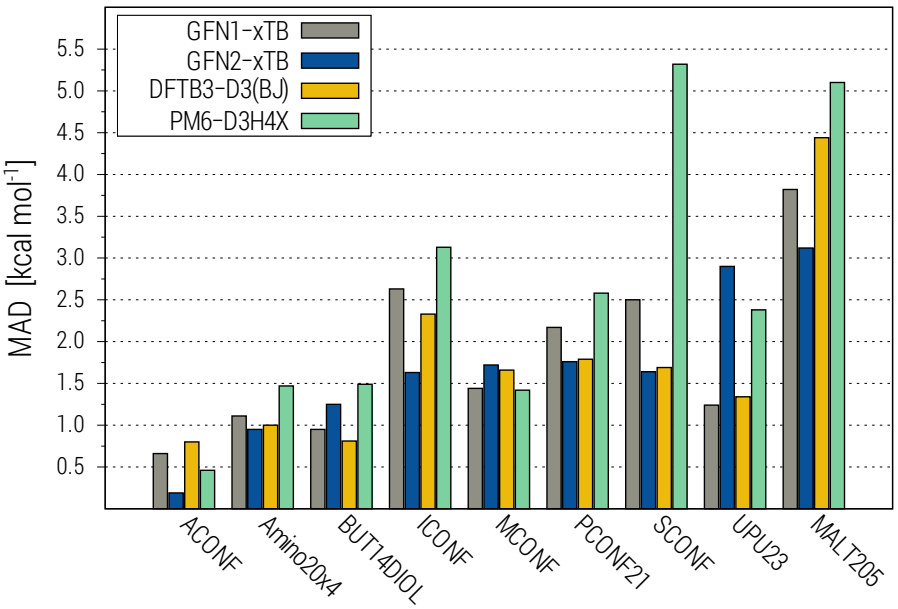

Figure 3.6.: MADs for conformational energies evaluated for GFN1–xTB, GFN2–xTB, DFTB3-D3, and PM6-D3H4X on the ACONF, Amino20x4, BUT14DIOL, ICONF, MCONF, PCONF21,SCONF, and UPU23 subsets of the GMTKN55 database and the MALT205 benchmark.

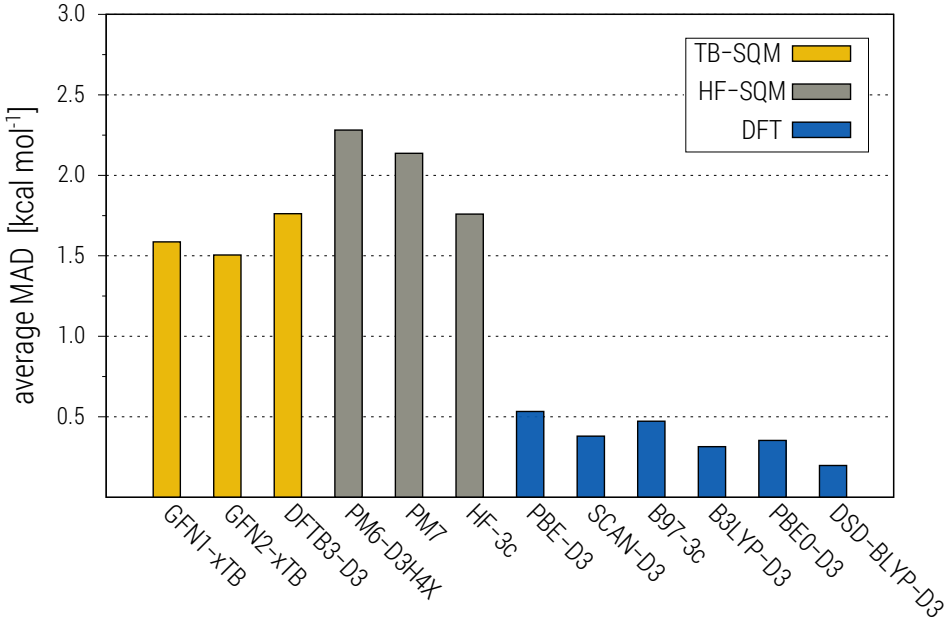

Figure 3.7.: Averaged MADs for the subsets shown in Fig. 3.6, evaluated for different levels of theory.

As Fig. 3.7 depicts, this balance in general is much better at the DFT level than at any semiempirical level. The MAD of conformational energies calculated by semiempirical methods is on average more than three times higher than at even a "cheap" DFT level. At the DFT level, the

PES appear to be much smoother and consistently shaped. The same observation was also made in other publications.[328] Nevertheless semiempirical methods allow much shorter computation times while still maintaining a reasonable level of accuracy, which is sufficient for qualitative results. Several recent studies also show that the GFNn–xTB methods are among the best performing semiempirical methods for conformational energies and geometries.[293,329] The good trade-off between accuracy and computational cost enables the use of GFNn–xTB for the generation of conformers with the iMTD-GC workflow. Although it is technically possible, any higher level theoretical method, even low cost DFT with a small basis set, would be much to expensive for the vast amount of required geometry optimizations and evaluations.

3.4. Selection of Default Thresholds and Settings in CREST

Any (semi-)automated screening procedure requires the application of pre-defined thresholds of various kinds. The by far most important threshold is the energy window, *i.e.*, the maximum relative energy up to which structures are considered further. This is naturally related to the population of the structures at finite temperature, which is calculated according to Eq. 3.2. For conformational energies at a semiempirical level the default size of this window is 6 kcal mol^{-1} in CREST, which is a reasonable but still conservative choice for many systems.[41,329] This window is applied even though significant Boltzmann populations at 298 K are obtained only up to approximately 2 kcal mol^{-1}. The larger "save choice" value should account for the non-parallel PES of the semiempirical and higher level QM methods. In fact, the results from benchmark studies suggest (*cf.* Fig. 3.6), that in some cases it may be necessary to increase the energy window to 10 kcal mol^{-1} or more. For applications that involve chemical changes such as the protonation, even larger energy windows have to be applied in order to recover all thermodynamically accessible structures. By default, the corresponding energy threshold for the protonation, deprotonation and tautomerization applications is 30 kcal mol^{-1}. Other thresholds are applied for the identification of conformers and rotamers as discussed in Section 3.3.1. These thresholds are used to quantify the difference between two structures according their relative energies (E_{thr}), atomic RMSD (RMSD$_{thr}$), and rotational constants (B_{thr}). The default values are given in Tab. 3.1. The selection of default threshold values has a significant influence on the

Table 3.1.: Overview of various default thresholds applied in the CREST program for structural comparisons.

threshold	value
E_{win} energy (conformers)	6.0 kcal mol^{-1}
E_{win} energy (prot./deprot./taut.)	30.0 kcal mol^{-1}
E_{thr} energy (between conformers)	0.1 kcal mol^{-1}
RMSD$_{thr}$	0.125 Å
B_{thr} (rot. constant)	15.0 MHz

performance and results of an automated procedure, particularly on the computational wall-times. Hence, it is important to carefully choose and adjust these settings. If for example a smaller energy window is chosen, more structures will be discarded, which leads to shorter computation times but also to less complete ensembles. Or, as another example, if the RMSD threshold is increased, more structures will be (falsely) identified as the same conformer, also leading to smaller ensembles and consequentially lower computational cost. Other settings, such as the simulation time of the MTD simulations, have a more direct influence on the performance of the workflow and are system dependent. Due to this dependence such settings have to be determined dynamically for each simulation. In case of the MTD time t_{mtd} (in picoseconds) the scaling is chosen to be dependent on an effective number of atoms N_{eff},

$$t_{\mathrm{mtd}} = 0.1\,(N_{eff} + 0.1\,N_{eff}^2)\,, \qquad \text{where } 5\,\mathrm{ps} \leq t_{\mathrm{mtd}} \leq 200\,\mathrm{ps}. \tag{3.4}$$

This is justified because larger molecules typically require longer simulations in order to undergo conformational changes. The effective atom number is obtained from the total atom number N and a flexibility measure according to $N_{eff} = N\xi_{\mathrm{f}}$, where ξ_{f} is given by

$$\xi_{\mathrm{f}} = \sqrt{\frac{1}{N_{\mathrm{bonds}}}} \left(\sum_{i}^{N_{\mathrm{bonds}}} \left(1 - \mathrm{e}^{-5(\mathbf{B}_{\mathrm{AB}}-2)^{10}}\right)^2 \frac{4}{N_{\mathrm{A}}^{\mathrm{neigh}} N_{\mathrm{B}}^{\mathrm{neigh}}} \left(R_i^{(\mathrm{f})}\right)^2 \right)^{\frac{1}{2}} . \tag{3.5}$$

Here, the summation runs over all *non*-terminal bonds i with the involved atoms A and B (*i.e.*, A,B $\in i$), $\mathbf{B}_{\mathrm{AB}}$ is the Wiberg-Mayer bond order[330,331] between the two atoms as obtained from a GFNn–xTB calculation, and $N_{\mathrm{A,B}}^{\mathrm{neigh}}$ are the numbers of neighboring atoms of A and B. The predefined factor $R^{(\mathrm{f})}$ is 1 if the bond i is not part of a ring and < 1 (depending on the ring size) if it is. In total, the flexibility measure ξ_{f} can be $0 < \xi_{\mathrm{f}} < 1$, where values close to 1 indicate an highly flexible system, *e.g.* n-alkanes, and values $\ll 1$ indicate rigid systems. For other proposed flexibility measures, see *e.g.* Refs. 332,333.

Another important system dependent variable is the bias potential V_{bias} (see Eq. 3.3). Since the RMSD is a normalized variable, potentials of the same order of magnitude would be obtained for any system. However, larger molecules do require higher bias to undergo conformational changes. Therefore, the potentials V_{bias} have to be scaled by the system size, which in practice is done by scaling the pre-factors in Eq. 3.3 by the number of atoms, *i.e.*, $k_i = k_i' N_{at}$. Finding and optimizing a reasonable set of pairs for the variables k_i and α_i is a non-trivial task and was done by hand.[41] Different combinations of the two factors will act differently upon any system and hence a set of all combinations of three different k_i' and four different α_i (yielding in total 12 MTD simulations, see Fig. 3.5) is applied. Additionally, two further simulations are performed with extreme value combinations. For specialized applications such as the NCI–iMTD (for non-covalently bound complexes, see Section 3.7.2) different sets of parameters are employed. The default combinations of k_i' and α_i are listed in Table 3.2.

The influence of a single bias potential on a two dimensional PES of 1-bromo-3-chloropropane,

Table 3.2.: Combinations of the parameters k_i' and α_i as applied by default in the CREST program. k_i' is multiplied by the number of atoms N_{at} in order to obtain k_i of Eq. 3.3.

MTD	k_i' [mEh]	α_i [Bohr^{-2}]
1	3.00	1.300
2	1.50	1.300
3	0.75	1.300
4	3.00	0.780
5	1.50	0.780
6	0.75	0.780
7	3.00	0.468
8	1.50	0.468
9	0.75	0.468
10	3.00	0.281
11	1.50	0.281
12	0.75	0.281
13	1.00	0.100
14	5.00	0.800

obtained by rotation around two dihedral angles, is shown in Fig. 3.8. For this small molecule the entire conformational PES is described by the two dihedral angles φ and θ. As can be seen from Fig. 3.8b the surface is symmetric with two global minima (θ, φ) at $(68°, 66°)$ and $(292°, 294°)$, which correspond to the different enantiomers of 1-bromo-3-chloropropane. Since two enantiomers are distinguishable by their atomic RMSD it is possible to "fill" one of the global minima with a bias potential while retaining the other. As a consistency check, by choosing the enantiomer $(292°, 294°)$ (Fig. 3.8a) as the center for V_{bias}, the biased PES in Fig. 3.8c is obtained, which has only the global minimum at $(68°, 66°)$. Furthermore, in this example a combination of k_i and α_i similar to setting 10 of Tab. 3.2 was employed, which provides valuable insight about their effect on the PES.

However, for small molecules (less 20 to 30 atoms) it is often not necessary to perform many MTD runs and some reduced run-types exist for speeding up the sampling procedure in CREST (*i.e.*, with the command line keywords "`-quick`","`-squick`", and "`-mquick`").

3.5. Computational Details

All shown screening procedures are implemented in a computer program called CREST which makes use of the `xtb` program. CREST makes full use of single node (OMP) parallelization in order to execute several independent `xtb` calculations at once. All calculations executed with `xtb` were performed using the 6.2 release version of the program. DFT calculations were performed with the `TURBOMOLE.7.3.1` program. The resolution-of-identity (RI) approximation for the Coulomb integrals[132] was generally applied using the matching default auxiliary basis

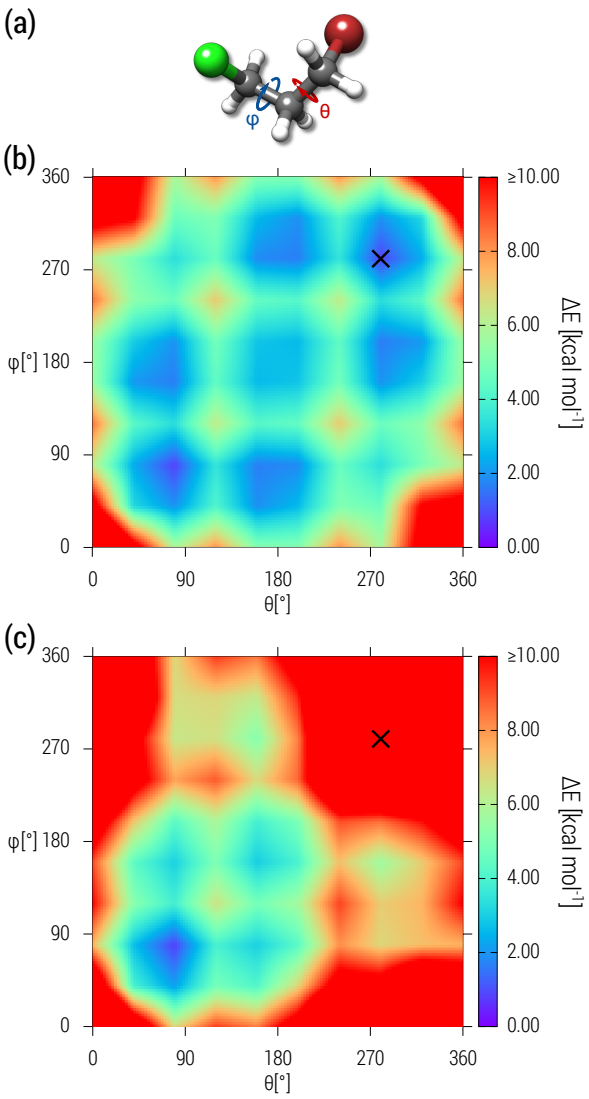

Figure 3.8.: Comparison of the unbiased and biased two dimensional PES of 1-bromo-3-chloropropane at the GFN2–xTB level. **(a)** Molecular structure and the dihedral angles φ and θ used for the construction of the two dimensional PES of 1-bromo-3-chloropropane. **(b)** Unbiased PES at the GFN2–xTB level. **(c)** PES under influence of a single bias potential V_{bias} with $k_i = 0.03$ Eh and $\alpha_i = 0.3$ Bohr^{-2} at the GFN2–xTB level. The center of V_{bias} is is marked by "**X**", which corresponds to the global minimum structure shown in **a**.

sets.[334] The integration of the exchange-correlation contribution was evaluated on the numerical quadrature grids *m4*. The default convergence criteria for single-point energies were 10^{-7} Eh.

3.6. Conformational Search Examples

3.6.1. Conformations of (S)-citronellal

Citronellal is an acyclic monoterpene with a characteristic lemon scent. This molecule was chosen as it represents a typical organic molecule concerning size and flexibility and is therefore well suited to demonstrate the standard application of CREST.

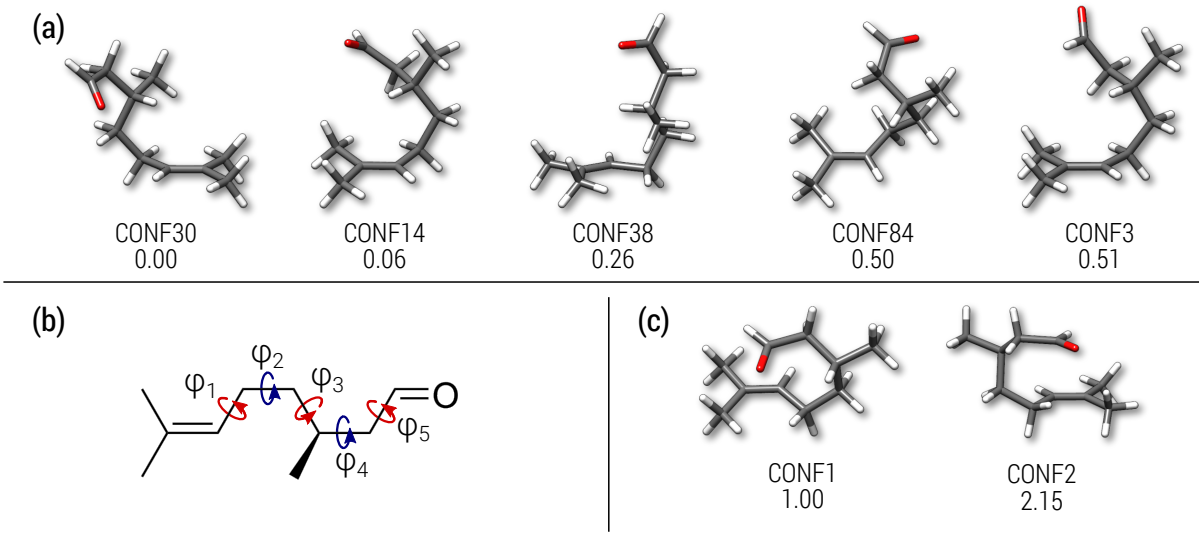

Figure 3.9.: Citronellal gas-phase conformers depicted with calculated enthalpies at 0 K in kcal mol^{-1}. **(a)** 15 conformers were experimentally identified by microwave spectroscopy and all 15 conformers were found in the calculated CRE. Only the five highest populated conformers are shown. Enthalpies were calculated at 0 K as the sum of PBEh-3c[158] energies and zero point vibrational energies from GFN2–xTB. **(b)** Lewis structure of (*S*)-citronellal with unhindered non-methyl-dihedral angle rotations highlighted. **(c)** Conformers obtained initially at xTB level are too high in energy at DFT level and not found under experimental conditions.

The (*S*)-stereoisomer and its conformers were investigated. Because of the acyclic geometry and with five freely rotatable C-C bonds it is to be expected that the conformational space of citronellal is relatively large. Fig. 3.10b shows a part of the multidimensional unbiased PES of citronellal along two dihedral angles. The two dimensional PES shows many deep pocketed minima connected by high energetic barriers of approximately 5-6 kcal mol^{-1}. In Fig. 3.10c, the 2D-PES in a MTD run (with bias potential) is shown, which reduces the complexity of the PES in terms of accessible conformations considerably.

The gas-phase iMTD-GC calculation generated 262 conformers, within the conservatively chosen energy window of 6 kcal mol^{-1}. As already mentioned, large energy windows allow the compensation of non-parallel energy surfaces of SQM and DFT (see Fig. 3.7) thereby preventing the loss of potentially low lying conformers at DFT level after re-optimization. The conformers can be classified into two main types: chain-like and globular-folded conformations. By visual inspection of the CRE it becomes apparent that the relative orientation of the aldehyd

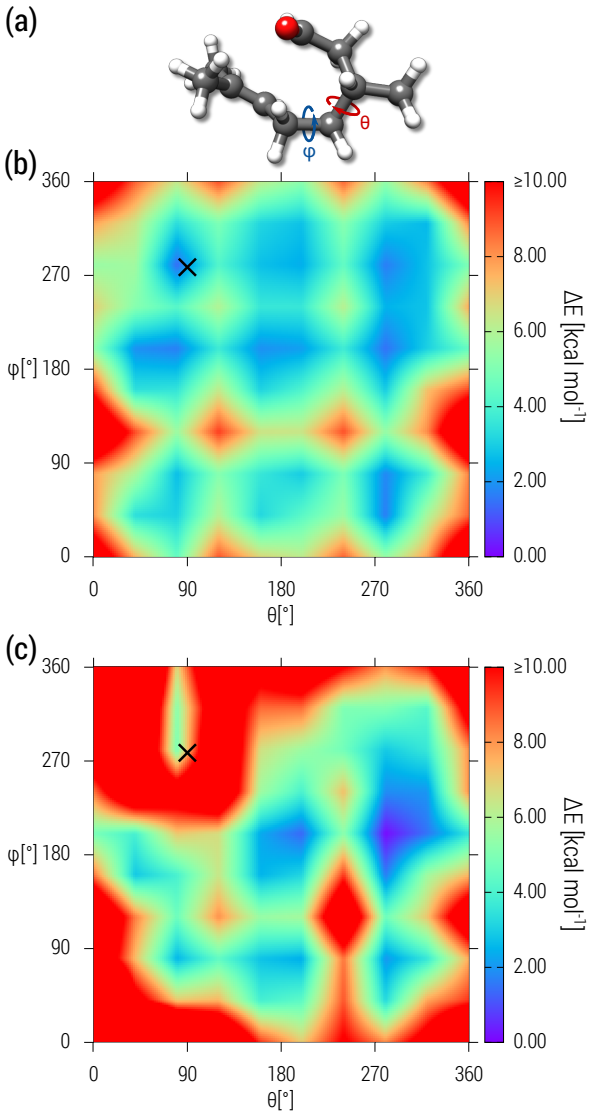

Figure 3.10.: Comparison of the unbiased and biased two dimensional PES of citronellal at the GFN2–xTB level. **(a)** Citronellal with indicated dihedral angles φ and θ for the construction of the two dimensional PES. **(b)** Unbiased PES of citronellal. **(c)** PES with a single bias potential V_{bias} ($k_i = 0.06$ Eh and $\alpha_i = 0.15$ Bohr^{-2}). The center of V_{bias} is is marked by "**X**", which corresponds to the global minimum structure shown in **a**. The V_{bias} center appears slightly shifted to the right of the minimum, which is an artifact of the large φ and θ grid of 10° per turn and the color interpolation of the plotting program.

functional group has a large influence on the conformation of citronellal and the globular-folded conformation is stabilized by intramolecular non-covalent interactions of the aldeyhd and C-H groups. The conformational flexibility of citronellal has been experimentally investigated by the group of M. Schnell using chirped-pulse Fourier-transform microwave (CP-FTMW) spectroscopy where 15 gas-phase conformers were identified.[335] Rotational constants contain information about the entire geometry of the molecule and therefore conformational aspects as well. Since

rotational constants are rather method dependant, the CRE is re-optimized at the PBEh-3c (DFT) level to get a good comparison to the experiment. Conformers with small mean deviations of the rotational constants between theory and experiment were visually inspected for matching geometries in the literature.[335] All 15 experimentally identified conformers are found in the iMTD-GC-CRE. The first five of the 15 conformers are shown in Fig. 3.9a.

Post-optimization of the ensemble at a higher level is important as also shown by the fact that the 30th conformer of the initial conformer search at GFN2–xTB level corresponds to the lowest lying conformer of the refinement at PBEh-3c level, highlighting a noticeable re-ranking. The conformers shown in Figure 3.9c are identified by iMTD-GC//GFN2–xTB to be the highest populated and were also predicted by the computational investigation of Schnell *et al.* Interestingly, the conformers CONF1 and CONF2 could not be identified within the experimental spectrum. Experiments with different carrier gases suggested that conformational relaxation towards more stable conformers is facilitated due to collision in the supersonic expansion. The absence of CONF1 and CONF2 can be explained by conformational relaxation and their too high enthalpy after refinement at DFT level.

After demonstrating that relevant conformers were found within the gas-phase CRE, the completeness of the ensemble concerning the populated conformers is investigated by comparing calculated and experimental ^{1}H-NMR spectra in solution. To this end, a new CRE was generated in chloroform and the ^{1}H-NMR spectrum was calculated (see Fig. 3.11) with the procedure detailed in Ref. 40. The spectrum calculated with only one conformer clearly shows that the multiplet splittings of the proton signals are not correct. Only if the whole populated ensemble of 35 conformers is taken into account, the qualitatively correct splittings are obtained. The differences between the experimental and calculated chemical shifts partly stem from the neglect of zero-point effects, vibrational averaging, and errors in the respective DFT calculations. It has to be stressed that in addition to the conformers the identification of rotamers is crucial for calculating NMR spectra. Rotamers are very important since they are necessary to describe the correct averaging of NMR parameters due to the fast interchange of nuclei at the slow time scale of the NMR experiment. The overall good agreement between the calculated and experimental multiplicities indicates that all major conformers were found. This in turn highlights the sophistication and robustness of the iMTD-GC algorithm for generating conformer–rotamer–ensembles (CRE).

3.6.2. Conformations of Macrocyclic Molecules

The treatment of macrocylic rings with chemoinformatic conformational sampling algorithms is challenging since it requires special treatments or heuristic rules.[228] Molecular dynamic-based approaches prove to be useful as their application is straightforward and no additional adjustments are required for these compounds.[240,301,341] In particular, aMD based approaches seem to be promising as the PES modification helps to overcome even high energetic ring-interconversion barriers.[319] Therefore, it can also be expected for the MTD approach to yield similarly good results. To assess the performance of iMTD-GC for macrocycles, a conformational search was

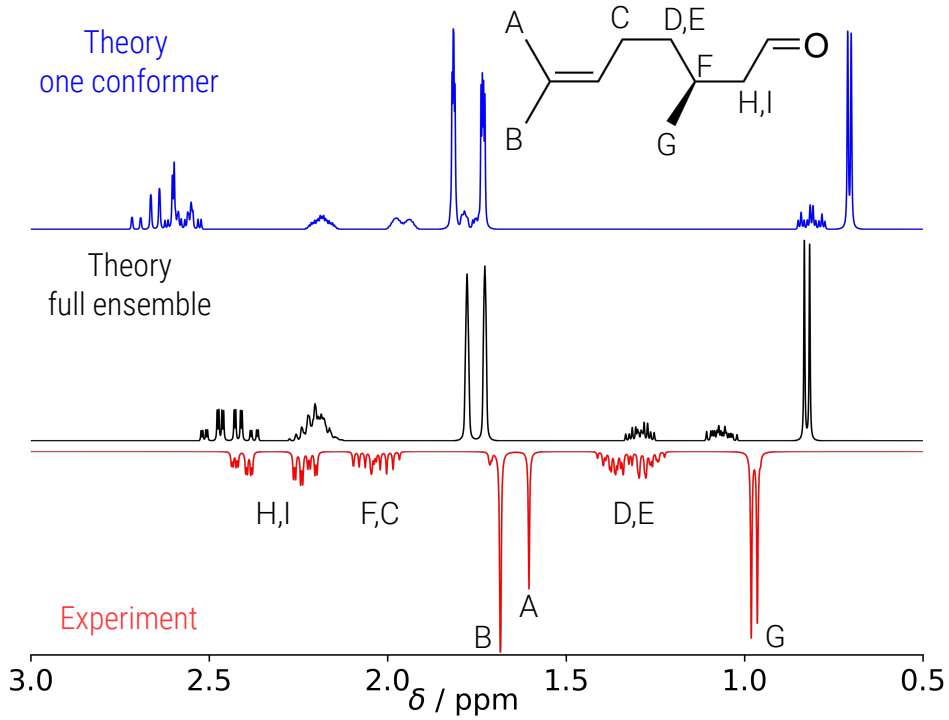

Figure 3.11.: Comparison of calculated and experimental ^{1}H-NMR spectra of citronellal with the focus on the aliphatic region. The upper theoretical spectrum (in blue) is calculated with only one conformer. The theoretical spectrum (in black) better reproduces the experimental spectrum and is calculated with the full CRE consisting of 35 conformers. The lower experimental spectrum[336] (red, inverted) was measured in $CHCl_3$ at 400 MHz, 298 K and the theoretical spectra were calculated with PBE0/def2-TZVP[COSMO][337,338,339]//PBEh-3c[DCOSMO-RS][340] for the coupling and shielding constants. The assignment of the multiplets is indicated by capital letters.

conducted for three different macrocyclic systems taken from Ref. 301. Crystal structure geometries of the three compounds were obtained from the Cambridge Crystal Structure Database (CCSD)[342] with the IDs POXTRD, CAMVES, and CHPSAR. The geometries were optimized in the gas-phase with GFN2–xTB. Their composition and size represent typical organic systems in the target range of the iMTD-GC. The (gas-phase) conformer ensembles can be expected to be sufficiently diverse if: A) a high RMSD to the input (crystal structure) conformation is observed, B) a large number of conformers within the default energy window is obtained, and C) there is a large energetic difference between the input structure and the lowest conformer. Results according to these criteria are depicted in Fig. 3.12.

For all three systems the defined criteria for a diverse ensemble are fulfilled. The smallest macrocycle POXTRD has a huge number of distinct conformers, while the other two examples (CHPSAR, CAMVES) have smaller CREs. By visual inspection of the ensembles of CHPSAR and CAMVES, pairs of hydrogen bonds are identified within the ring-systems. The hydrogen bonding strongly stabilizes a few selected conformational motifs, leading to more compact en-

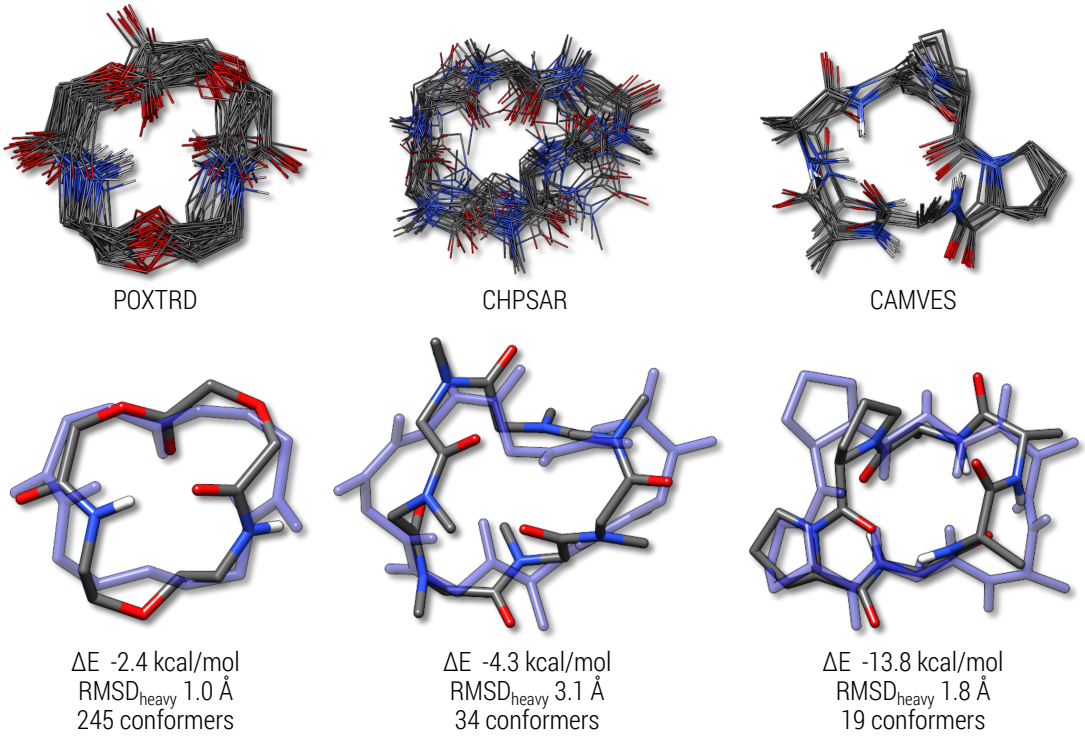

Figure 3.12.: iMTD-GC gas-phase ensembles of three macrocycles are shown. Below overlays between GFN2–xTB optimized geometries from the crystal structure and the lowest lying conformer are presented. For POXTRD only the first 45 conformers of the ensemble are shown. Energy differences (ΔE) between the lowest lying conformer and optimized input geometry are given in kcal mol^{-1}. The RMSD$_{heavy}$ between the lowest lying conformer and input geometry is given in Å.

sembles compared to POXTRD within the energy window of 6 kcal mol^{-1}. The results show that ring-interconversion poses no problem for the iMTD-GC procedure and highlights that different molecular classes can be treated with the same set of search parameters.

3.6.3. Conformations of Ac-Ala$_{19}$-LysH$^+$

As a larger example we chose the protonated peptide Ac-Ala$_{19}$-LysH$^+$, consisting of 20 amino acids (220 atoms). With the size and flexibility of this system we approach the current practical limit of the iMTD-GC workflow conducted at a SQM level, although the application to larger systems would easily be possible at a FF level. From combined theoretical and experimental studies it is known that the conformation of this molecule depends on the protonation site.[343–345] It was found that the protonation at the C-terminus stabilizes an α-helical form of the peptide, which is shown in Fig. 3.13. The alternative N-terminal protonation at lysine destabilizes the helical conformation and leads to more compact structures. Furthermore, the lysine protonated conformations are preferred in the gas-phase, where a single unique conformer was suggested.[345] Hence, the Ac-Ala$_{19}$-LysH$^+$ system is an ideal molecule to evaluate the performance of iMTD-GC for larger systems. Three conformational searches were conducted: starting A) from the

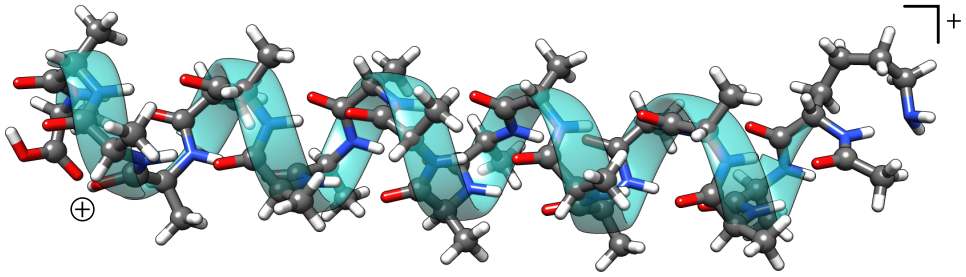

Figure 3.13.: The most stable conformer of Ac-Ala$_{19}$-LysH$^+$ with protonation at the C-terminus. The protonation site is marked by the "⊕"-sign.

α-helical conformer protonated at the C-terminus, B) from a N-protonated helical structure, and C) from the lowest-energy conformation proposed in Ref. 345.
Starting from the C-terminal protonated conformations (case A), the findings in the literature can be confirmed already at the GFN2–xTB level. The entire ensemble consists of 54 α-helical structures that differ only by the orientation of Lys. The most stable conformer is shown in Fig. 3.13. No energetically close lying folded structures were found on the PES where the protonation site is at the C-terminus. During the iMTD-GC search several of these structures are created and then sorted out according to the energy threshold.
In the second conformational search (case B) the most stable helical conformer is taken as an input structure and the protonation site is artificially set to the N-terminal lysine group. With this setup only folded conformations are generated in the ensemble (126 conformers). Here, the most stable conformers are very similar to the "unique" conformer from the literature. For an energy based comparison, the "unique" structure identified as lowest conformer in Ref. 345 is taken as the reference point.

A comparison between the most stable conformer of the iMTD-GC ensemble and the reference structure is shown in Fig. 3.14. The folded conformers contain α- and 3_{10}-helical segments. However, at the GFN2–xTB level the most stable conformers generated by iMTD-GC are up to 1.64 kcal mol^{-1} lower in energy than the reference structure.
When the conformational search is started from the reference structure (case C), a smaller ensemble compared to the previous searches is obtained by iMTD-GC (56 molecules). The energetically lowest conformers of ensemble B and C have similar structures and are up to 1.7 kcal mol^{-1} more stable than the reference. For better comparison between the iMTD-GC structures and the reference all conformations were re-optimized at the PBEh-3c level and single-point energies were calculated at the higher PBE0-D4/def2-TZVPD level.[174] The relative energies for selected conformers are plotted in Fig. 3.15. Although the obtained conformers from the second conformational search are very similar to the reference structure and are favored by a few kcal mol^{-1} at the GFN2–xTB level, at the hybrid DFT level the reference structure is still preferred. However, by conducting the third conformational sampling (case C), the literature ensemble could be extended. At the PBE0-D4/def2-TZVPD level, there are three conformers in the new ensemble which are energetically lower (by 0.33 to 0.37 kcal mol^{-1}) than the reference.

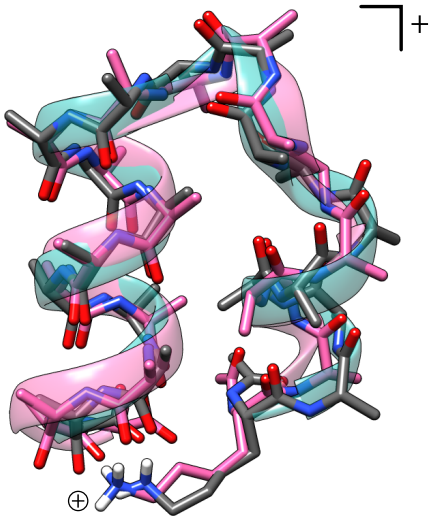

Figure 3.14.: Overlay between the most stable conformer of Ac-Ala$_{19}$-LysH$^+$ taken from literature (magenta) and the highest populated conformer of ensemble B generated by iMTD-GC (gray/green). The N-terminal protonation site is marked by the "⊕"-sign.

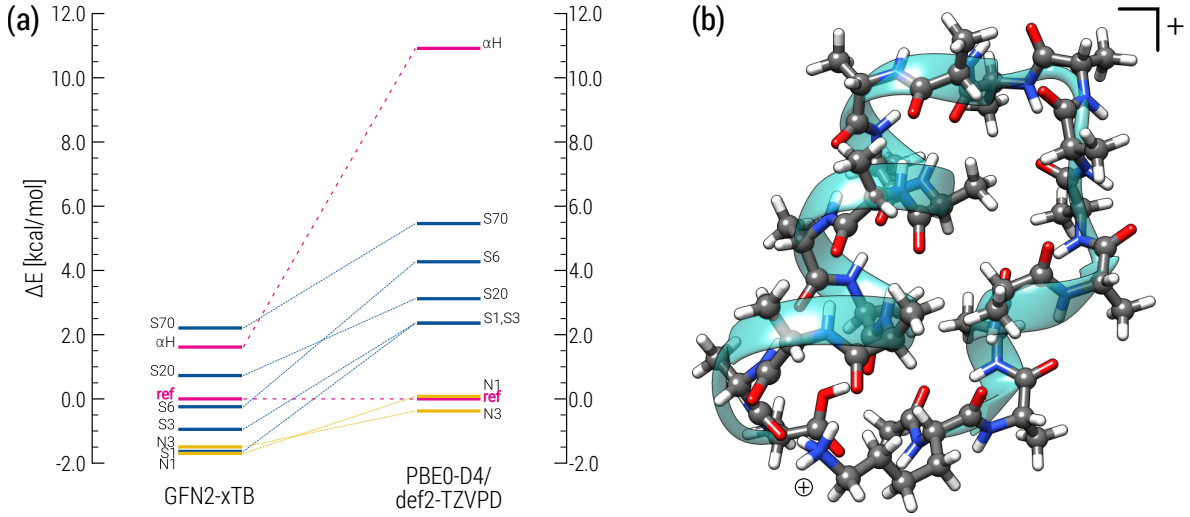

Figure 3.15.: **(a)** Comparison of relative energies for selected conformers at the GFN2–xTB and PBE0-D4/def2-TZVPD//PBEh-3c levels. Conformers labeled by "S" were generated starting from the N-protonated helical conformer (case B). Conformers labeled by "N" were generated starting from the reference structure "ref" (case C). αH is the most stable α-helical conformer. **(b)** Overall most stable iMTD-GC conformer at DFT level of Ac-Ala$_{19}$-LysH$^+$, labeled N3.

Furthermore, there are four other conformers which are less than 0.1 kcal mol^{-1} above the reference and thus would be significantly populated. Compared to the PBE0-D4 results, most conformations at the GFN2–xTB level are energetically close lying. This indicates a much flatter PES at the SQM level compared with the DFT level. Since energies of up to 0.4 kcal mol^{-1} are well within the error margin of conformational energies at a hybrid DFT level, it is not possible

to determine if these new conformers are truly the most stable structures of Ac-Ala_{19}-$LysH^+$. However, the ensemble indicates that there is not just one unique populated conformer of the peptide in the gas-phase but rather an ensemble of several energetically and structurally close conformations that are protonated at the lysine terminus. Furthermore, the example shows that qualitative results can be obtained with the iMTD-GC//GFN2–xTB method at comparatively low computational cost. In the original study several first-principles simulations based on replica-exchange molecular dynamics (REMD)[346–349] at the DFT (PBE) level[128] were performed in order to find the low energy conformations.[345] This approach is extremely expensive, even without including PBE0 single-point calculations. In contrast, the iMTD-GC//GFN2–xTB sampling was conducted within a couple of days on a single workstation and yielded similar low-energy conformations.

3.6.4. Conformers of Metal-Organic Systems

Metal-organic compounds can be routinely calculated with DFT but the computational cost of applying the iMTD-GC procedure with DFT as underlying electronic structure method is high as already mentioned. Combining the GFNn–xTB SQM methods with the iMTD-GC algorithm has the advantage of low computational cost and enables treatment of organometallic systems. In fact our approach is the only routinely available for that purpose on the market. The possibilities are demonstrated for two metal-organic examples in the following chapter.

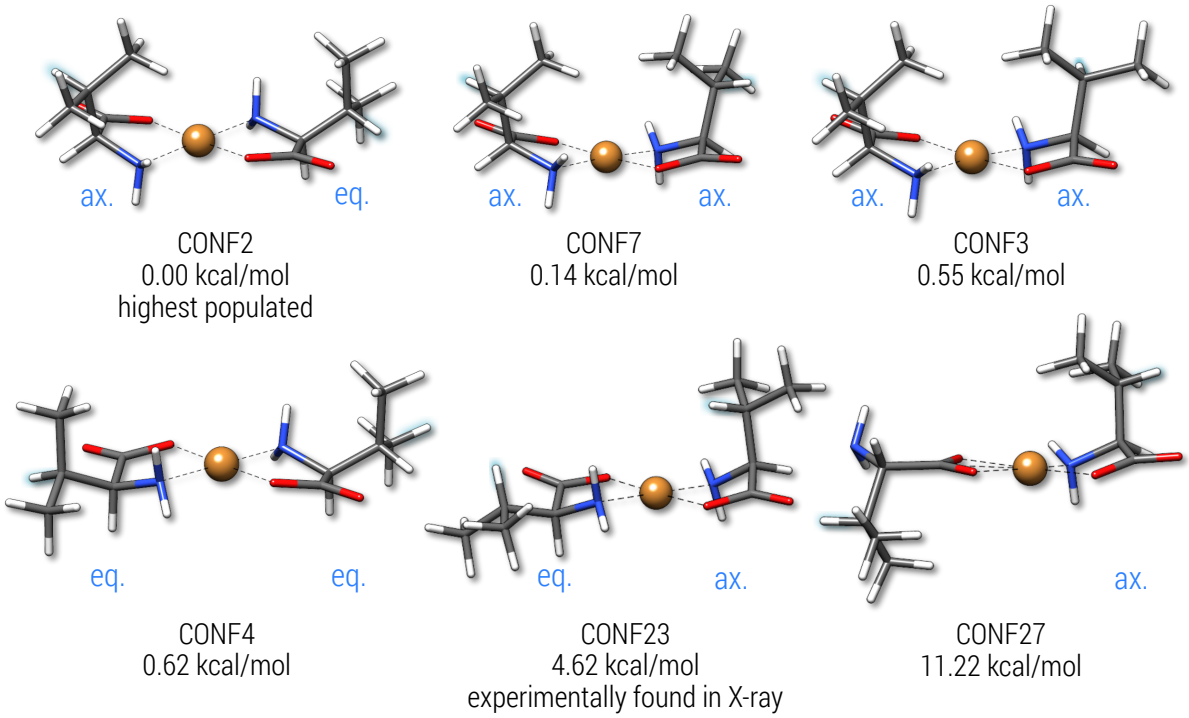

Figure 3.16.: Conformers of *trans*-Cu^{II}(L-Valine)$_2$ optimized in gas-phase at B97-3c[180] level of theory. The free energies given below the respective conformer name are in kcal mol^{-1} (@ 298.15 K).

The first example is *trans*-Cu^{II}(L-Valine)$_2$. Each chelate ring can have an axial or equatorial conformation of the isopropyl-group. The valine residue can exhibit various conformations, identifiable by comparing the highlighted methine-proton position in Fig. 3.16, where six representative conformers are depicted. The generated structures were refined at the B97-3c level, where the first four conformers are populated and conformers CONF23 and CONF27 are higher lying conformations. Conformer CONF23 is not populated in the gas-phase but closely resembles the experimentally observed conformation in the crystal structure.[350] This indicates its stabilization in the solid phase due to packing (or other related) effects.
As a second example, the κ^2-tris(pyridyl)methanol-diacetylplatinum(II) complex was investigated. The chelate complex was studied in implicit methanol and the ensemble consists of 68

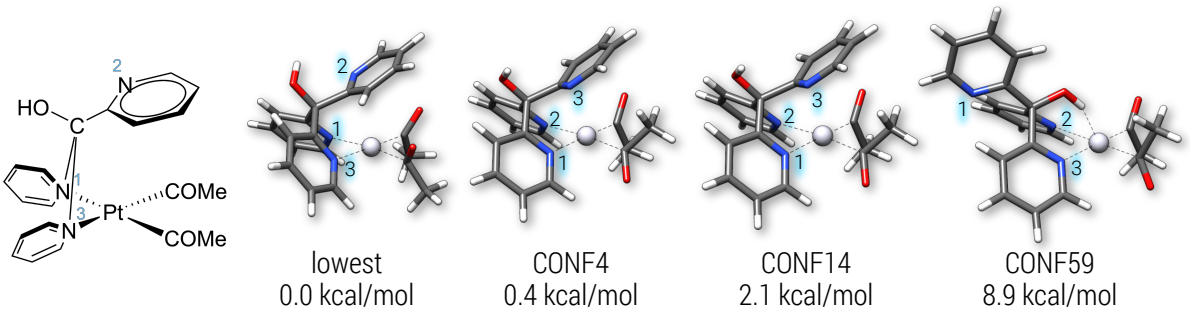

Figure 3.17.: Lewis structure of [Pt(COMe)$_2$(2-py)$_3$COH] and selected lowest and higher lying conformers calculated at GFN2–xTB[GBSA(MeOH)] (GBSA = Generalized Born (GB) with solvent accessible surface area(SA)) level of theory (ΔE in kcal mol^{-1}) are shown. The numbering at the nitrogen atoms is aiding in the distinction of the coordination centers.

conformers within an energy window of 10 kcal mol^{-1}. The tris(pyridiyl)methanol ligand forms a κ^2-chelate complex with the Pt(II)-ion and has a free coordination site at the third uncoordinated pyridyl group. In Fig. 3.17, a selection of conformers is shown, highlighting the flexibility and dynamic processes that may occur in solution. The acetyl groups (COMe) are almost freely rotatable and the uncoordinated pyridyl ring easily can rotate as well. The numbering at the coordinating pyridinyl-nitrogens, illustrates that the chelate complex can open up and re-coordinate to the free center. This is particularly visible in conformer 59, where the hydroxy group coordinates to the Pt-ion and the pyridyl-group is facing away from the metal center. The conformer search gives valuable insight into the flexibility of this Pt(II)-chelate-complex, which has been confirmed experimentally.[351] The wall computation time for this calculation was 12 minutes on 40 cores (Intel Xeon Gold 6148 CPU @ 2.4 GHz).

3.7. Specialized Applications

The general setup of the conformational search algorithm as a combination of molecular dynamics simulations and quantum chemical structure optimization enables several specialized

applications. Atoms can be constrained in the input structure or removed from the bias potential. Furthermore, other potentials than V_{bias} can be included in the calculations.

3.7.1. Constrained Conformational Sampling

The first specialized application of the CREST program is the constrained conformational sampling. Single atoms or parts of the molecular structure can be fixed and retain their geometry during the calculations. This makes it possible to, *e.g.*, screen for conformational changes only occurring in some domains of a molecule. Atoms that are constrained must, however, not appear in the bias, since this would counteract the constraining potential.

Tyrosine Conformation on a Graphene Surface

Bio-sensing of α-amino acids using nanomaterials is a vital research for which a detailed knowledge of the amino acid conformations at the material-interface is essential.[352,353] The conformational search of L-tyrosine at a model graphene surface is demonstrated here. The graphene sheet consists of 216 carbon atoms and has D_{6h} symmetry. For the conformational search with CREST, the graphene layer is constrained and all graphene atoms are removed from the RMSD criterion. Otherwise, this would lead to the dissociation of the complex and a strong deformation of the graphene monolayer. Fig. 3.18a shows the lowest lying L-tyrosine conformation found at the graphene surface.

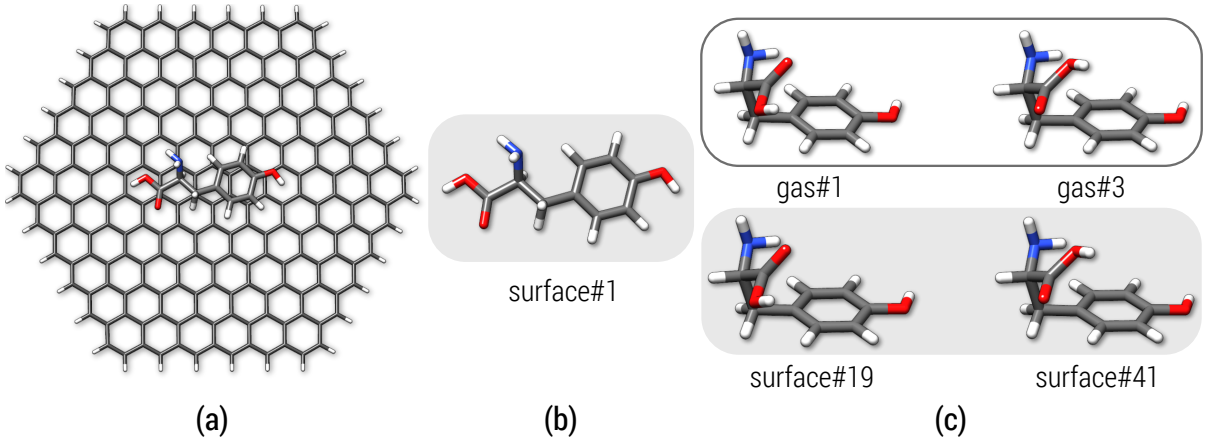

Figure 3.18.: Tyrosine conformations on a model graphene surface and in the gas-phase. **(a)** Most stable tyrosine conformer on the graphene surface. **(b)** Most stable tyrosine conformer depicted without graphene. **(c)** Comparison between gas-phase and interface conformers.

On the surface 108, different tyrosine conformers were found. All low lying conformers show the π-π interaction motif via parallel alignment of the phenyl-ring to the graph sheet. Higher-energy conformations bend away from the parallel phenyl-ring arrangement. Other observed interaction motifs are C(α/β)-H...π and N-H/O-H...π interactions. The interaction energy (E_{int})

for the most stable conformer and the graphene mono-layer amounts to -2.8 kcal mol^{-1} at the GFN2–xTB level. To highlight the difference between a constrained conformational search at an interface and in the gas-phase, tyrosine conformers were created in the gas-phase without any constraints. Here, 30 conformers were found. Comparing the geometries taken from the graphene-based ensemble and the gas-phase it is seen, that conformations are created at the graphene interface, which are not present in the gas-phase ensemble.

As depicted in Fig. 3.18b and 3.18c some low lying gas-phase conformations can be found on the graphene surface, but they are not the lowest populated conformers, *e.g.*, the lowest gas-phase tyrosine conformer is equal to the 19th conformer from the graphene-tyrosine ensemble and is 1.3 kcal mol^{-1} higher in energy than the lowest graphene-tyrosine conformer. The graphene potential clearly influences the tyrosine conformations and hence it is essential to create conformations in their genuine environment.

Conformers of Transition-States

Studying reaction kinetics is usually done by investigating activation energies of rate determining steps. Reactions are analyzed with the assumption that the reaction kinetics are termed by the transition-state (TS) free energy, relative to the free energy of the reactants. It is assumed that the reaction proceeds through the TS, which is linked to the energetically most favorable substrate. If the substrate can interconvert quickly between its low energy conformations, the reaction is governed by the Curtin–Hammett principle.[219,354] In this case, the reaction proceeds through the lowest TS which is not necessarily connected to the lowest lying substrate conformer. Hence, for accurate results it may be necessary to search the chemical reaction space and find low-lying TS conformations. Constraint conformer searches can be applied to a previously found TS. The procedure is demonstrated for the enzyme COMT (catechol-*O*-*m*ethyl transferase), which catalyzes the methyl group transfer from *S*-adenosyl-L-methionine (SAM) to a catechol ion[355,356] (see Fig. 3.19a for the Lewis structure of the TS). First, the basic S_N2 reaction was modeled by taking the active site of the enzyme and saturating all capped bonds with methyl groups. A TS guess was optimized and checked for the reaction mode by performing a harmonic frequency calculation. To preserve the TS vibrational mode in the conformational search, the atoms dominantly contributing to this mode were constrained. In this system, the catechol oxygen, the carbon of the transferred methyl group, and the sulfur of the SAM group were fixed. To retain the magnesium-cation coordination the Mg-ligand distances were constrained as well (only one constraint per ligand). Additionally, the water and the amid ligand were constrained to an O_{H_2O}-Mg-O_{Amide} angle of 180°. For the TS conformational search only the constraining of the breaking and forming bonds is necessary and all other constraints are used to keep the active site of the enzyme intact. The iMTD-GC approach generates 141 conformers within 6 kcal mol^{-1}. Overall, the calculation takes 46 minutes on 40 cores (Intel Xeon Gold 6148 CPU @ 2.4 GHz). The conformers are good estimates for the further optimization into the TS at GFN2–xTB[GBSA(MeOH)] level. Here, 138 true TS are obtained. During the optimization of the TS, geometries can converge into the same TS geometry and have to be sorted out.

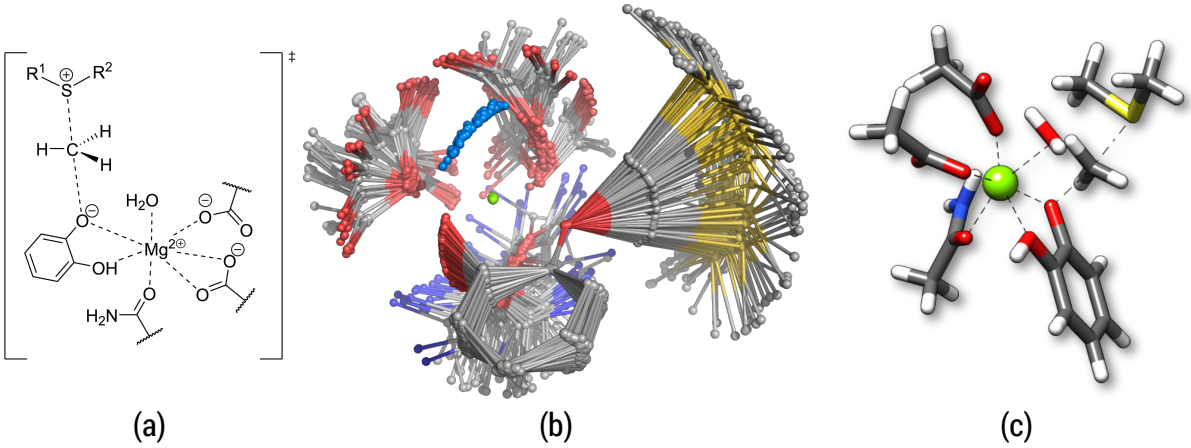

Figure 3.19.: TS of the active site of the COMT enzyme. **(a)** Lewis structure of the S_N2 methyl-group transfer reaction. **(b)** TS ensemble of 91 TS which are optimized after the iMTD-GC conformer search (atoms in light blue depict the water oxygen, all hydrogen atoms are omitted for clarity). **(c)** Lowest lying TS at GFN2–xTB[GBSA(MeOH)] level.

After sorting, 91 unique conformers within 6.1 kcal mol^{-1} remain. Overlays of the optimized TS ensemble and the lowest lying TS are depicted in Fig. 3.19b and 3.19c. The procedure provides a semi-automated approach of finding lower lying TS by relaxing the ligand geometries.

For the COMT example the lowest-lying TS conformer is 13.5 kcal mol^{-1} lower in energy than the initial TS. Comparing the barrier from the lowest substrate conformer and the lowest TS-conformation with the reaction barrier from the single structure approach (see Fig. 3.20) the barrier height is reduced by 8.7 kcal mol^{-1} when using the ensemble based protocol. Although the work was conducted at a semiempirical level it should be stressed that the procedure is generally applicable and that it is also possible to refine the TS at DFT-level.

3.7.2. Aggregate Sampling

With the MTD based approach it is possible to also screen for different conformations of non-covalently bound aggregates and complexes (NCI–iMTD). This is a special run-type mode in which an ellipsoidal shaped potential is added as a constraint to the MTD simulations. The additional potential is necessary to avoid the dissociation of the non-covalently bound complexes. However, to obtain unbiased conformations and aggregate structures the ellipsoidal constraint is removed during the geometry optimization. The energy contribution E_{pot} given by the ellipsoid potential is defined as

$$E_{pot} = \sum_{i}^{N} \left(\frac{|\mathbf{R}_i - \mathbf{O}|}{R_{i,pot}} \right)^{10} , \tag{3.6}$$

where the summation runs over all atoms N. $\mathbf{R}_i$ are the Cartesian coordinates of atom i, $\mathbf{O}$ is the center of the potential (*i.e.*, the origin), and $R_{i,pot}$ is the radius of the potential parallel

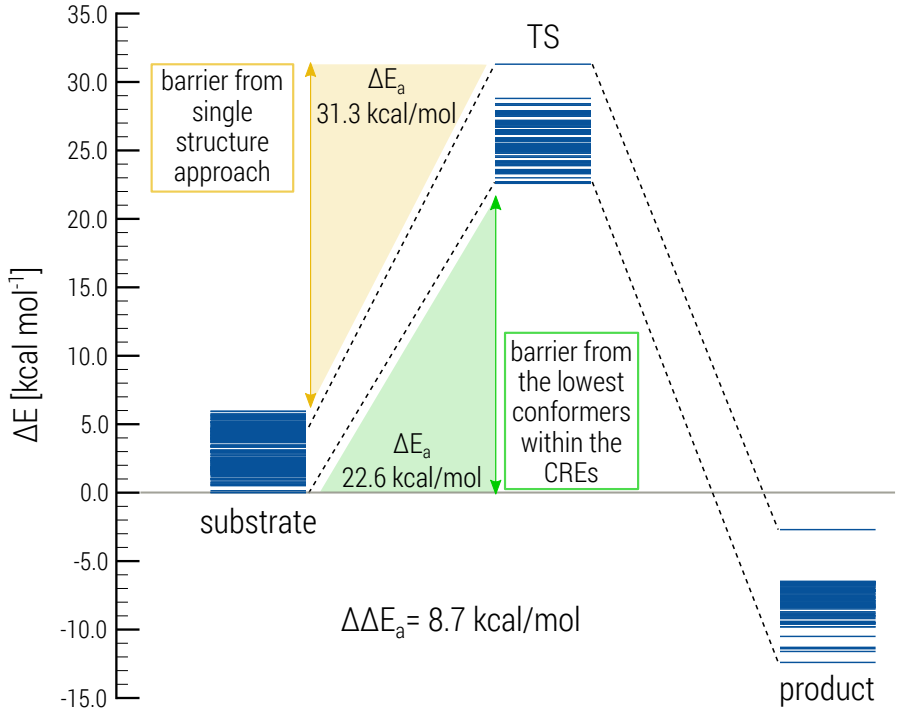

Figure 3.20.: Reaction barriers of the S_N2 methyl-group transfer. The reaction barrier from the single structure approach (yellow) and barrier from the lowest conformers of the CREs (green) are compared. Energies of the conformers relative to the lowest substrate conformation are illustrated by horizontal blue lines. All energies were calculated at the GFN2–xTB[GBSA(MeOH)] level.

to $\mathbf{R}_i - \mathbf{O}$. If such a potential is combined with a strong RMSD bias, chemical reactions can be enforced, resulting in a mode similar to the nano-reactor presented in Ref. 41. However, the aim of the NCI–iMTD procedure is to generate a low-energy ensemble for which the parameters are adjusted accordingly. In the following two examples demonstrate the application of the NCI–iMTD procedure.

Water Hexamer

Since each non-covalently interacting fragment can also have different conformations on its own even for small systems a large number of complexes is possible. In the general case one would search first with iMTD-GC the monomer conformations before NCI–iMTD is started. Hence, as a primary example small molecular clusters such as $(H_2O)_6$ are used here, where fragment conformations are irrelevant. In the first step the ellipsoidal potential is automatically generated from scaled principal rotation axes. For the water hexamer the resulting potential is schematically shown in Fig. 3.21. Within a 6 kcal mol^{-1} energy window 69 different gas-phase aggregates are found for $(H_2O)_6$. Many of the generated structures are isomers that differ only by the direction of their hydrogen bonds and otherwise show a similar geometry. Hence, only a selection of six noteworthy structures is shown in Fig. 3.22. The structures shown in Fig. 3.22a - 3.22d

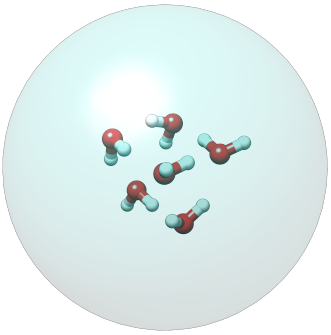

Figure 3.21.: Visualization of the almost spherical ellipsoidal potential around a $(H_2O)_6$ cluster.

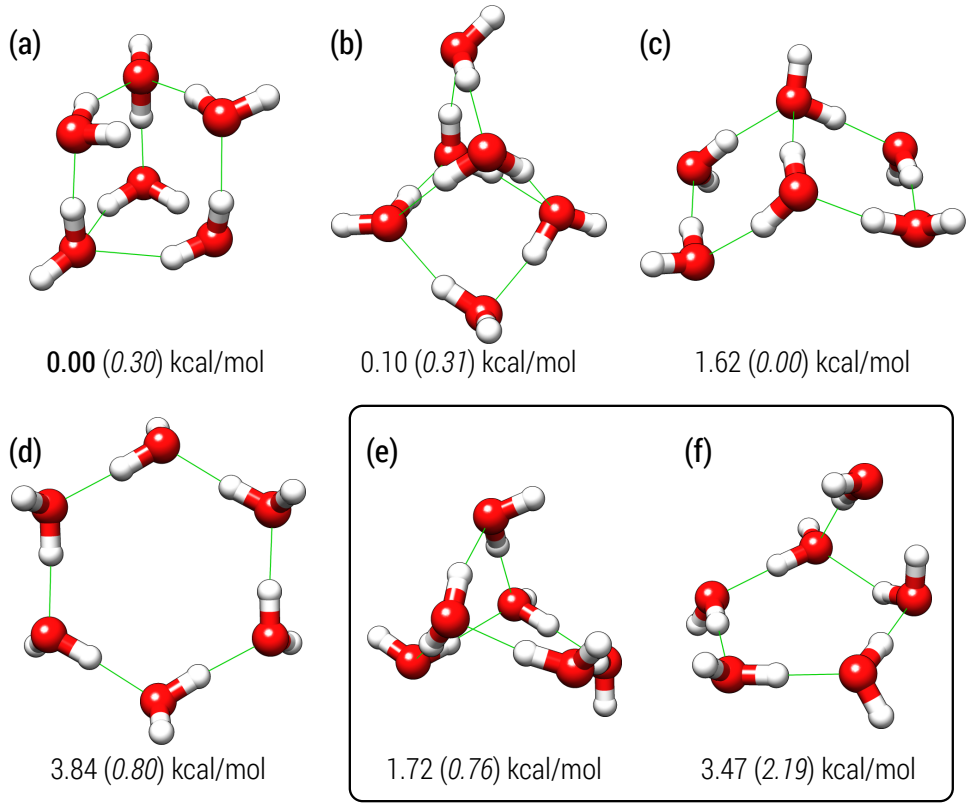

Figure 3.22.: Six different aggregates that are automatically generated for $(H_2O)_6$. Relative energies (ΔE) at the GFN2–xTB level are given below the respective structures. Energy values in parenthesis show relative energies at the PBEh-3c level.

are well-known from literature (*i.e.*, the prism-, cage-, book-, and cyclic-hexamer respectively) and are for example included in the WATER27 subset of the GMTKN55 database.[109,357] The two aggregates in Fig. 3.22e and Fig. 3.22f, which were chosen randomly from the ensemble, additionally highlight the structural diversity created by the NCI–iMTD procedure.

1-Naphtol Dimer

The second example is the 1-naphthol dimer, where various π–π stacking and hydrogen bonding motifs are possible. In the literature seven conformations were proposed based on a comparison of the experimental and theoretical rotational constants.[358] With the default NCI–iMTD mode, 88 unique aggregates of the 1-naphthol dimer are found at the GFN2–xTB level of theory within a 6 kcal mol^{-1} energy window. To figure out whether the seven proposed complexes are included in this ensemble, structures were pre-selected based on low mean relative deviations (MRD) of the calculated GFN2–xTB and experimental rotational constants from Ref. 358. Afterwards, these structures were visually inspected. The reference rotational constants were obtained for geometries calculated at B3LYP-D3(BJ)/6-31++g(d,p) level of theory taken from the original publication (see Ref. 358). For all seven complexes a corresponding (or at least closely related) structure was found with the NCI–iMTD mode at the GFN2–xTB level of theory. The geometries are shown in Fig. 3.23. From benchmark studies it is known that rotational constants for

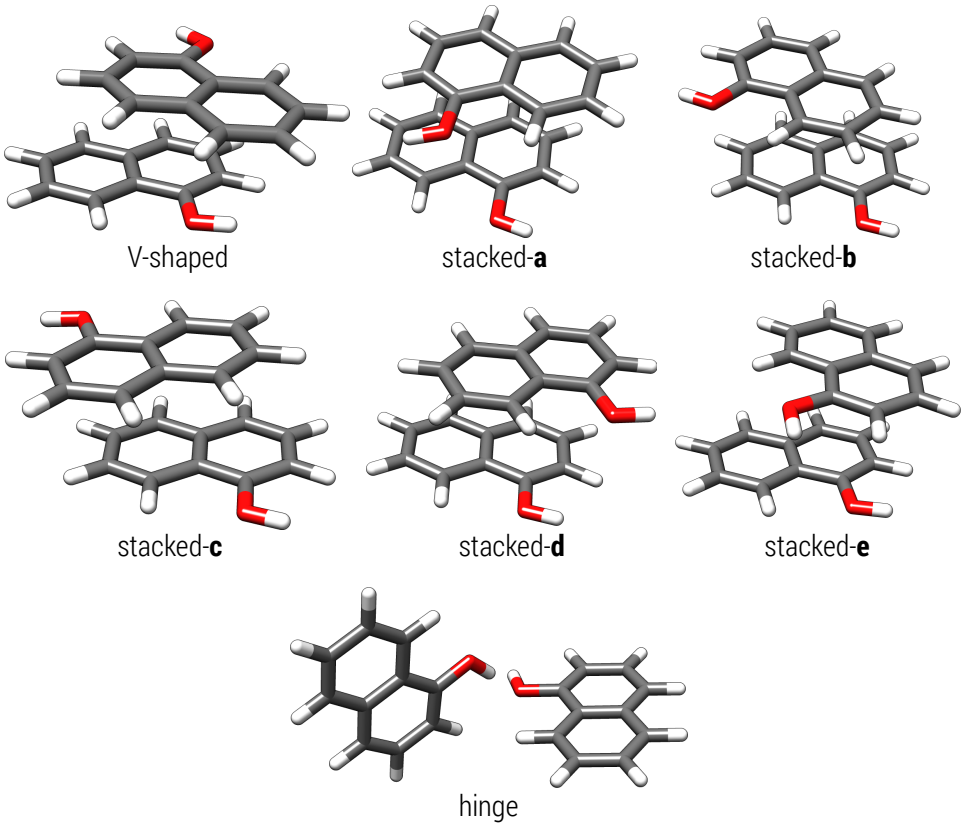

Figure 3.23.: Seven 1-naphthol dimer aggregates that were generated at the GFN2–xTB level of theory and match the structures proposed in the literature.

geometries calculated at a semiempirical level have larger relative deviations compared with geometries calculated by DFT.[359] Therefore, the GFN2–xTB geometries are not expected to reproduce the hybrid DFT rotational constants very well but rather serve as a guideline for the structure evaluation. The rotational constants for the seven predicted complexes of the

1-naphthol dimer are given in Tab. 3.3. The rotational constants calculated for the GFN2–xTB geometries deviate from the reference by 4–5 % on average. Many more aggregates are generated,

Table 3.3.: Rotational constants B_e for the seven predicted aggregates shown in Fig. 3.23.

	calc.[a] B_e/MHz			ref.[b] B_e/MHz			
system	A	B	C	A	B	C	MRD
V-shaped	491.5	282.3	271.0	468.2	284.6	257.9	3.6 %
stacked-a	450.0	317.7	291.9	449.6	301.2	286.2	2.5 %
stacked-b	457.0	307.0	295.2	479.0	273.5	272.2	8.4 %
stacked-c	479.3	290.9	282.7	454.9	298.6	286.0	3.0 %
stacked-d	466.3	300.3	291.8	480.4	276.0	270.8	6.5 %
stacked-e	428.1	333.1	281.9	421.0	326.5	270.7	2.6 %
hinge	609.7	125.0	120.8	594.1	131.3	125.6	3.8 %
exp. (B_0)	—	—	—	462.4	275.9	252.5	—

[a] Calculated for GFN2–xTB geometries. [b] Taken from Ref. 358.

and none of the seven isomers is the most energetically favored complex of the ensemble at the semiempirical level. However, the trend in the aggregate stability identified by Jäger *et al.*,[358] that the V-shaped forms are more stable compared to the stacked and hinge forms is reproduced for the GFN2–xTB ensemble. Comparing directly to the experimental rotational constants (see Tab. 3.4), a total of five structures from the NCI–iMTD ensemble fit to the experimental values. All five geometries have rotational constants matching the experiment but a clear identification based only on the very similar rotational constants is not possible. Four of these structures are V-shaped homologues. In fact, the most stable of them is the V-shaped complex that was identified by the authors Ref. 358 and is shown in Fig. 3.23. Only one of the structures is not V-shaped but appears to be a more symmetric form of the stacked **b**-complex. The predicted complexes (except the already known V-shaped form) are shown in Fig. 3.24. The

Table 3.4.: Rotational constants B_e for the four new predicted 1-napthol aggregates shown in Fig. 3.24.

	B_e/MHz			
system	A	B	C	MRD
exp.[a] (B_0)	462.4	275.9	252.5	—
new-1	456.9	303.2	266.5	5.5 %
new-2	484.6	288.7	267.7	5.2 %
new-3	502.6	276.1	261.9	4.1 %
new-4	486.1	287.6	269.3	5.3 %

[a] Taken from Ref. 358.

new generated aggregates support the original conclusion that the true conformation of the 1-naphthol dimer is a V-shaped type complex.

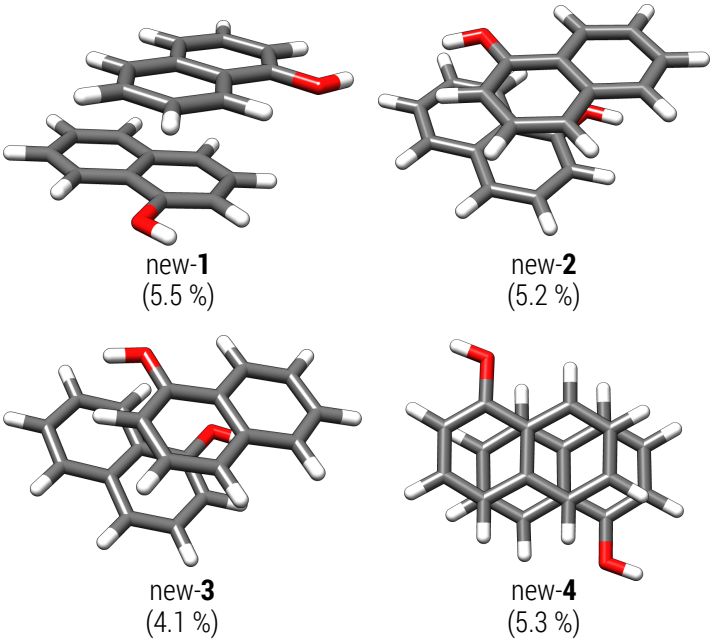

Figure 3.24.: Four new aggregates are predicted for the 1-naphthol dimer generated at the GFN2–xTB level of theory. The MRD of the calculated rotational constant from the respective experimental value is given in parenthesis.

3.7.3. Automated Protonation/Cationization

The procedure for the automated protonation was already discussed in Ref. 220 and is only briefly outlined here. The computational protocol has since been optimized and was implemented into CREST. The workflow is similar to the concept used within the chemical reaction networks,[274,276] where the reactive sites are determined first and then a reactive species is placed in close vicinity during a geometry optimization. Here, reactive sites are π– and lone-pair (LP)–centers that are obtained from localized molecular orbitals (LMOs) and the reactive species is a proton (H^+). The geometry optimization leads to a set of different protomers. Sorting the protomer ensemble is based on relative proton affinities, in the same way that the CRE sorting depends on conformational energies. GFNn–xTB is able to describe these proton affinities sufficiently accurate.[220] The procedure is schematically outlined in Fig. 3.1 and in Fig. 3.25 for the benzocaine molecule, which is a prominent example for protomers in the literature.[360–363] In the gas-phase, the O-protonated benzocaine molecule is favored over the N-protonated species (at GFN2–xTB level). All possible protonation sites in the aromatic ring are also obtained with the automated procedure but are still not populated. However, if the calculation is performed with implicit water solvation, both the N- and O-protonated species are obtained and populated. This corresponds to the experimental finding that N-protonation can (only) occur in the gas-phase under the influence of microsolvation.[363] In a modified version of the procedure also the cationization of molecules, for example with a sodium ion Na^+, is possible. The setup is the same as for the protonation site screening, but the corresponding ion is placed at the π– or LP–center, instead of a proton. An example from the literature[364] is the cationization of adeno-

sine to $[Ade+Na]^+$. At the GFN2–xTB level ten different cationized structures are generated within a conservative 30 kcal mol^{-1} energy window. The two energetically lowest structures are found within 10 kcal mol^{-1} and agree with the predicted relative stabilities in the literature.[364] Also reasonable higher-energetic coordination motives are generated by the automatized screening procedure. Gas-phase structures of the two lowest and one higher energetic $[Ade+Na]^+$ complex are given in Fig. 3.26.

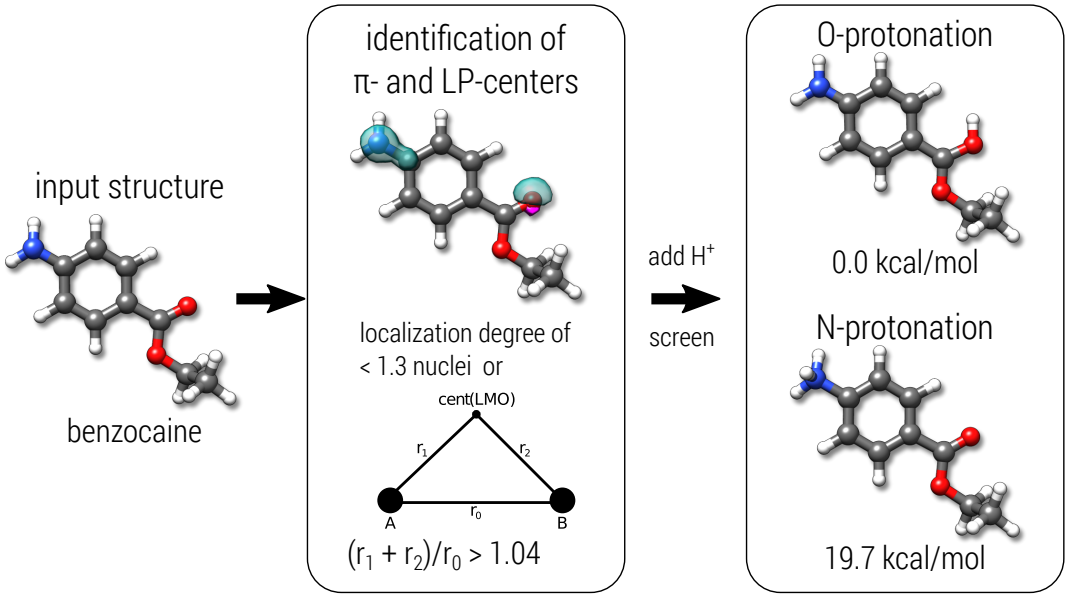

Figure 3.25.: Schematic procedure of an automatized protonation of the benzocaine molecule.

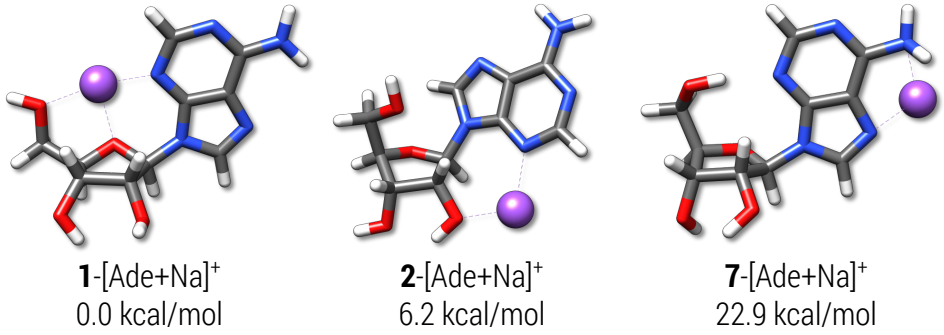

Figure 3.26.: Three of overall ten $[Ade+Na]^+$ structures generated at the GFN2–xTB level within a 30 kcal mol^{-1} energy window. Relative energies at GFN2–xTB level are given below the corresponding structure.

The preferred coordination sites found, *i.e.*, **1**-$[Ade+Na]^+$ and **2**-$[Ade+Na]^+$ in Fig. 3.26, are correct and were already discussed in the literature.[364] By default the screening procedure only yields topologically unique structures without specific attention being paid to the conformation and hence may not lead to the global minimum. Small conformational differences can, *e.g.*, result from a different conformation of the input structure (here adenosine). Additional con-

formational searches for the products were not performed but in real applications should be conducted.
The trivial counterpart to the automated protonation is the automated deprotonation. Since the screening only involves the removal of protons to generate input deprotomers and the results are comparable to the automated protonation, it will not be discussed here. Overall the protonation/cationization procedure provides an automated approach of finding relevant protomers and coordination sites which can be used in further computational studies.

3.7.4. Automatized Tautomerization and Isomerization

Tautomerism is a widespread phenomenon that influences the chemistry of molecules with readily interchangeable isomers. In the most common type of this inter-conversion the isomers only differ in the position of a proton, which is called prototropy.[354] Knowledge about a molecules possible tautomeric behavior is of great importance, since the isomers can strongly differ in their physical and chemical properties.[253,365] It is also highly important in structural databases and hence, many chemoinformatic approaches exist for the identification of tautomers.[254,366,367]
The automated quantum chemical tautomerization (or isomerization) is a standalone feature implemented in CREST. The protocol involves a sequence of protonating and deprotonating steps and was first applied in Ref. 221 for the calculation of pK_a values in water. If larger

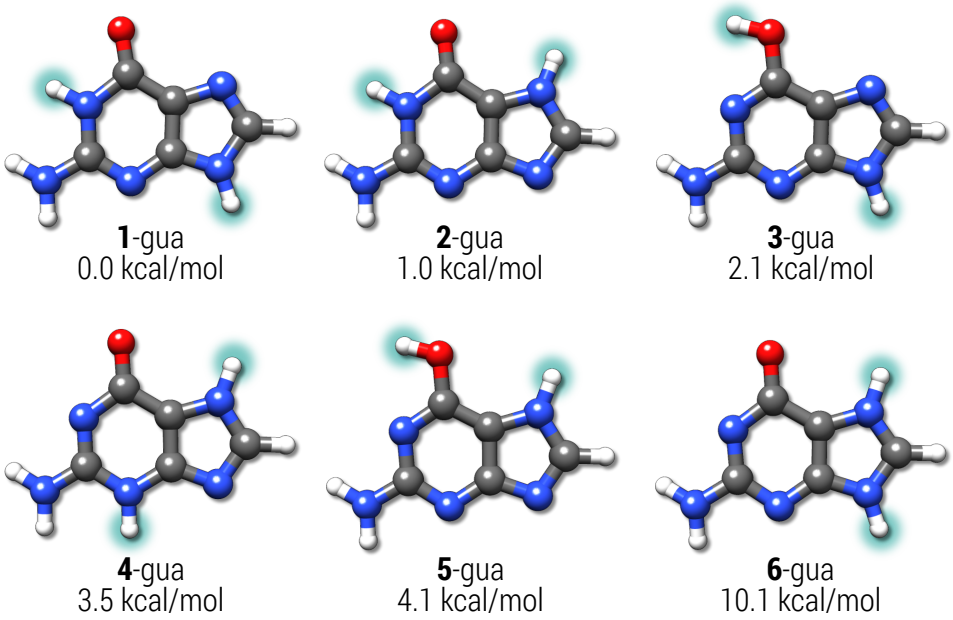

Figure 3.27.: Six lowest prototropic tautomers of guanine generated at the GFN2–xTB[GBSA(H_2O)] level within a 30 kcal mol^{-1} energy window. Relative energies at the GFN2–xTB[GBSA(H_2O)] level are given below the corresponding structure. Proton positions that are changed during the tautomerization are highlighted.

topological changes are induced by this procedure, the obtained structures have to be considered structural isomers instead of prototropic tautomers. A starting geometry is protonated

as described in section 3.7.3 and each resulting protomer is deprotonated at every position. With this procedure all stable structures containing a single permutation of proton positions with respect to the input are obtained. Typically, these "first order" prototropic tautomers are already a good estimate of a molecules' tautomerism. However, even for simple molecules such as guanine shown in Fig. 3.27, further permutation of protonation sites can lead to additional tautomers. Hence, the screening procedure is an iterative sequence of protonation and deprotonation which is performed twice or more. Figure 3.27 shows automatically generated low-energy tautomers of guanine. The tautomerization was started from the keto form of guanine, dubbed **1**-gua in Fig. 3.27. In the first protonation/deprotonation iteration, the tautomers **2**-gua, **3**-gua and **6**-gua are obtained, which differ only by the position of a single hydrogen atom (relative to **1**-gua). Further iterations also yield the structures **4**-gua and **5**-gua with two permutated hydrogen positions. Typically, two iterations of protonation and deprotonation are sufficient to recover relevant low-energy tautomers. In case of the guanine ensemble at the GFN2–xTB[GBSA(H_2O)] level all experimentally known low-energy tautomers (Fig. 3.27) as well as higher-energetic structures discussed in the literature were recovered.[368–372]
As already mentioned the QM based protocol enables its application also to metal containing molecules. In terms of tautomerism this includes the typical prototropic case (*e.g.*, at the ligands), as well as less common phenomena such as proton-hydride tautomerism. An example for the latter is the tautomerism of $[(Cp^*)Rh(bpy)H]^+$ as shown in Fig. 3.28, which is part of an experimentally suggested catalytic cycle.[373] The GBSA implicit solvation model for acetonitrile was employed to resemble the experimental conditions. Within a 10 kcal mol^{-1} window there are

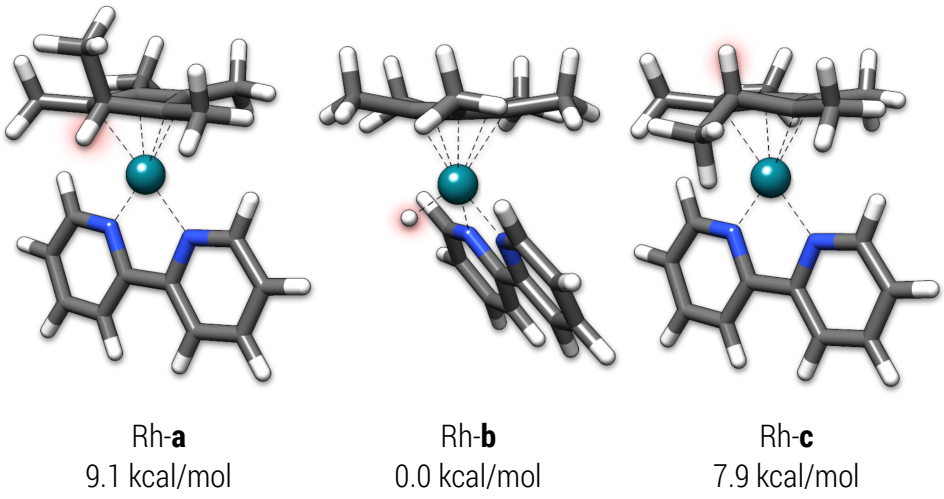

Figure 3.28.: Selected proton-hydride tautomers of $[(Cp^*)Rh(bpy)H]^+$ obtained automatically by using Rh-**a** as an input. Relative energies at the GFN2–xTB[GBSA(MeCN)] level are given below the structures.

three distinct proton-hydride tautomers of $[(Cp^*)Rh(bpy)H]^+$ at the GFN2–xTB[GBSA(MeCN)] level (some broken structures were discarded). Rh-**a** is taken as input, where the proton is bound to the Cp* ligand, facing downwards to the rhodium atom. The procedure automatically recov-

ers the hydride structure Rh-**b** $[(Cp^*)RhH(bpy)]^+$, which is in fact the most stable tautomer and another prototropic tautomer Rh-**c**, in which also the Cp* ligand is protonated, but the proton is facing away from the metal center. While Rh-**c** is approximately 1.2 kcal mol^{-1} more stable than Rh-**a**, only the latter is a reactive species in the catalytic cycle.[373] However, since there are reactive centers (LMOs) also on the up-facing side of Cp*, the procedure generates Rh-**c**.

In general, the automated sequence of protonation/deprotonation allows to access *any* tautomer that converges to a local minimum during the geometry optimization. The reason for this is that no barriers for the addition or secession of protons are taken into account and hence thermodynamically unfavorable (*i.e.*, high energetic) isomers can be produced. Therefore, the procedure recovers typically not only prototropic tautomers but also other structural isomers. Due to the exploratory nature of the approach it is currently up to the user to decide which of the generated structures shall be considered as tautomers and which as structural isomers. The exploration has, however, the advantage that more complicated types of tautomerism, such as ring-chain-tautomers, are often found by the same workflow. In contrast, many chemoinformatic tautomerization tools require additional heuristics for the treatment of ring-chain-tautomerism.[253,254]

3.8. Troubleshooting

The algorithms used in CREST and `xtb` are physically very plausible and have been implemented in a way to provide robust simulations under various conditions. Nevertheless, also because they can be applied to almost any system composed of all common elements from the periodic table, seeming discrepancies of the results to corresponding experiments or expectations may occur. Assuming that they are not rooted in technical problems and a proper CREST ensemble file for the chemically correct system has been written, we here want to discuss briefly common error sources.

First, one should check if the real and simulated systems are close to each other. The mostly applied continuum solvation models usually give good results but may fail for very polar or ionic situations. The conformation/protomer/tautomer ensembles in the solid, liquid (solution), or gaseous state may strongly differ from each other due to packing or solvation effects. One should not expect in general that the conformation found in a X-ray diffraction experiment corresponds to the lowest one in solution. The only "clean" way to compare theoretical and experimental results is under gas phase, low-temperature conditions. Although this very general and seemingly trivial statement holds for practically all computational chemistry work, it is nonetheless repeated here. Note that we always assume equilibrium conditions meaning that effects *e.g.* by kinetic trapping are excluded so that species may be missing in the theoretical ensemble.

If deviations occur in "fair" comparisons, *i.e.*, the simulated conditions match the experimental ones, their cause is mostly rooted in an inaccurate GFN*n*–xTB PES. This can be checked by a re-ranking of the GFN*n*–xTB ensemble at a reasonable, dispersion-corrected DFT level

(GGA or better a hybrid using at least a triple-zeta AO basis set). If the relative energies for DFT optimized structures differ substantially from the GFNn–xTB results (strong re-ordering), one has to be cautious. Another option to shed light on this case is to employ other, already existing tight-binding variants, *e.g.*, GFN1–xTB for comparison. In the future this problem is may be solved by applying faster computers (*i.e.*, running CREST on the DFT PES) or better TB methods (we are working on this). At this point, however, the only general recommendation is to employ in critical cases larger energy windows and to re-evaluate more structure candidates at higher level.

Less often problems are encountered from the CREST algorithms themselves. If the system is large and the PES is complicated, the applied finite run time in the MTDs may not allow sufficient exploration of a huge structural space. The best way in our opinion to tackle this issue is to employ many different initial structures for CREST. They can be obtained from chemical intuition or other algorithms but could also be generated by CREST using the various "`-quick`" run modes.

Furthermore one should be aware that missorting of the CRE is possible due to the threshold based approach. Hence, in cases of chemical systems with very dense conformational ensembles (*i.e.*, conformations with $\Delta E \ll 0.1\,\mathrm{kcal\,mol^{-1}}$ and very similar rotational constants) it is important to check the influence of different sorting thresholds on the final CRE. This is often the case for large and/or flexible molecules with a huge conformational space, but can also be encountered in smaller systems.

3.9. Conclusion

We presented a variety of automated quantum chemical screening procedures for the efficient exploration of the low-energy chemical space. The main focus herein is the generation and separation of the different isomers that are referred to as conformers. Different conformations contribute to various physical observables, such as NMR shifts and coupling constants, reaction barriers or pKa values. Therefore, the knowledge about a molecules' ensemble of different conformations is a valuable information required for accurate computational modeling. We have shown a computational workflow for the generation of conformers based on a metadynamics (MTD) approach with a self-similarity energy penalty V_{bias} that utilizes the atomic Cartesian RMSD as a collective variable. Furthermore, the procedure includes a genetic structure crossing (GC) step and was implemented in an iterative algorithm, which is conveyed in the abbreviation iMTD-GC. The induced directionallity of the chemical space exploration due to the RMSD bias allows for shorter simulation times compared with conventional MD based sampling approaches. Savings in the computational cost are in turn invested for the re-optimization of geometries at a low-cost QM level. The iterative *ansatz* hereby helps to explore the conformational space with respect to the global minimum structure, as it only terminates if no lower energy conformers are found. The algorithm is general in the sense that it will work on every level of theory for the underlying PES. However, the metadynamics sampling shows its potential especially in

combination with SQM methods that make the evaluation of thousands of molecular geometries feasible. For this purpose, the robust and reliable GFN2–xTB method was employed. Even at a relatively cheap level of quantum chemistry structures are generated that often match quite well with the experimentally observed conformations. The performance was assessed for several systems with up to 220 atoms in direct comparison with either experiment or high level theoretical data. The investigated systems include typical organic compounds, organo-metallic complexes, and non-covalently bound clusters. The conformational search algorithm can also be modified by applying additional constraints, *e.g.*, fixing of different bond lengths. This constrained conformational sampling can be a valuable tool for obtaining better TS geometries and barriers.

Additional screening workflows for the generation of non-covalently bound complexes as well as protomer and tautomer ensembles are also implemented in the presented CREST program. These procedures benefit from the SQM treatment in the same way as the conformational sampling, leading to an efficient exploration of the respective chemical space. The examples discussed here were investigated mainly at the GFN2–xTB level of theory, which is sufficient for a qualitative discussion of the presented procedures. Hence, the resulting ensembles can be used and evaluated in different ways. In the first case, the fast exploration of the potential energy surface at SQM level provides information to enhance the chemical understanding of the system. In the second scenario, if large-scale computational studies shall be conducted, the ensemble is a good starting point for further refinement (re-optimization) at the DFT or WFT level.

Overall, the results show that the procedures implemented in CREST in combination with low-cost QM methods provide a generally applicable workflow for sampling the low-energy chemical space. The straight forward handling of the program make its standard application easily feasible and hence can be an excellent staring point for chemical investigations. Ongoing work includes the optimization of faster (entirely force field based) variants of the conformational screening algorithm to extend the scope of CREST to much larger systems and its application for unknown compound identification workflows.

Acknowledgments

This work was supported by the DFG in the framework of the “Gottfried Wilhelm Leibniz-Prize” awarded to S.G. and by the DFG in the framework of the priority Program No. SPP 1807 “Control of Dispersion Interactions in Chemistry”.

Supporting Information

Some additional supporting information can be found in Appendix A2. The open-source CREST program and source code can be downloaded from `https://github.com/grimme-lab/crest`.

Part III.

Statistical Thermodynamics of the Low-Energy Chemical Space:

Calculation of Absolute Molecular Entropies and Heat Capacities

The entropy is one of the fundamental quantities in thermodynamics and commonly associated with a state of uncertainty or disorder. Historically, "entropy" was introduced as a term in 1865 by R. Clausius,[374] who later became dean at the University of Bonn. Important connections of the entropy to statistical mechanics have been made by Boltzmann and Gibbs,[213,214] and today it is still an active field of research, *e.g.*, in information theory.[375] Part III hence follows a long

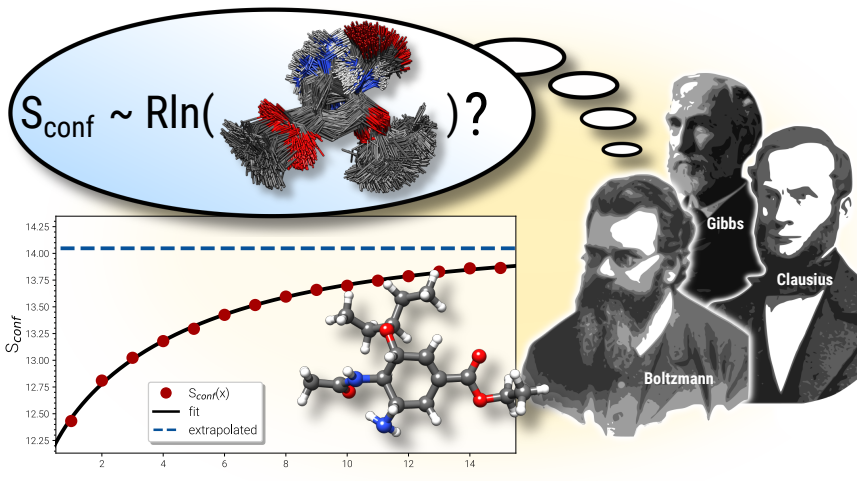

Figure III.1.: Introduction figure to the article presented in Chapter 4. [*Chem. Sci.* **2021**, *12*, 6511–6568]

tradition of research and is dedicated to accurate calculations of the absolute molecular entropy. As outlined in Chapter 2.3, the entropy of a molecule can be calculated from the vibrational partition function in a rigid-rotor harmonic-oscillator approximation. Among other things, this leads to a neglect of anharmonicities and missing degrees of freedom (DOF) due to other molecular conformations. A detailed overview of the respective thermodynamical entropy expressions and previous developments is provided in the chapter below. Mathematical formulations for these terms are well known,[43,376] but especially for the complicated conformational contribution no generally applicable workflows existed so far. Hence, an automated and numerically stable workflow has been implemented in CREST that calculates the missing conformational entropy at GFNn–xTB or GFN–FF level from extended sampling of the conformational space. As a novelty, conformational entropies are extrapolated from intermediate ensembles during execution of the CREST algorithm and frequencies (calculated by DFT *and* GFNn–xTB/FF) are scaled and interpolated[42] between vibrational and rotational partition functions.

In comparison with experimental data it is shown that with the presented approach excellent accuracy, much better than "chemical accuracy" of 3 cal mol^{-1} K^{-1}, can be achieved for absolute entropies (and the related heat capacities) of flexible molecules. Due to its GFNn–xTB/FF origin, the workflow can routinely be applied to systems up to roughly 200 atoms and (almost) arbitrary elemental composition. This is demonstrated for a set of drug-like molecules and some prototypical chemical reactions involving a large change of DOF.

4. Calculation of Absolute Molecular Entropies and Heat Capacities Made Simple

Philipp Pracht* and Stefan Grimme*

This article is part of the "2021 Chemical Science HOT Article Collection" and the "2021 ChemSci Pick of the Week Collection"

Received 1st of February 2021, Published online 24th of March 2021

Reprinted (adapted) with permission from†
Pracht, P.; Grimme, S. *Chem. Sci.* **2021**, *12*, 6551–6568.

DOI 10.1039/d1sc00621e

Own manuscript contribution

- Further development of the CREST code
- Performing and supervising the computations
- Interpretation of the computed data
- Writing the manuscript

*Mulliken Center for Theoretical Chemistry, Institut für Physikalische und Theoretische Chemie, Rheinische Friedrich-Wilhelms-Universität Bonn, Beringstraße 4, 53115 Bonn, Germany
†Reproduced with permission from the Royal Society of Chemistry.

Abstract

We propose a fully-automated composite scheme for the accurate and numerically stable calculation of molecular entropies by efficiently combining density-functional theory (DFT), semiempirical methods (SQM), and force field (FF) approximations. The scheme is systematically expandable and can be integrated seamlessly with continuum-solvation models. Anharmonic effects are included through the modified rigid-rotor-harmonic-oscillator (msRRHO) approximation and the Gibbs-Shannon formula for extensive conformer ensembles (CEs), which are generated by a metadynamics search algorithm and are extrapolated to completeness. For the first time, variations of the ro-vibrational entropy over the CE are consistently accounted-for through a Boltzmann-population average. Extensive tests of the protocol with the two standard DFT approaches B97-3c and B3LYP-D3 reveal an unprecedented accuracy with mean deviations $<1\,\mathrm{cal\,mol^{-1}\,K^{-1}}$(about <1–2%) for the total gas phase molecular entropy of medium-sized molecules. Even for the hardship case of extremely flexible linear alkanes ($C_{14}H_{30}$–$C_{16}H_{34}$), errors are only about $3\,\mathrm{cal\,mol^{-1}\,K^{-1}}$. Comprehensive tests indicate a relatively strong variation of the conformational entropy on the underlying level of theory for typical drug molecules, inferring the complex potential energy surfaces as the main source of error. Furthermore, we show some application examples for the calculation of free energy differences in typical chemical reactions.

4.1. Introduction

A main goal of computational chemistry is to realistically model various chemical reactions and predict their products. While those reactions are usually carried out at room temperature in solution, quantum mechanical (QM) calculations are primarily conducted for isolated molecules at absolute temperature zero. In order to compare theory with experiment, additional corrections and computational steps are required. Calculations of thermodynamic properties at finite temperatures are essential and if we neglect here the issue of solvation, the basic problem is an efficient computation of the molecular entropy.[3,32]

As for most other thermodynamic properties, QM computations of the entropy are commonly based on frequency calculations in the harmonic oscillator (HO) approximation. This is then usually extended by the rigid-rotor model, giving rise to the rigid-rotor-harmonic-oscillator (RRHO) approach. A comparison of entropies calculated in this way to experimental values for small molecules reveals an insufficient accuracy already for relatively rigid molecules mainly due to anharmonicity effects.[377–380] Because RRHO errors are often systematic, a common strategy is linear or multi-parametric scaling of the HO vibrational frequencies to mimic the effect of anharmonicity.[381–387] However, even frequency scaling is unable to account for all of the missing contributions to the entropy.

Approaches that compute the absolute entropy can be roughly categorized into two major classes. The first go beyond the HO approximation and explicitly account for anharmonicities in the description mainly for low-frequency, torsional normal modes. For example, this can be

done by construction of one-dimensional (1D) potential energy surfaces (PES) along the respective normal modes, as in the uncoupled normal mode approach of Sauer and coworkers.[388–390] This scheme was later adapted by Head-Gordon *et al.*[380] to include a separate treatment of vibrational and torsional modes (UM-VT). Advances have also been made for approaches that investigate coupled torsional motions.[391–393] Another method that includes the torsional anharmonicity via 1D-PES and takes multiple structures into account is the MS-T approach (and its variants), developed by Truhlar and coworkers.[394–396] Good results can be achieved with all of the above schemes, but in practice the construction of the PES and the relevant modes is technically involved, often only possible for relatively small molecules and unfeasible for routine computational chemistry workflows.

A stronger focus on multiple minima (molecular configurations/conformers) leads to the second class of approaches. Here, thermodynamic properties are approximated only by considering the *unique* minima on the PES, which in the molecular case are the different conformations. In the context of the mode following (MF) approaches discussed above, this can be understood because anharmonic torsional modes describe the transition between low-lying conformations.[40,237] Although entropies and heat capacities are thermodynamic features encoded rather globally in the shape of the PES[397,398], conformations can be used to map the problem to well-defined points on the PES. More specifically, part of the absolute entropy is computed by an informational thermostatistic partition function (Gibbs-Shannon entropy[375,399]) that only depends on a given Boltzmann probability distribution of the conformers. This idea was pursued in the so-called "minima mining" approaches,[376,400–402] where effects of anharmonicities are partially absorbed into the conformational entropy. As for the MF methods, a wide variety of different schemes exist,[43,403–405] such as the so-called mutual information expansion (MIE)[406,407], or the maximum information spanning tree (MIST)[408,409] procedures. More recent developments were introduced by Suárez and coworkers.[410–412] In their approach, the thermodynamic quantities are obtained from snapshots along an extended molecular dynamics (MD) trajectory, which are associated with unique molecular conformations. The vibrational contributions are averaged over all snapshots, while the configurational entropy is calculated via an MIE. This is doable at a force field (FF) level, but will become cumbersome for medium sized drug-like molecules at higher theoretical levels. Note that essential parts of these schemes depend solely on structure based descriptors (dihedral angles). Other studies in the literature,[413] employ some kind of flexibility measure to empirically derive molecular entropies and even more recently Hutchison *et al.* have used structural descriptors to develop a promising machine learned estimation of conformational entropy.[414]

In this study, we introduce an improved scheme that is developed from the minima mining approach and is designed to work in an almost "black box" fashion in combination with modified RRHO calculations. Herein, for the calculation of conformational entropies the recently developed GFN2–xTB[36,39] tight-binding MO and GFN–FF[80] force field methods are employed to keep computational cost under control and improve the PES description in comparison to many standard FFs. Both methods are consistently available for all elements in the periodic table

up to radon (Z=86). Below, we will first start with a general overview of the partitioning of entropies and heat capacities, followed by a description of technical novelties and the automated procedure used for the conformational part. After discussing general observations with regard to entropy calculations, benchmark results for entropies and heat capacities are presented in comparison with experimental gas phase values. In the last section we apply our scheme to some biochemically relevant systems (drug molecules) and discuss a few prototypical chemical applications.

4.2. Theory

The absolute molecular entropy in the Born-Oppenheimer approximation consists of translational (trans), rotational (rot), and vibrational (vib, also termed internal) parts

$$S = S_{trans} + S_{rot} + S_{vib}\,. \tag{4.1}$$

The most complicated vibrational contribution can be further decomposed according to

$$S_{vib} = S_{HO} + S_{anharm} + S_{conf}\ , \tag{4.2}$$

where HO denotes the harmonic oscillator value, S_{anharm} its anharmonic correction and S_{conf} is the conformational entropy arising from the population of different conformational minima. This last term is relevant for many chemically important and often non-rigid molecules like alkanes or typical drugs. Its efficient computation is the main point of this work. The corresponding partitioning and formulas can be derived analogously for the heat capacity C_p for which only the finally used equation is reported below (see Eq. 4.13).

If S_{anharm} is neglected or as usually absorbed into a scaled S_{HO} term or partially accounted for by S_{conf} (see below), Eq. 4.1 can be rewritten as

$$S = S_{RRHO} + S_{conf}\ , \tag{4.3}$$

where S_{RRHO} refers to the usual rigid-rotor-harmonic-oscillator approximation for the rotational/translational and internal parts, respectively. In the following, in order to avoid terminology problems,[43] we denote all parts of the entropy that are *not* included in S_{RRHO} (or S_{msRRHO}, see below) of a given reference structure as *conformational* or *configurational* entropy and will use the terms interchangeably. The decomposition used above is physically motivated by the fact that some vibrational anharmonicity effects, at least for not too large distortions, maintain the equilibrium structure (bond stretching and many angle bendings), while many torsion motions lead to new (conformational) minima with low barriers. This partitioning of the entropy into vibrational and conformational parts was first introduced by Karplus *et al.*, and has since been used in many studies.[43,401,404,415–417]

A well-known problem of RRHO-based entropy calculations is that S_{vib} tends to infinity for

vibrational frequencies approaching zero. In actual calculations for larger, flexible molecules, many low-frequency vibrational modes appear which are often better characterized by internal rotations of functional groups rather than by stretching or bending vibrations. They are in a typical range of 5–50 cm^{-1} and can spoil the computed entropy due to artificial numerical errors and their strong anharmonicity components. Correction schemes exist which explicitly treat such modes anharmonically in a coupled or uncoupled form[380,396]. These methods require the costly computation of one-dimensional (1D) PES as well as definition of special internal coordinates. In our opinion, while such methods can be beneficial and accurate for small to medium sized and not too flexible molecules ($\approx$ 20–30 atoms), they are not viable for a robust and rather general treatment for systems with hundreds of atoms.

In 2012, one of us proposed to modify the treatment of the low-frequency part of the vibrational spectrum by taking a so-called rotor-approximation and continuously interpolating between a rigid-rotor and vibrational description for each mode.[42] Herein, the vibrational entropy of a harmonic oscillator with frequency ν at temperature T is given by

$$S_V = R \left[\frac{h\nu}{kT} \frac{e^{-h\nu/kT}}{(1-e^{-h\nu/kT})} - \ln(1-e^{-h\nu/kT}) \right] . \tag{4.4}$$

The rigid-rotor entropy for a free rotor is given by

$$S_R = R \left[\frac{1}{2} + \ln \left\{ \left(\frac{8\pi^3 \mu' kT}{h^2} \right)^{1/2} \right\} \right] , \tag{4.5}$$

where μ' describes the dependence on the average molecular moment of inertia B_{av} and the frequency of the normal mode

$$\mu' = \frac{\mu B_{av}}{\mu + B_{av}} , \tag{4.6}$$

with $\mu = \frac{h}{8\pi^2 \nu}$. In Eqs. 4.4–4.6, h is Planck's constant, R is the gas constant, and k is Boltzmann's constant. The final continuously interpolated S_{mRRHO} entropy ("m" for modified) is then given by a sum over all normal modes

$$S_{mRRHO} = S_{trans} + S_{rot} + \sum_{i}^{modes} \left[\frac{S_V}{1+(\frac{\tau}{\nu_i})^\alpha} + \left(1 - \frac{1}{1+(\frac{\tau}{\nu_i})^\alpha} \right) S_R \right] , \tag{4.7}$$

with $\alpha = 4$ (introduced with the damping function in Ref. 140). This does not involve any computational overhead compared to a standard HO calculation and merely requires the definition of a vibrational energy threshold τ below that the rotor entropy instead of the vibrational one is continuously taken. A related (but discontinuous) treatment has been proposed by Truhlar.[418] A typical value used by us since years in standard thermochemical studies is $\tau = 50\, cm^{-1}$. In this work, we consider τ for the first time as an adjustable parameter to account for part of the non-conformational anharmonicity effects. Furthermore, calculated harmonic frequencies are linearly scaled by a factor ν_{scal}, as is common practice[381–383] to account for deficiencies

of the underlying method employed for the PES calculation and further anharmonicity effects mainly in the high-frequency part. The only two empirical parameters included are adjusted to reproduce experimental entropies for a benchmark set of mostly rigid molecules (see below). For better distinction this modified RRHO treatment is in the following denoted by S_{msRRHO} ("s" for scaled).

The major aim of this work was to find a robust approximation to S_{conf} which is already significant for medium flexible molecules (see section 4.4.4). We build upon the original idea of Gilson and co-workers[376] termed "minima mining" or "mixture of conformers" strategy, which has later been applied to organic molecule entropy calculations by DeTar[401] and Guthrie[402]. The basic formula reads

$$S_{conf} \approx S_{mix} = -R \sum_{i}^{conf} p_i \ln p_i \tag{4.8}$$

and approximates S_{conf} by the conformer mixing entropy S_{mix} summed over a conformer ensemble. The thermal populations p at absolute temperature T are given by

$$p_i = \frac{g_i e^{-E_i \beta}}{\sum g_j e^{-E_j \beta}} \ , \tag{4.9}$$

where $\beta = \frac{1}{kT}$, E_i is the energy of the equilibrium structure of conformer i, and g_i is a general state degeneracy. The conformational entropy depends on the level of theory through the calculated populations entering the Gibbs-Shannon entropy formulation in Eq. 4.8, which in turn depend directly on the equilibrium (free) energies. But also for other configurational entropy approaches, that are usually cited as being *purely informational*,[43,411] there exists a bias towards the underlying method used for the generation of molecular structures, for example by MD simulations. This is especially problematic for very crude approximations of the conformational entropy, *e.g.*, based only on the number of conformers N_{conf} according to $S_{conf} \approx R \ln (N_{conf})$. This approximation is used in some studies[402,419] and is appealing due to its simplicity. However, while this formulation may be used for very simple molecules, it breaks down for more complex PES. Further discussion of this point is given in Appendix A3.

The sum in Eq. 4.8 is taken over all significantly populated, *distinguishable* structures representing a so-called generalized Boltzmann distribution.[399] The problem of this procedure (also termed Gibbs-Shannon entropy based procedure) is that not only an almost complete conformer ensemble has to be found but additionally, it should be "pure", *i.e.*, free of so-called rotamers. In this case for molecules with non-degenerate electronic ground states, all g_i are unity. Rotamers are structures indistinguishable by any nuclear spin-independent quantum mechanical observable. They arise from rotation around covalent chemical bonds (or other inversion-type processes) that interchange nuclei belonging to the same group of nuclides, as for example the interchange of protons at a methyl group by rotation.

In this work, we propose and implement for the first time an automatic algorithm that generates a theoretically proper ensemble of unique conformer structures required for the accurate computation of S_{conf}. For the conformer search problem, we employ our recently described

CREST program[33] (abbreviated from Conformer-Rotamer Ensemble Sampling Tool), which is based on metadynamics simulations employing on-the-fly computed quantum mechanical tight-binding PES.[33,41] We assume at this point that the conformer-rotamer ensembles (CRE) obtained from CREST are sufficiently complete and the energies E_i are accurate. If this is really the case for very flexible molecules (*e.g.* long alkanes) can be tested by comparison of computed and experimental entropies and heat capacities (see Sec. 4.4.2 and 4.4.3). Note that our approach works with any (on-the-fly computed) PES and hence, at least in principle, the errors introduced by the underlying method for the PES and the other approximations to the entropy problem could be decomposed.

The CREST algorithms were originally developed to generate rotamer containing ensembles and the related nuclei-exchange information for the simulation of NMR spectra[40]. Hence, it seems straightforward not only to identify rotamers, but to extend the algorithm to automatically compute the proper degeneracy number g_i. However, as mentioned above, conformer ensembles (CE) must be free from the indistinguishable rotamers to be compatible with entropy calculations. Therefore, g_i are treated as unity in the usual case.

The only exception here are symmetrical molecules that can form "enantiomeric" (*i.e.*, in principle distinguishable) conformers through rotation of bonds. A typical case is the gauche conformer of *n*-butane. These geometrical enantiomers are degenerate and would be falsely classified as rotamers in our previous implementation. Effectively, this introduces a factor of $g_i' = \{1, 2\}$ instead of g_i in the degeneracy, depending on if the formation of a geometrical enantiomer is possible. Our new approach considers this problem for the first time in a correct and automated way. Inserting this into the standard entropy expression for degenerate states[215] leads to

$$S'_{conf} = R\left[\ln \sum g_i' e^{-E_i\beta} + \frac{\sum g_i'(E_i\beta)e^{-E_i\beta}}{\sum g_i' e^{-E_i\beta}}\right] . \tag{4.10}$$

The correct S_{msRRHO} entropy is a population average over the CE, analogously to other physical observables. Unfortunately, the many costly DFT geometry optimizations and frequency calculations will quickly become the computational bottleneck for moderately sized systems. Therefore, as a further approximation, we compute S_{msRRHO} at the DFT level for the lowest conformer and add the respective ensemble contribution as a thermostatistical average over all populated conformers at a less computationally demanding, lower theoretical level. The arising $\overline{S}_{msRRHO}$ term is given by

$$\overline{S}_{msRRHO} = \left(\sum p_i S_{msRRHO,i}\right) - S_{msRRHO,ref} , \tag{4.11}$$

where $S_{msRRHO,i}$ is the absolute msRRHO entropy of the conformer calculated at the low force field or SQM level to avoid very many (high level/DFT) HO calculations. $S_{msRRHO,i}$ and the free energies (G_i) are only explicitly calculated for the lowest $\geq 90\%$ populated (based on initial total energies E_i) conformers while for all others, the average is taken. The populations p_i refer to Eq. 4.9 and are calculated using G_i from the corresponding msRRHO calculations. For

convenience, we subtract the entropy of a reference structure $S_{msRRHO,ref}$ in Eq. 4.11 such that $\overline{S}_{msRRHO}$ can be added directly taken as a further correction to the S_{mRRHO} result taken from any standard quantum chemistry code. $S_{msRRHO,ref}$ typically refers to the DFT reference structure, for which vibrational frequencies are calculated at the SQM or FF level. To avoid changes to the geometry and appearance of imaginary vibrational modes, we here additionally make use of a new procedure called Single Point Hessian (SPH),[22,420] for which some details are given in Appendix A3. Note that if $\overline{S}_{msRRHO}$ is calculated at the same level as S_{msRRHO}, one would arrive at the correct population average because S_{msRRHO} and $S_{msRRHO,ref}$ exactly cancel each other. The treatment would then be exact.

Thus, our final working equation for the molecular entropy is given by

$$S_{conf} = S'_{conf} + \overline{S}_{msRRHO} \ . \tag{4.12}$$

The corresponding formula for the heat capacity at constant pressure is

$$C_{p,conf} = R \left(\frac{\sum_i g_i \left(E_i\beta\right)^2 e^{-E_i\beta}}{\sum_i g_i e^{-E_i\beta}} \right) - R \left(\frac{\sum_i g_i \left(E_i\beta\right) e^{-E_i\beta}}{\sum_i g_i e^{-E_i\beta}} \right)^2 \ , \tag{4.13}$$

and the enthalpy is

$$[H(T) - H(0)]_{conf} = RT \frac{\sum_i g_i \left(E_i\beta\right) e^{-E_i\beta}}{\sum_i g_i e^{-E_i\beta}} \ . \tag{4.14}$$

Note that g_i is used in C_p and $H(T) - H(0)$ instead of g'_i. In our opinion, basing S_{conf} (and related properties) directly on a given level of theory via the Gibbs-Shannon entropy of an ensemble (Eq. 4.8 and 4.10) provides a genuine understanding of the quantity in accordance with chemical intuition. Furthermore, it can be very well coupled to automated conformational search tools, which are anyway necessary for accurate computation of other physical observables.

4.3. Implementation and Computational Details

4.3.1. Extrapolation to Ensemble Completeness

For very flexible systems (*e.g.* long alkanes), the number of accessible conformers Ω is roughly proportional to $\Omega \approx 3^R$, where R is the number of freely rotatable bonds (commonly associated with the number of sp^3-sp^3 carbon single bonds).[419] In principle, all conformers, *i.e.*, the complete ensemble and the respective energies are required for the calculation of S_{conf} but even for only moderately sized systems this number is prohibitively huge.

Practically, the obtained ensemble quality depends mostly on the run time t of the (biased) molecular dynamics (MD) in CREST. Basically, it is the number of optimized snap-shot structures gathered over all runs and will converge to a complete CE with the length of the conformational search. On the other hand, the conformational entropy also exhibits predictable behavior with regard to increasing ensemble completeness. If the lowest energy conformer is known, adding higher-lying conformers to the ensemble can only increase the entropy. If many

of the low-energy structures are already found, the entropy increase for additional states is smooth and it seems possible to extrapolate to completeness without explicit knowledge of all conformers. The pre-requisite for this is the generation of enough intermediate points, *i.e.*, consecutive conformational ensembles with systematically improved quality. A smooth and continuous convergence of the entropy to its maximum value can only be observed if conformers are added consistently from all regions of the PES (see Sec. 4.4.2 for examples).

In the implementation of the algorithm, information from incomplete CEs of consecutive iterations is used for an extrapolation of the entropy according to

$$S'_{conf}(x) - S'_{conf}(0) = p_1 \left(1 - \exp\left(p_2 x^{p_3}\right)\right) \ , \tag{4.15}$$

where x is the iteration number, and $S'_{conf}(0)$ refers to the result of the first initial conformer ensemble from the new CREST workflow (see Sec. 4.3.2). The parameters p_1, p_2 and p_3 are fitted automatically to the available data points from each entropy sampling run employing the Levenberg-Marquadt[421,422] algorithm. In summary the extrapolation can be seen as an unsupervised learning procedure used to correct for incompleteness.

4.3.2. Algorithmic and Technical Details

The conformational entropy calculation as described above is performed with the recently published CREST program.[33] A special run type was implemented for this purpose, where the focus is set to an extensive sampling around the global and low-lying local minima. Ideally the calculation of S_{conf} should be conducted from the already known global minimum conformer, *e.g.*, obtained from another conformational search with default settings in CREST. The enantiomer degeneracy number g_i is obtained automatically as described in detail in Appendix A3. For the msRRHO part, any quantum chemical method or even force fields can be applied. Here, we use the composite DFT method B97-3c[180] and the well-known B3LYP-D3 functional[55,172,173] in a standard def2-TZVP basis.[338] Molecular symmetry numbers are automatically determined for each conformer entering $\overline{S}_{msRRHO}$ and should be also included in the DFT frequency evaluation.

The few simple steps required for the calculation of the absolute entropy are

1. Run CREST in default mode on a starting structure to find the lowest conformer
2. Optimize the geometry of this conformer with DFT, compute the Hessian matrix from the DFT structure and use the HO vibrational frequencies to calculate S_{msRRHO}
3. Run CREST in entropy mode on the lowest-energy conformer and employ the DFT reference structure for $\overline{S}_{msRRHO}$, resulting in S_{conf}
4. Compute $S = S_{msRRHO} + S_{conf}$

Note that for large systems step two could in principle also be conducted at a low theory level (SQM or FF). However, because step three is usually the computational bottleneck, it

is recommended to take S_{msRRHO} from a more accurate DFT treatment. In general, this partitioning allows systematic improvements of the scheme because the different contributions can in principle be calculated at any level of theory.

If no low-lying conformers (relative energy <1-2 kcal/mol at ambient temperature) are found in the first step, the entropy run is not necessary and the plain S_{msRRHO} value can be taken. The default setup for the metadynamics bias potentials in the entropy mode and further technical settings were empirically determined on a few test cases similar to the optimization of the run parameters in a conventional conformer search run[41] (see CREST documentation and source code[423]). Note that the MD runs are by default initiated with random numbers and hence the details of the obtained CE vary stochastically. For larger, very flexible molecules with a complicated PES this can amount to stochastic variations of 2–5% for S_{conf} (see also Section 4.4.4 for discussion).

The general workflow for the computation of S_{conf} in CREST is outlined in Fig. 4.1. The

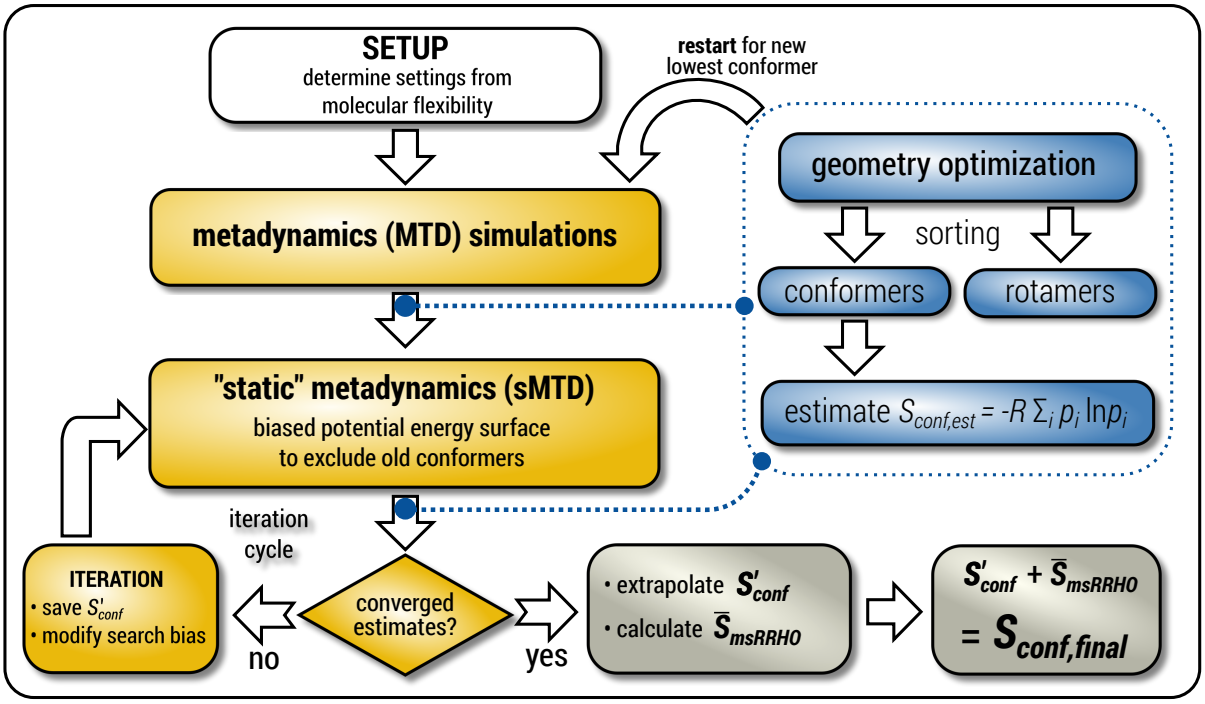

Figure 4.1.: Schematic representation of the workflow used for the computation of S_{conf}. See text for details.

procedure is designed to work fully automatic and to provide intermediate ensembles for entropy extrapolation as described above. For the input structure, the run time t of the biased MD is determined automatically from a covalent and non-covalent flexibility measure (see Sec. 4.4.4 and Appendix A3). To create an initial structural ensemble, 24 metadynamics (MTD) simulations are conducted with several different bias parameters as in the default CREST runtype. The structural ensemble obtained from this step is later used as the reference to calculate $S'_{conf}(0)$ (see Eq. 4.15). Structures are sorted according to their relative energy, structural Cartesian RMSD,

and rotational constants to distinguish between unique conformers and degenerate rotamers, as described in Ref. 33.

From the CEs two sets of structures are extracted *via* a combined principle component analysis (PCA)[424,425] and k-Means clustering[426,427] approach, using dihedral angles as geometrical descriptors. The first set of structures, which always consists of 36 structures, is used as input for further metadynamic simulations. The other set consists of a number of structures that depends on the molecular flexibility and current ensemble size. This second ensemble is used to generate a global bias potential in the metadynamics simulations and, in contrast to the initial MTD simulation, is not updated with new bias structures. The idea here is to apply this new unchanged bias similar to a global potential used in classical umbrella sampling[428] or basin-hopping algorithms[429,430] to efficiently block entire energy basins of the PES and direct the conformational search to new minima. For better differentiation, this is referred to as static metadynamics simulation (sMTD). The ensemble obtained by sMTD is merged with the previous ensemble and a preliminary conformational entropy $S_{conf,est}$ is determined. If no change (within a 0.5 % threshold) in $S_{conf,est}$ and the total number of unique conformers (within 2 %) is observed, the final conformational entropy is calculated. Otherwise, a new iteration of 36 sMTDs is conducted using input structures and static bias structures determined from the updated ensemble. Furthermore, with each iteration the number of static bias structures is increased. This procedure is repeated until convergence is reached both with regards to $S_{conf,est}$ and the number of unique conformers in the ensemble. For the final calculation of S'_{conf}, an extrapolation as described in Sec. 4.3.1 is conducted. This new algorithm in CREST can also be used for normal conformer search with the keyword `--v4`. The default convergence thresholds were conservatively chosen to provide good reproducibility (see section 4.4.4), but can manually be adjusted.

A problem may appear if the rather approximate PES used in CREST (here GFN2–xTB or GFN–FF) is substantially different from the DFT PES (here B97-3c or B3LYP-D3/def2-TZVP). This is indicated by different lowest-energy conformers and significant energetic re-ordering of the CREST ensemble obtained with the GFN methods after refining (re-optimizing) it with the respective DFT methods. In such cases, we suggest to use the S_{msRRHO} value obtained for the lowest DFT conformer and corresponding S_{conf} from the GFN ensemble. If the lowest GFN and DFT conformer structures agree qualitatively, this approximation seems to be reasonable according to our experience.

Ideally, the PES employed for the initial conformational search and the one used for automatic S_{conf} calculation should be the same. Here, we employ the GFN2–xTB tight-binding method[39] and the recent general force field GFN–FF[80] and compare the results. The latter speeds-up the CREST calculations by a factor of 10–30 for typical cases with 50–100 atoms. The S_{msRRHO} value is always computed with B97-3c and a frequency scaling factor ν_{scal} of 0.97, or B3LYP-D3/def2-TZVP with a frequency scaling factor ν_{scal} of 0.98. Test calculations employing GFN2–xTB in this step yield somewhat less accurate results and, because the calculation of S_{conf} is the computational bottleneck, do not reduce the overall computational times significantly. In all

frequency calculations, a S_{msRRHO} cut-off value of $\tau = 25\,\mathrm{cm}^{-1}$ was employed. τ and ν_{scal} (for the DFT methods) were adjusted to perform equally well in combination with both GFN–FF and GFN2–xTB. CREST is essentially a driver for the `xtb` program[431] which is used for all GFN calculations. For the DFT calculations, TURBOMOLE 7.4[432,433] is used throughout.

4.3.3. Benchmark Sets

For the initial tests and determination of the empirical parameters τ (msRRHO cut-off) and ν_{scal} (DFT frequency scaling factor) we employ the benchmark set of Li, Bell and Head-Gordon (LBH).[380] This LBH set consists of 39 organic molecules ranging from ethane (smallest) to n-octane (largest) and is shown in Appendix A3. For cross-validation we extended this set by 23 similar, but mostly larger molecules ranging from cyclohexane (smallest) to n-dodecane (largest). This set is termed AS23 (Absolute Entropy) from now on and is described also in Appendix A3. The corresponding experimental gas phase reference entropies and $C_p(T)$ values are taken from Refs. 434,435. Studies are available in the literature presenting much larger collections of experimental reference data, *e.g.*, in Ref. 419. However, these databases contain mostly small, rather rigid systems (*e.g.*, substituted aromatic compounds) which are not in the focus of our study. Nonetheless, the combined LBH and AS23 sets should sufficiently representative for benchmarking absolute entropies. To show possible limitations of our approach a set of maximally flexible linear alkanes (up to $C_{18}H_{38}$) is investigated separately.

For the heat capacities, we additionally test the temperature dependence in a typical range of 200–1500 K, while for entropies only the value at 298 K is considered. For this a subset of the LBH molecule set is used, as described in Ref. 380. Note that the numerical values and errors for entropy and C_p are similar and thus, the conclusions for the temperature dependence of the latter should also apply for the entropy.

Furthermore, in Sec. 4.4.4 we present a case study for 25 pharmaceutical (Clinical Drug) molecules, denoted CD25. There are no experimental entropy values available for this set, but differences between the ensembles (*e.g.*, gas phase versus implicit solvation) and different PES employed to calculate the entropy can be studied theoretically. We suggest this set also as a challenging test for other approaches.

4.4. Results

4.4.1. General Considerations

The absolute entropy is a complicated property which includes various terms of different magnitude that can be qualitatively interpreted.[43,376] As an example the suggested partitioning of the absolute entropy for two molecules is shown in Tab. 4.1.

The largest portion of the entropy results from the vibrational, rotational, and translational degrees of freedom (DOF), as commonly obtained by standard quantum mechanical frequency calculations employing the RRHO approximation. Contributions from translational and rota-

Table 4.1.: Contributions to the total molecular entropy for n-decane and tamiflu. RRHO and msRRHO values correspond to the B97-3c level of theory, S'_{conf} and $\overline{S}_{msRRHO}$ were calculated at the GFN2–xTB level. Relative contributions are given in percent next to the respective contribution.

	S cal mol^{-1} K^{-1}			
contribution	n-decane		tamiflu	
RRHO	116.4		169.0	
msRRHO	117.3	(89.9%)	173.4	(91.6%)
vib.	47.2		95.4	
rot.	29.4		34.9	
trans.	40.8		43.1	
anharm. (msRRHO-RRHO)	0.9		4.4	
S'_{conf}	12.5	(9.6%)	13.7	(7.2%)
$\overline{S}_{msRRHO}$	0.7	(0.5%)	2.3	(1.2%)
sum	130.5	(100.0%)	189.4	(100.0%)
exptl.	130.4		—	

[a]Values taken from Ref. 380.

tional DOF have the same order of magnitude (about 30-40 cal mol^{-1} K^{-1} in Tab. 4.1) for all chemical systems of about this size (mass). In contrast, vibrational contributions quickly exceed several hundred cal mol^{-1} K^{-1} for molecules >100 atoms. In the important drug-size regime, the vibrational entropy is clearly the largest contribution and hence its accuracy depends also on how good anharmonicities are described. As defined in Sec. 4.2, the effect of anharmonicities can be estimated from the difference between the entropy calculated by the new msRRHO and standard RRHO scheme (*i.e.*, without modifying τ and frequency scaling). Looking at the two example molecules, decane shows only a relatively small RRHO-msRRHO difference of 0.9 cal mol^{-1} K^{-1} while tamiflu exhibits a much higher anharmonic contribution of 4.4 cal mol^{-1} K^{-1}. This is in line with chemical intuition, as one would expect many more anharmonic ro-vibrational modes for a complicated drug molecule like tamiflu than for a rather simple linear structure composed of only CH and CC bonds. In any case, the anharmonicity is non-negligible and must be accounted for by either τ and ν_{scal} or some more elaborate, explicit scheme. With increasing flexibility of the molecule the configurational contribution increases drastically and in fact, S_{conf} can be taken as a molecular flexibility measure (see Sec. 4.4.4).

For decane and tamiflu the conformational entropy S'_{conf} accounts for 12.5 and 13.7 cal mol^{-1} K^{-1}, respectively. Though decane (32 atoms) is smaller than the drug molecule tamiflu (50 atoms), their conformational entropy values are rather similar. The simple explanation for this is the higher flexibility of decane, which is typically indicated by a larger relative contribution of S'_{conf} to the absolute entropy for similar sized structures. In general S'_{conf} will be close to zero for the most rigid molecules or molecules with only a few distinct conformers, but adds a significant portion (ten or more percent) to the absolute entropy for highly flexible molecules.

The last contribution to S_{conf} is the population average $\overline{S}_{msRRHO}$. This term may provide

insight about the variation of S_{msRRHO} within the ensemble. It will be small if all contributing conformers have a similar ro-vibrational entropy as the reference structure (*e.g.* for decane with $0.7\,\mathrm{cal\,mol^{-1}\,K^{-1}}$), or yields a large contribution in the opposite case (tamiflu, $2.3\,\mathrm{cal\,mol^{-1}\,K^{-1}}$). For the latter, computed msRRHO entropies can vary by several entropy units for different conformations rather independently of the chosen τ or ν_{scal} values. An example is provided in Fig. 4.2, where S_{msRRHO} was calculated for 299 (random) conformers of tamiflu at two different theoretical levels (GFN–FF and B97-3c). Here, entropies at the GFN–FF level are overestimated

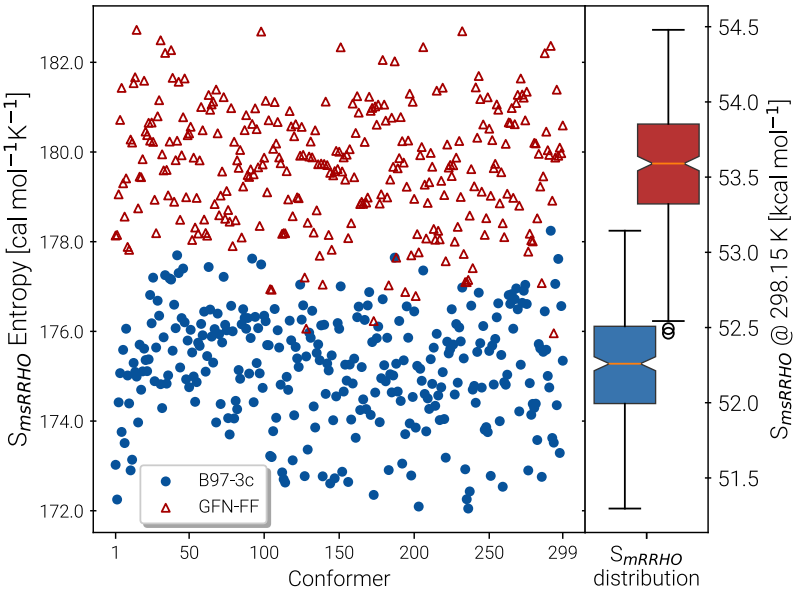

Figure 4.2.: Spread of entropies calculated in the msRRHO approximation at the GFN–FF (red) and B97-3c (blue) level. On the right side box plots for the two methods are given for an easier visualization of the metric averages and shifts.

by $4\,\mathrm{cal\,mol^{-1}\,K^{-1}}$ on average compared to the more accurate B97-3c level. Both methods show a similar spread of the S_{msRRHO} values, which range approximately $6\,\mathrm{cal\,mol^{-1}\,K^{-1}}$ from lowest to highest value thus reconfirming the use of $\overline{S}_{msRRHO}$. Hence, the validity of an approximate $\overline{S}_{msRRHO}$ obtained at SQM or FF level depends on the performance for *relative* msRRHO entropies and may be used if a shifted (*cf.* Eq. 4.11) population average similar to the higher reference DFT level is expected.

Another novelty of our approach is the extrapolation of S'_{conf} to the ensemble completeness as discussed in section 4.3.1. The corresponding procedure requires systematically and smoothly improving CE quality in each iteration. In practice, the required number of iterations is very molecule specific but convergence is typically achieved within 5–15 iterations (see Fig. 4.3 for some examples). The entropy difference between the last iteration and the extrapolated value is often relatively small but very significant for very flexible systems with huge ensembles. For example the CE of *n*-octadecane contains over half a million conformers within $6\,\mathrm{kcal\,mol^{-1}}$ at the last iteration. In a more typical case the entropy gain due to the extrapolation is smaller

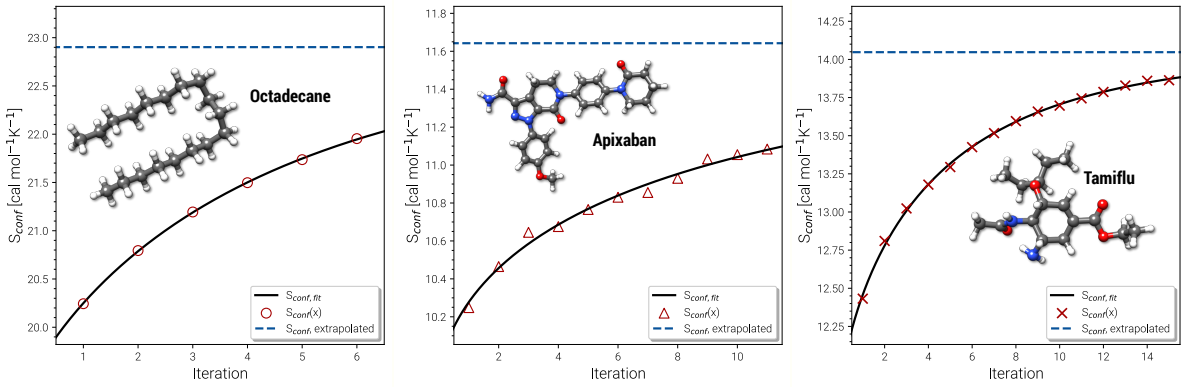

Figure 4.3.: Examples for the extrapolation of conformational entropy at the GFN–FF level of theory. The iteration number x refers to the sMTD iteriation cycle depicted in Fig. 4.1.

than one entropy unit ($1\,\mathrm{cal\,mol^{-1}\,K^{-1}}$). Apixaban and tamiflu depicted in Fig. 4.3 are such examples, but nonetheless exhibit different convergence behavior. For small molecules the extrapolation is mostly not necessary because the entire ensemble will be found during the initial sampling procedure. From another viewpoint, the extrapolation scheme might rather be seen as a technical supplement for reduction of stochastical noise between the iterations and consequently, an improved prediction the final S_{conf} value. Note, that $3\,\mathrm{cal\,mol^{-1}\,K^{-1}}$ "entropy units" refer to the usual $1\,\mathrm{kcal\,mol^{-1}}$ chemical accuracy at room temperature. Thus, with an accuracy for S better than about 1-2 $\mathrm{cal\,mol^{-1}\,K^{-1}}$, the electronic energies of the molecules from DFT or wave function theory (WFT) become the accuracy bottleneck in typical thermochemical calculations.

4.4.2. Benchmarking Absolute Entropy

Recently, Head-Gordon *et al.* published the LBH set containing 39 organic molecules and their experimental gas-phase entropies, which provides an excellent reference for the evaluation of absolute entropies.[380] For a more thorough evaluation the set was extended by the AS23 molecules. Entropy values for the two sets were calculated for four combinations of theory levels. These are S_{msRRHO} contributions obtained with either B97-3c or B3LYP-D3/def2-TZVP and the conformational entropies calculated at GFN–FF or GFN2–xTB level and with τ and ν_{scal} values as described above. Parity plots for the different levels of theory with reference to the experimental data are given in Fig. 4.4 and the corresponding statistical data are provided in Tab. 4.2.

The excellent performance of our approach is obvious from both Tab. 4.2 and the parity plots (Fig. 4.4). To the best of our knowledge, the RMSD of $0.79\,\mathrm{cal\,mol^{-1}\,K^{-1}}$ calculated at the B97-3c+S_{conf}(GFN–FF) level refers to the best performance of a theoretical method for this benchmark set ever reported in the literature. For comparison, the best performing method discussed in Ref. 380, (UM-VT, a DFT based MF approach) has a RMSD of $1.24\,\mathrm{cal\,mol^{-1}\,K^{-1}}$. For the combined LBH+AS23 set the errors are slightly larger (RMSD of 1.1–1.3 $\mathrm{cal\,mol^{-1}\,K^{-1}}$).

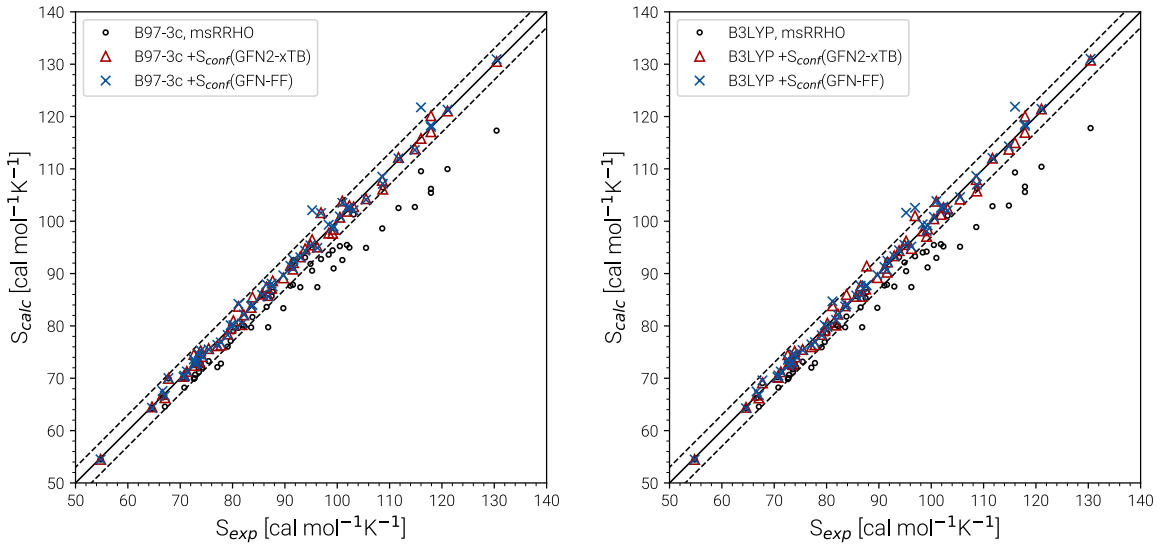

Figure 4.4.: Parity plots for calculated and experimental entropies for all molecules of the LBH and AS23 set. The combinations of B97-3c and B3LYP-D3/def2-TZVP S_{msRRHO} values with GFN2–xTB and GFN–FF S_{conf} values, respectively are shown. For reference also the plain S_{msRRHO} entropies are plotted. The solid line corresponds to perfect correlation between theory and experiment. Error bars of 3 cal mol^{-1} K^{-1} are given as dashed lines and correspond to chemical accuracy at $T = 298$ K.

Table 4.2.: Mean deviation (MD), mean average deviation (MAD), root-mean-square deviation (RMSD), and standard deviation (SD) for absolute entropies obtained at different theoretical levels in comparison to experimental data. All values correspond to standard entropies at 298.15 K in cal mol^{-1} K^{-1}. Three outliers have been removed for the final GFN–FF results (see text).

S_{RRHO}	B97-3c		B3LYP-D3/TZ		UM-VT[a]
S_{conf}	GFN–FF	GFN2–xTB	GFN–FF	GFN2–xTB	
LBH set					
MD	0.32	0.23	0.23	0.09	-0.52
MAD	0.59	0.65	0.60	0.65	0.86
RMSD	0.84	0.91	0.85	0.93	1.24
SD	0.79	0.89	0.83	0.93	1.14
full set					
MD	0.21	0.15	0.24	0.07	—
MAD	0.73	0.83	0.73	0.92	—
RMSD	1.09	1.19	1.16	1.29	—
SD	1.08	1.19	1.15	1.30	—

[a]Values taken from Ref. 380.

Yet, all of the four tested method combinations are well below the targeted chemical accuracy of 3 cal mol^{-1} K^{-1}. A similar performance on a set of 128 experimental absolute entropies was reported by Guthrie[402] using B3LYP/6-31G**, with an RMSD of 1.29 cal mol^{-1} K^{-1}. Larger,

flexible molecules in this set are identical with the ones in the LBH+AS23 set. However, Guthries benchmark set is mainly composed from rather rigid structures for which the S_{RRHO} entropy is already quite accurate.

For both B97-3c and B3LYP-D3, deviations between the calculated S_{msRRHO} (or S_{RRHO} values, data not shown) and the experimental value increase with the size and flexibility of the molecule. Only by including the conformational contributions it is possible to reach chemical accuracy. Overall, the different method combinations show fairly similar performance, although some trends can be recognized. A good performance of B3LYP-D3 is unsurprising as it is well known to be among the best performing DFT functionals for the calculation of vibrational properties[381,382] and was basically constructed for this purpose.[55] Although the (computationally cheaper) B97-3c method performs slightly better than B3LYP-D3/def2-TZVP, this is sensitive to the choice of τ and ν_{scal} and furthermore depends on the technical settings of the DFT calculations, like the choice of the grid or SCF convergence thresholds.[436] Therefore, a clear preference for one out of the two tested methods is difficult to draw.

The same is true when comparing the two assessed methods for calculating S_{conf}. S_{conf} strongly depends on the shape of the PES which can be rather different between a force field and a quantum chemical method. Since GFN2–xTB has the more physically reasonable PES of the two methods, usually a better performance should be expected. However, GFN–FF seemingly outperforms GFN2–xTB in combination with both B97-3c and B3LYP-D3 but this is mainly due to the removal of three strong outliers (3,3-dimethylpentane, 3,3-diethyl-2-methylpentane and perfluorheptane) that were discarded from the GFN–FF error statistics. For all three molecules GFN–FF produces some artificially low-lying conformers resulting in an overestimation of the conformational entropy (7 %, 5 % and 3 % respectively). Only one additional outlier, triethylamine (TEA), is observed for the combined LBH+AS23 set, but since it is present for all four method combinations, it may not be attributed to a wrong conformational energy landscape. The origin of the error for TEA (overestimation by approximately 5 %) remains unknown, but it has not been removed from the statistics presented in Tab. 4.2. Without TEA the statistics would improve even further to low MADs and RMSDs of 0.77 and 1.04 $\mathrm{cal\,mol^{-1}\,K^{-1}}$ for B97-3c and 0.87 and 1.18 $\mathrm{cal\,mol^{-1}\,K^{-1}}$ for B3LYP-D3 in combination with S_{conf}(GFN2–xTB), respectively. The best overall result for the LBH+AS23 set after removing all outliers is obtained with B97-3c+S_{conf}(GFN–FF). Interestingly, our S_{msRRHO}+S_{conf} values tend to slightly overestimate compared to the experimental data, while the opposite holds for approaches that go beyond the harmonic approximation, such as UM-VT.[380] This is indicated by the mean deviation, which for the LBH benchmark set is always positive for our approach and always negative for different version of the methods presented in Ref. 380. Tentatively, this may be attributed to some missing (configurational) contributions in UM-VT and/or to our strict separation of harmonic vibrational terms and conformational terms. The latter mainly concerns low frequency modes that are correlated to conformational transitions and which were a key motivation for the mRRHO method with the rotor cut-off τ as an adjustable variable. In other schemes, for example the one introduced by Zheng and Truhlar,[396] attempts have been made to tackle this problem

by explicitly combining the rotational, vibrational, and confromational partition function.

Linear Alkanes

Computational and accuracy limits of the presented approach are explored for the example of n-alkanes of increasing size, up to $C_{18}H_{38}$ (see Fig. 4.5). Such extremely large flexible systems have not been considered before quantitatively.

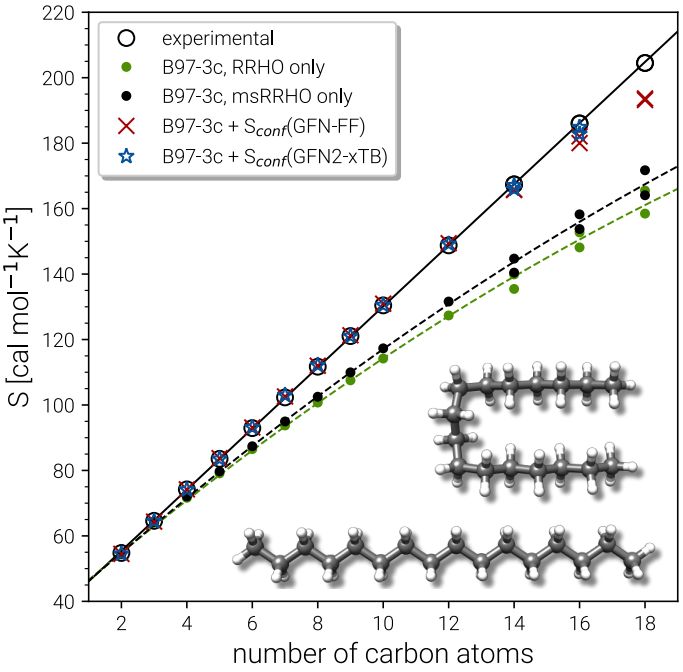

Figure 4.5.: Parity plot for calculated and experimental entropies for n-alkanes from ethane to octadecane. All values correspond to B97-3c S_{msRRHO}, either combined with GFN2–xTB or GFN–FF S_{conf}, or without the conformational contribution. For $C_{14}H_{30}$ up to $C_{18}H_{38}$ two values are shown each, which correspond to the competing linear and folded global minima (see text for details). As example the folded and linear minimum energy conformers for hexadecane are depicted.

The experimental entropy values[434,435] show a strict linear increase with the number of carbon atoms and the reproduction of this relation represents a challenging task for theoretical methods. Both the RRHO as well as the msRRHO models increasingly underestimate the entropy with growing system size leading to a strongly non-linear behavior and errors of more than 20% for the largest alkanes considered. The major part of this difference can be accounted for by S_{conf}. In fact, up to tetradecane ($C_{14}H_{30}$), the computed values are all still within chemical accuracy of $3\,\mathrm{cal\,mol^{-1}\,K^{-1}}$ upon adding the conformational term. However, other effects start to come into play at this system size. The global minimum of $C_{14}H_{30}$ and of smaller n-alkanes in the gas-phase always correspond to a linear (unfolded) structure. As intramolecular interactions, in particular London dispersion, become stronger with increasing system size, other conformers will be fa-

vored eventually. For $C_{14}H_{30}$ up to $C_{18}H_{38}$, a competing folded conformer (in which dispersion interactions are maximized) is observed.[437,438] The folded conformers are energetically similar to the respective linear structure but differ strongly in their msRRHO entropy. Depending on the applied theoretical level, either conformation could be the global gas-phase minimum, which makes the choice of S_{ref} in Eq. 4.11 ambiguous and could introduce errors. In the ideal case, the variations between different reference conformers in $\overline{S}_{msRRHO}$ and S_{msRRHO} would cancel and lead to the same conformational entropy regardless of the chosen global minimum. This is observed for $C_{18}H_{38}$ and S_{conf} calculated at the GFN–FF level and would always be the case if $\overline{S}_{msRRHO}$ (see Eq. 4.11) is calculated at the same level as S_{msRRHO}. For $C_{16}H_{34}$ variations between the different theory levels are larger and only the GFN2 conformational entropy for the folded conformer as reference is still within chemical accuracy. Nevertheless, accurate entropies of extremely flexible large alkanes have been consistently obtained for the first time and this can be considered as a major achievement even though some issues for $C_{18}H_{38}$ remain. The detailed reasons for the deviations for the "worst cases" $C_{16}H_{34}$ and particularly $C_{18}H_{38}$ are not fully clear at this point but originate tentatively from the S_{conf} part.

Technical size limitations of our approach should also be noted. The computational cost increases strongly with molecule size at high flexibility and can make the conformational entropy calculation unfeasible for larger molecules. At the GFN2 level, the S_{conf} calculation for $C_{16}H_{34}$ already takes a few hundred hours of computation time, and hence, we did not attempt to calculate $C_{18}H_{38}$ at this level of theory. With the much cheaper GFN–FF method, on the other hand, the entropy for both $C_{16}H_{34}$ and $C_{18}H_{38}$ can still be computed roughly "over night" on a standard CPU node with 14 cores. Somewhat larger (up to 100-200 atoms) but less flexible molecules (*e.g.*, typical drugs, see Sec. 4.4.4) are also feasible at the GFN–FF level due to the shorter MD run times required. Neither of these system sizes can routinely be treated by DFT based MF approaches. In summary, the combination of S_{msRRHO} calculations with the specialized conformational sampling procedure for S_{conf}, and the $\overline{S}_{msRRHO}$ averaging performs excellently and is on par with or even better than complicated and computationally demanding mode based approaches. Improvements of our approach may be necessary for molecules with a very large number of internal rotors at least if absolute values are considered and hence, a beneficial error compensation is not given.

4.4.3. Benchmarking Heat Capacity

Heat capacities and enthalpies (see Eqs. 4.13,4.14) depend less strongly on the ensemble partition function than the entropy. Hence, it is sufficient to calculate C_p and enthalpies $[H(T) - H(0)]$ only for a single converged ensemble without extrapolation. The performance of our approach was evaluated on a subset of the LBH benchmark with 44 experimental heat capacities for linear and branched alkanes at different temperatures between 300 and 500 K. For reference, we again compare with the UM-VT results provided in Ref. 380. Parity plots for the comparison with experimental data are shown in Fig. 4.6 and the corresponding statistical data are given in Tab. 4.3.

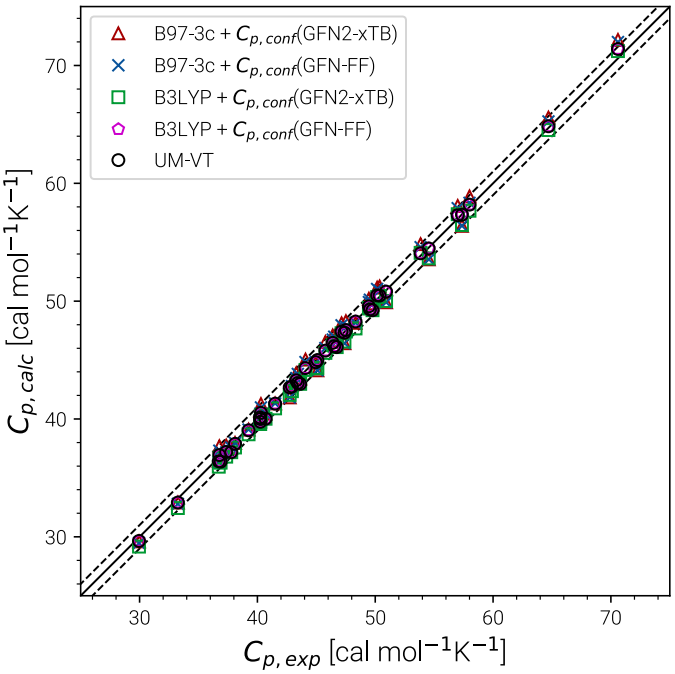

Figure 4.6.: Parity plots for calculated and experimental heat capacities for a subset of the LBH set. Method combinations of B97-3c and B3LYP-D3/def2-TZVP $C_{p,msRRHO}$ values with GFN2–xTB and GFN–FF $C_{p,conf}$ values are shown. UM-VT values were taken from Ref. 380.

Table 4.3.: Mean deviation (MD), mean average deviation (MAD), root-mean-square deviation (RMSD) and standard deviation (SD) for heat capacities obtained at different theoretical levels in comparison to experimental data. All values are given in cal mol^{-1} K^{-1}.

$C_{p,RRHO}$	B97-3c		B3LYP-D3/TZ		UM-VT[a]
$C_{p,conf}$	GFN–FF	GFN2–xTB	GFN–FF	GFN2–xTB	
MD	0.05	0.17	-0.39	-0.11	-0.05
MAD	0.47	0.57	0.47	0.25	0.68
RMSD	0.58	0.69	0.54	0.32	0.78
SD	0.58	0.68	0.38	0.31	0.79

[a]Values taken from Ref. 380.

Excellent performance is achieved for all assessed methods with RMSDs and SDs (much) smaller than 0.7 cal mol^{-1} K^{-1}. In Fig. 4.6, virtually all data points are within an error range of 1 cal mol^{-1} K^{-1}. The choice of the theoretical level used for the msRRHO calculations seems to be less important as both B97-3c and B3LYP-D3 perform well. Looking at the corresponding mean deviations B97-3c tends to slightly overestimate C_p while B3LYP-D3 shows the opposite trend. This is attributed to the choice of the frequency scaling factor and the cut-off value τ, which were adjusted for the computation of entropies. Accordingly, the results could be seen as

further evidence for the conceptional validity of this treatment. At ambient temperature absolute values of heat capacities are smaller than absolute values for entropies. The corresponding conformational contributions are mostly not the accuracy bottleneck for the heat capacities but can be significant at lower temperatures. For example in the LBH subset, the largest $C_{p,conf}$ values are obtained only for the most flexible systems (*n*-heptane, *n*-octane) and even then it accounts only to about 2–3 $\mathrm{cal\,mol^{-1}\,K^{-1}}$. However, it should be noted that the errors in the standard RRHO treatment will quickly exceed the desired 3 $\mathrm{cal\,mol^{-1}\,K^{-1}}$ range.

Temperature Dependence of the Heat Capacity

As $C_{p,conf}$ converges to zero with increasing temperature (all conformers are equally populated for $T \to \infty$), the accuracy of the calculated heat capacity for large T depends mostly on the underlying frequency calculation. *n*-Octane is shown as an example in Fig. 4.7a, in comparison with experimentally derived[439] heat capacities for in the temperature range from 300 to 1500 K. For temperatures below 500 K, the RRHO approach systematically underestimates the C_p values, which is improved by the msRRHO treatment. To reach chemical accuracy for this temperature regime, adding the conformational contribution is mandatory. With increasing temperature the unmodified RRHO value starts to overestimate the experimental C_p. Because the msRRHO treatment always increases the heat capacity in comparison to the RRHO value, no improvement is obtained with our approach for very high temperatures. For *n*-octane at 1500 K this leads to an overestimation of 7 $\mathrm{cal\,mol^{-1}\,K^{-1}}$ in comparison to experiment. However, it should be noted that the high temperature reference values in Fig. 4.7 are derived indirectly from low temperature experimental data[439,440] and hence these data points may have a larger uncertainties than the low temperature ones. In fact, other references can be found that differ from the here shown data and are slightly closer to the computed values.[441] In the chemically important temperature regime of up to 500 K, where our approach is very accurate, a significant conformational contribution to the total C_p value is obtained (for a few examples see Fig. 4.7b). The temperature dependence of $C_{p,conf}(T)$ is very characteristic for each molecular structure and may contain maxima/minima in the curves. Extrema of $C_{p,conf}(T)$ can be associated with large changes of the individual conformer populations and may be interpreted as conformational phase transitions. For a more general review of interpretations of PES related heat capacity features see the work of Wales (Ref. 397). The linear chain-like molecules in Fig. 4.7b (decane, octane and hexanethiol) only have a single maximum in the range 100–200 K. Around 200 K, many folded, higher energetic conformations start to be populated, while at lower temperatures only very linear structures are obtained. The global maximum of $C_{p,conf}$ depends on the molecule specific energetic distribution of the conformers within a given energy window. For example, the CE of hexanethiol and octane consist of about the same number of conformers (150 and 152 structures respectively within 6 $\mathrm{kcal\,mol^{-1}}$), but differ with regard to their relative conformational energies. Molecular characteristics become even more pronounced for complicated molecules, *e.g.*, tamiflu and penicilin , where often multiple extrema are obtained for $C_{p,conf}(T)$ (see Fig. 4.7b).

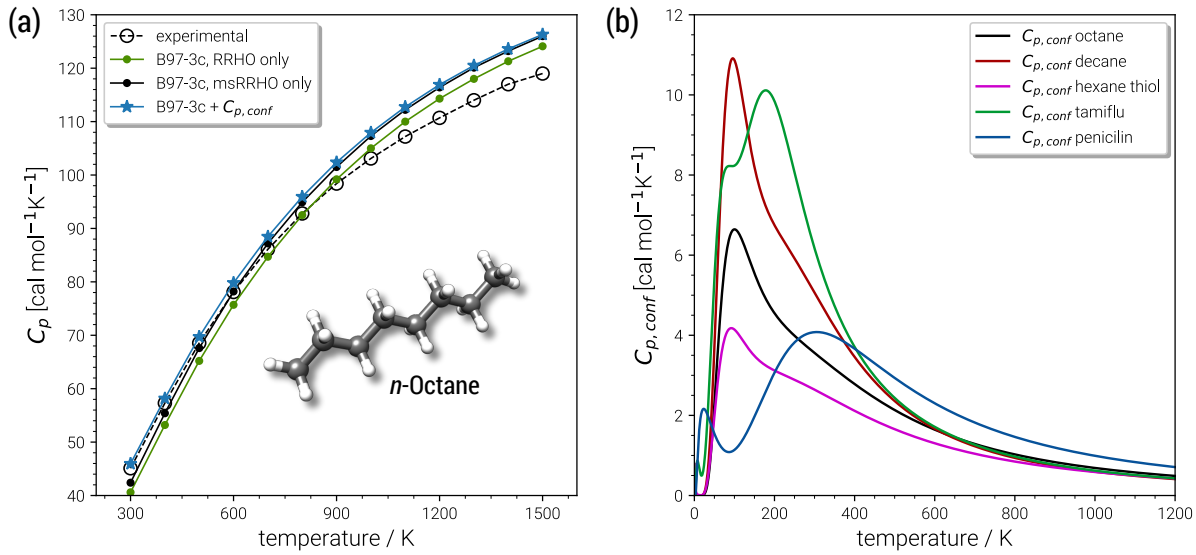

Figure 4.7.: **(a)** Heat capacities calculated for *n*-octane in the temperature range 300 to 1500 K. **(b)** Temperature dependence of the conformational heat capacity shown for octane and other example molecules from the AS23 and CD25 sets. (ms)RRHO values correspond to the B97-3c level and CE were obtained at the GFN2–xTB level.

4.4.4. Case Studies

Drug Molecules

After demonstrating the excellent performance of the presented approach to calculate absolute entropies in section 4.4.2, we now turn our attention to biochemically more important systems. The CD25 set is introduced, containing 25 commercial drug molecules with 28 to 98 atoms. For these molecules no experimental entropy and C_p values are available to compare with. Nonetheless also a purely theoretical investigation of the CE and respective entropies may yield important insights. Note that a comprehensive evaluation of the entropy for such important molecules with a highly accurate method is missing in the chemical literature.

Due to their similar size and elemental composition, similar S_{conf} values may be expected for typical drugs. This is not the case as can be seen from the entropies calculated for the CD25 set, shown for the GFN2–xTB and GFN–FF levels in Fig. 4.8. Conformational entropies in the CD25 set range from close-to-zero to over 20 cal mol^{-1} K^{-1}. The reason for this is rooted in the very diverse and complicated PES of the molecules. Compared to the smaller and chemically rather similar molecules in the LBH and AS23 set, the molecules in the CD25 set show a variety of functional groups and intramolecular non-covalent binding motifs. This leads to a fine balance of covalent and non-covalent forces which characteristically shape the overall PES. Certain energy basins (a collection of related minima), for example, could be strongly favored because of intramolecular hydrogen bonding and thus reduce the overall number of energetically accessible minima. In such cases, an accurate description of the respective potentials is required and the computed S_{conf} value is strongly dependent on the underlying theoretical method.

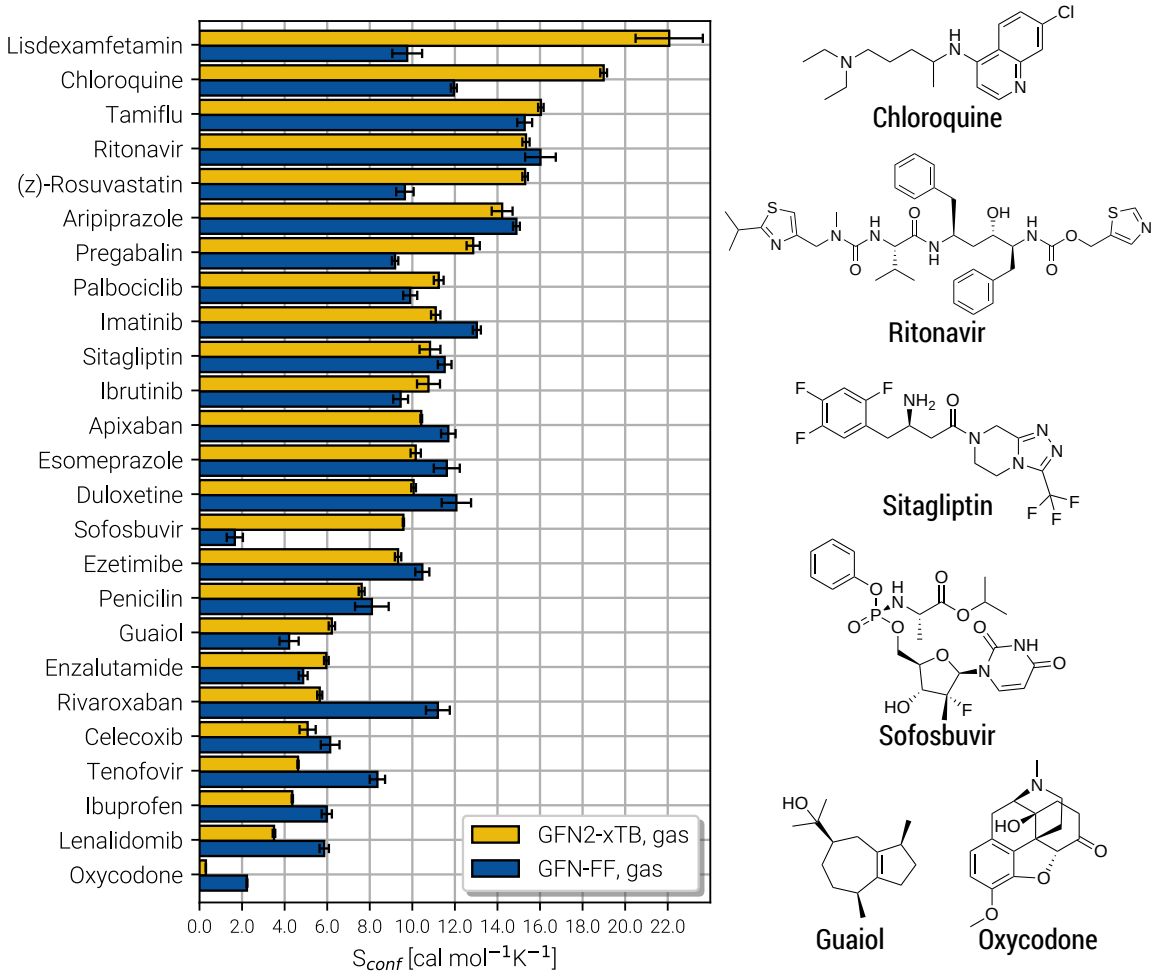

Figure 4.8.: Calculated S_{conf} values for a set of 25 clinical drug molecules at the GFN2–xTB and GFN–FF levels of theory sorted according to increasing value. Averaged values (shown as horizontal bars) and their standard deviations (shown as errors) have been determined by multiple executions of the above described algorithm, as described in the text below. On the right side Lewis structures of some of the molecules are shown (see Appendix A3 for all molecules).

With a few notable exceptions, the conformational entropies calculated with GFN2–xTB and GFN–FF only differ by 1 to 2 cal mol^{-1} K^{-1} and therefore provide the same semi-quantitative description of the PES. The exceptions are cases in which GFN2 produces much larger CE (chloroquine, lisdexamfetamin, pregabalin, rosuvastatin, sofosbuvir) than GFN–FF, or *vice versa* (rivaroxaban, tenofovir). For the most rigid molecule (oxycodone), only a single conformer is significantly populated ($p_i = 0.98$ at 298 K) at the GFN2 level, while three conformers are populated at the GFN–FF level, resulting in a larger entropy. For the other cases with larger differences between both methods, the interpretation is difficult because of a large number of significantly populated structures (about hundreds) in the CE. A better understanding would be provided by an improved theoretical description, *i.e.*, the ensemble calculated by DFT or WFT

but this is unfeasible due to the extremely high computational effort. Instead, one could refer to other qualitative descriptors when interpreting conformational entropies at a low theoretical level. Because the entropy is correlated with molecular structural features, one such descriptor could be the flexibility measure ξ_f, which is used for determining the simulation length settings in CREST.[33] This comparison of ξ_f and the S_{conf} is shown in Fig. 4.9 and in Appendix A3. Note that conformational entropies must be normalized to system size (number of atoms N_{at}) in order to be comparable in between molecules.

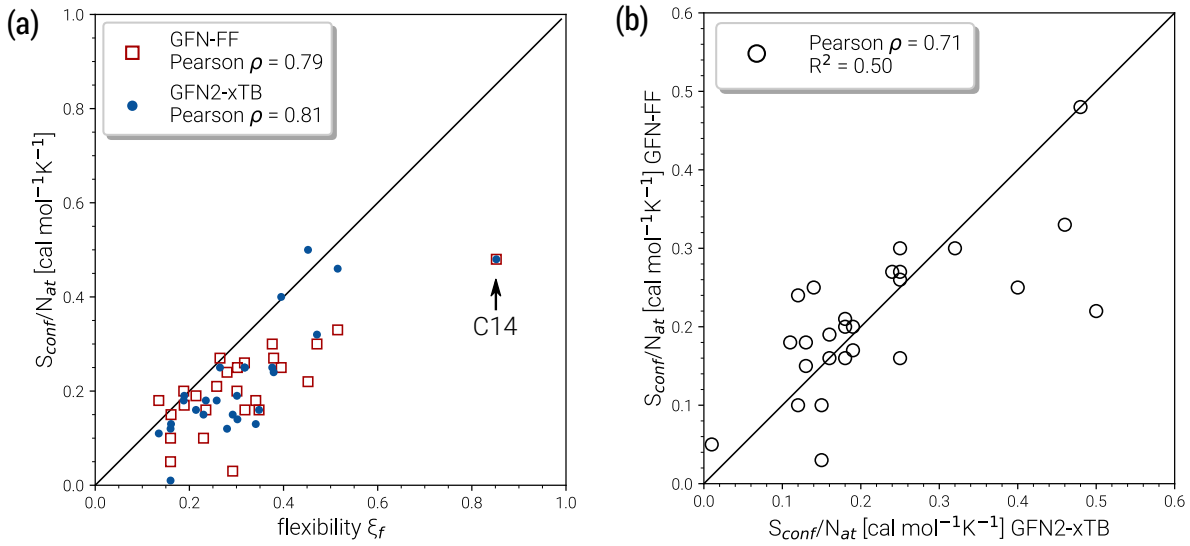

Figure 4.9.: Correlation plots for the molecules of the CD25 set. **(a)** Correlation between S_{conf}/N_{at} and the empirical flexibility measure ξ_f. **(b)** Correlation of the S_{conf}/N_{at} values at GFN–FF and GFN2–xTB level. The respective Pearson correlation coefficients ρ are shown in the legends.

Both methods show a relatively high correlation with the empirical flexibility ξ_f in (Fig. 4.9a). The only outlier here is tetradecane, denoted as "C14" in the figure, which is chemically different from the drug molecules and was added only as an upper bound reference for the flexibility. When quantified via the well-known Pearson correlation coefficient ρ, it can be seen that GFN2–xTB ($\rho = 0.81$) corresponds slightly better with ξ_f than GFN–FF ($\rho = 0.79$). This indicates a better description of the few critical cases mentioned above at the tight-binding level. The correlation of S_{conf}/N_{at} between the two methods (Fig. 4.9b, $\rho = 0.71$) again shows the intrinsic theory level dependence of the configurational entropy but is devoid from any deeper interpretation. Nonetheless, these examples demonstrate that the conformational entropy can be nicely correlated with purely structure based features of an ensemble or even empirical descriptors, which is why schemes such as the MIE[406] and MIST[408] have been proven to work comparatively well.

Finally, the CD25 set was employed to evaluate the robustness and reproducibility of the presented approach. As discussed above the stochastical nature of the MD runs leads to slightly varying results for different runs started on the same input structure. Hence, all of the 25

molecules were run several times in repetition and averaged to obtain S_{conf} and its standard deviation (SD) shown in Fig. 4.8. On average over the 25 systems, GFN2–xTB and GFN–FF yield SD values of 0.25 cal mol^{-1} K^{-1} and 0.35 cal mol^{-1} K^{-1} respectively. The only significantly larger SD of 1.6 cal mol^{-1} K^{-1} is obtained for the lisdexamfetamin molecule at GFN2–xTB level, which results from a large and complicated CE leading to convergence problems in S'_{conf}. In general GFN2–xTB has the more accurate PES of the two methods and produces more consistent results. Both GFN2–xTB and GFN–FF show reproducibility errors much below chemical accuracy and hence are appropriate for routine computations of S_{conf}. The much shorter computation times of GFN–FF might favor its default application for large systems and also enables the averaging over multiple entropy calculations to eradicate statistical differences (which would be rather costly at the GFN2–xTB level).

Chemical Applications

In this last section we give a few chemical examples, where absolute entropies are used to compute reaction entropies and Gibbs free energies.

Adsorption processes are important for a variety of applications, such as heterogeneous catalysis[442] where the entropy change can be measured via calorimetric experiments. Here, a rather well studied class of reactions is the adsorption of n-alkanes onto zeolites.[443] As an example the adsorption entropy of n-butane, n-pentane, and n-hexane (Fig. 4.10) in a H-ZSM-5 zeolite cut-out was calculated with GFN–FF.

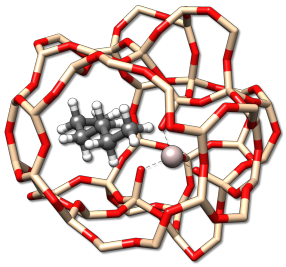

Figure 4.10.: The n-hexane molecule adsorbed by a H-ZSM-5 zeolite. Hydrogen atoms used for the saturation of the zeolite have been omitted for better visibility.

For a given zeolite structure cut-out (*e.g.*, obtained from a crystal structure and saturated with hydrogen atoms) thermodynamic properties can be obtained with the (ms)RRHO approach. Sampling of the configurations in CREST then simply requires some additional geometrical constraints, as was discussed in previous work.[33,444] This is necessary because the zeolite chunk shall mimic a solid and its structure would be strongly deformed or even broken by the metadynamic simulations and geometry optimizations at GFN level. The configurational problem is of course complicated by the combinatorical nature of different conformers at different adsorption sites, but in the present case the total system size is small enough to not pose major problems. Adsorption entropies are directly calculated from absolute entropies by

$\Delta S = S_{alkane/zeolite} - S_{alkane} - S_{zeolite}$ (see Tab. 4.4) and assessed with respect to experimental values.

Table 4.4.: Adsorption entropies (in cal mol^{-1} K^{-1}) for small linear alkanes on H-ZSM-5 zeolite cut-outs, calculated fully at the GFN–FF level of theory. Experimental adsorption entropies were obtained from Ref. 443.

adsorbed molecule	ΔS_{msRRHO}	ΔS_{conf}	$\Delta S_{ads,calc.}$	$\Delta S_{ads,exp.}$
n-butane	-34.1	3.1	-31.0	-24.9
n-pentane	-36.5	4.1	-32.4	-28.2
n-hexane	-38.1	2.8	-35.3	-28.9

The final calculated $\Delta S_{ads,calc.}$ shows deviations of only 4.2 to 6.4 cal mol^{-1} K^{-1} compared to experiment and show the same qualitative trend of adsorption strength (butane < pentane < hexane). While this trend is also reproduced already by S_{msRRHO}, it is important to notice that the configurational contribution accounts for roughly 10 % of the overall adsorption entropy and furthermore shifts ΔS_{msRRHO} in the direction of the experimental value. Because the zeolite is identical for all structures and configurations, all msRRHO entropies are similar and the term $\overline{S}_{msRRHO}$ consequently is $\ll$1 cal mol^{-1} K^{-1}. Therefore the main part of ΔS_{conf} can be attributed to S'_{conf} and qualitatively interpreted. Here, *n*-butane has the smallest amount of conformers but many configurations (adsorption orientations) in the zeolite while it is *vice versa* for *n*-hexane, leading to a similar contribution of $\Delta S_{conf} \approx 3$ cal mol^{-1} K^{-1} in both cases. For *n*-pentane on the other hand, both the conformational and configurational space are large and hence it shows the largest ΔS_{conf} value of the three systems. The calculated $\Delta S_{ads,calc.}$ are in very good agreement with experiment, considering that all results were obtained at a cost efficient force field level and none of the values exceed a deviation of 2 kcal mol^{-1} at 298 K. Note that the full calculation for each of the final ΔS_{ads} values only took about 1.5–2 h on a standard desktop computer (4 cores on a Intel i7-7700K 4.2 GHz CPU).

A more common usage for S_{conf} is to improve the calculation of reaction free energies. The conformational entropies and enthalpies are converted to ensemble free energies G_{conf} *via* the usual relation $G = H - TS$ and can be added directly to the G_{msRRHO} values of all reactands and products of the reaction. In general, a significant change of the DOF in the course of the reaction can cause significant entropic effects and a non-negligible effect on the reaction free energy.

Three examples (**A**, **B**, and **C**) are shown in Fig. 4.11 and the corresponding reaction energy differences are shown in Tab. 4.5.

Reaction **A** is the cyclization of a 1,5-diene into the perfume molecule β-georgywood.[445] Ring-closure reactions are often associated with a decrease of DOF, and hence an entropic destabilization is expected. This view is supported by the computed free energies, where the addition of ΔG_{conf} decreases the reaction free energy from -10.3 kcal mol^{-1} to -8.7 kcal mol^{-1}. For the typical "chemical accuracy" of 1 kcal mol^{-1}, adding the conformational term would

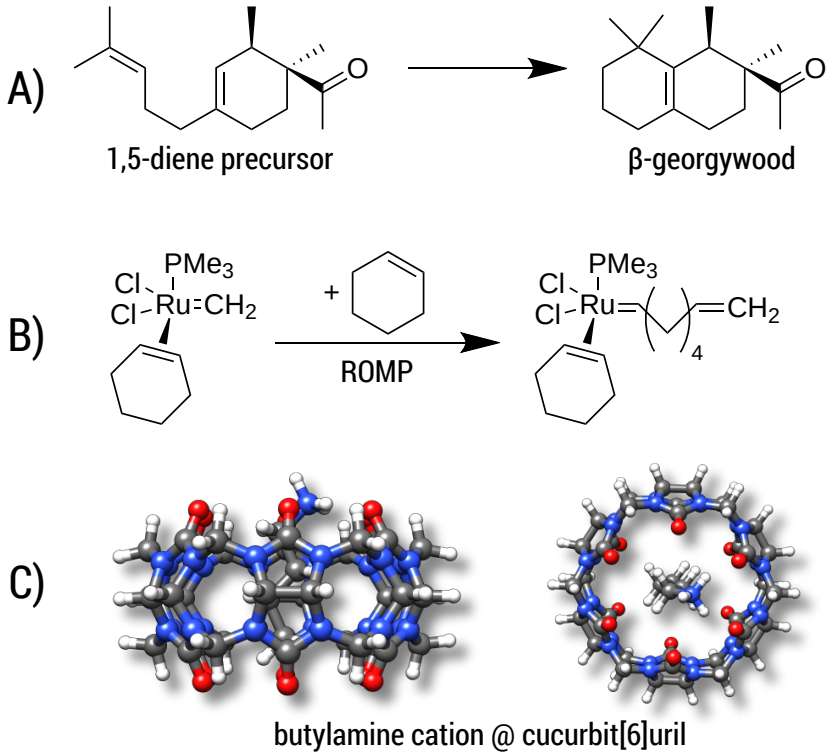

Figure 4.11.: Example reactions with large entropic contributions. A) cyclization of a 1,5-diene to the β-georgywood compound, B) simplified catalytic reaction of a ring-opening metathesis polymerization (ROMP), C) complexation of butylammonium in cucurbit[6]uril.

Table 4.5.: Energy differences for the reactions shown in Fig. 4.11. All values are given in kcal mol^{-1} and were obtained at the B97-3c level with conformational contributions calculated at GFN2–xTB level. Free energies correspond to 298.15 K.

	reaction energies		
reaction	ΔE	ΔG	$\Delta G + \Delta G_{conf}$
A	-15.0	-10.3	-8.7
B	-8.1	4.6	2.8
C	-82.0	-64.8	-64.3

therefore be necessary. Note, that ring-closures are common in many syntheses and biochemical processes (*e.g.* terpene chemistry,[446] or, as an example from a previous section, the synthesis of oxycodone[447]) and therefore will profit from a better description by our method.

Reaction **B** is a simplified catalytic reaction of a ring-opening metathesis polymerization (ROMP).[184] ROMP was pioneered by the groups of Chauvin, Grubbs and Schrock and are among the most important catalytic reactions in industrial chemistry.[448,449] The reaction free energy balance of **B** is positive as a result of the sterically undemanding PMe_3 ligand, but nonetheless the influence of G_{conf} is nicely demonstrated. Here, due to a loss of DOFs (two

reactants form one product molecule), ΔG becomes initially positive, which is counteracted by a DOF gain in G_{conf} of the product. The effect of the ensemble treatment has the same origin as in the ring-opening reaction **A**, but in this case favors the formation of the product by about 1.8 kcal mol^{-1}. This example furthermore shows the capability of GFN2–xTB (and GFN–FF), which can be routinely be applied to transition-metal containing systems.

The influence of configurational entropy can also be studied for non-covalent associations. Reaction **C** shows the binding of butylammonium in cucurbit[6]uril.[450,451] Binding affinities for small cations in cucurbiturils are well studied,[452] but for more flexible guest molecules such as butylammonium, entropic effects may become important. The association free energy changes from -64.8 kcal mol^{-1} to -64.3 kcal mol^{-1} upon addition of ΔG_{conf} in the gas phase. On first sight, the increase of about 0.5 kcal mol^{-1} seems negligible compared to the large overall value of about -64 kcal mol^{-1}. However, the latter value is quenched in solution[450,451] to about -6.9 kcal mol^{-1} indicating that under more realistic conditions ΔG_{conf} is indeed relevant.

All the examples discussed in this subsection have been modelled in the gas-phase, but the extension to solutions is easily possible by using implicit solvation models. Inclusion of solvation effects will modify the PES and therefore produce different ensembles (and conformational entropies) than in the gas-phase. A direct impact of this would be noticeable, *e.g.*, for phase-partition coefficients like logKow, which strongly depend on the respective ensemble.[453] Technically, such calculations are straightforward and are investigated currently in our laboratory.

4.5. Conclusion

An automated workflow for the calculation of absolute molecular entropies is presented. The molecular entropy is a fundamental thermodynamic quantity necessary for a complete understanding of molecular interactions. The main component of the absolute entropy is usually obtained from vibrational frequency calculations in the RRHO approximation, which for medium sized molecules (50–100 atoms) often underestimates anharmonicities for low-frequency modes and is missing configurational contributions arising from many accessible low-energy conformations. In the presented approach both sources of error are treated by a separation of the molecular entropy into a configurational (conformational) part and the entropy arising from translational, rotational, and vibrational degrees of freedom. For the latter, vibrational frequencies were obtained at the B97-3c and B3LYP-D3/def2-TZVP DFT level, employing a modified and scaled RRHO approximation (termed msRRHO) with two adjustable parameters τ and ν_{scal}. The conformational entropy is calculated from an ensemble of conformers using the well known Gibbs-Shannon entropy formula (S'_{conf}) and an population average over individual msRRHO contributions of the conformers ($\overline{S}_{msRRHO}$). We here make use of the fast and accurate GFN–FF and GFN2–xTB methods for the generation and energetic ranking of structures, driven by the recently introduced CREST program. The entire procedure is designed to work with only a few simple steps and minimal user input, which makes it routinely applicable to a broad range of systems.

The presented workflow was tested on a set of 62 experimental molecular gas phase entropies. An excellent performance (better than the chemical accuracy of $3\,\mathrm{cal\,mol^{-1}\,K^{-1}}$) was observed with MADs ranging from 0.73 to 0.92 $\mathrm{cal\,mol^{-1}\,K^{-1}}$ and SDs from 1.08 to 1.30 $\mathrm{cal\,mol^{-1}\,K^{-1}}$ respectively, depending on the combination of the DFT method with either GFN2–xTB or GFN–FF. Heat capacities were assessed on a set of linear and branches alkanes at different temperatures. The MAD and SD values are with $0.5\,\mathrm{cal\,mol^{-1}\,K^{-1}}$ even smaller than for absolute entropies but increase at very high temperatures $> 800\,\mathrm{K}$. The presented method performs better than related yet computationally significantly more costly approaches and to our knowledge provides the smallest errors for molecular entropies ever reported in the literature. This includes large, extremely flexible *n*-alkanes up to octadacene for which an unprecedented accuracy for the absolute entropy in comparison to experiment of about 5% was obtained.

Biochemically important systems and chemical applications were discussed on the basis of set of 25 drug molecules and four reaction examples, including the calculation of adsorption entropies, two reaction free energies and a non-covalent association free energy calculation. For the drug molecules, a correlation of molecular flexibility and the entropy was observed. The examples revealed a significant contribution of the configurational terms to the overall free energy, often exceeding the magnitude of chemical accuracy. In the future, a more thorough study of these effects across a wide range of chemical reactions is desirable.

In general, GFN2–xTB was found to provide (as expected) a more consistent description of the PES and hence the conformational entropy than GFN–FF. However, as calculations of S_{conf} tend to get very expensive for larger systems at GFN2–xTB or higher theoretical levels, GFN–FF is strongly recommended as the standard approach in routine treatments on common desktop computers. In theory, the basic components of the proposed scheme are systematically improvable by a better description of the PES. The modular partition of the absolute value into ro-vibrational and configurational parts enables a convenient replacement of the different methods, which provides a starting point for future studies. This also includes the extension to implicit solvation models that will allow to investigate molecular entropy differences between the gas-phase and solution or between different solvents.

Acknowledgements

This work was supported by the DFG in the framework of the “Gottfried Wilhelm Leibniz-Prize” awarded to S.G. .

Supporting Information

Some additional supporting information can be found in Appendix A3. The open-source CREST program and source code can be downloaded from `https://github.com/grimme-lab/crest`. Input structures used in this chapter can be downloaded from `https://github.com/grimme-lab/mol-entropy`.

Part IV.

Application of Efficient Quantum Mechanical Methods to Chemical Problems

In Part IV applications of low-cost quantum mechanical methods are discussed. One of the design purposes of GFN*n*–xTB methods is the calculation of vibrational frequencies, which are used in the following chapters for the computation of thermal free energy contributions G_{trv}^{T} and vibrational modes for the simulation of infrared (IR) spectra. However, it is shown that the SQM methods can also be used for some "off-target" properties, *i.e.*, the calculation of quantities which are *not* considered in the construction or parametrization of GFN*n*–xTB, but still can be calculated with reasonable accuracy.

Chapter 5 is dedicated to the calculation of gas-phase IR spectra from the so-called second harmonic approximation. Vibrational frequencies are obtained from the second derivatives of the energy with respect to nuclear coordinates. IR intensities are calculated from derivatives of the molecular dipole moment along the respective modes. Calculations of these properties can be conducted at a QM, SQM, or even FF level. This is of particular importance for IR spectra prediction in the context of unknown compound identification procedures,[44] where hundreds of expensive frequency calculations for a variety of systems are required and SQM or FF methods have the advantage. The performance for IR spectra simulation of GFN*n* SQM and FF methods and the composite DFT method B3LYP-3c is evaluated in comparison with over seven thousand experimental gas-phase references. As a general trend, it is found that IR spectra quality decreases here in the order B3LYP-3c $\gg$ GFN2–xTB $\gtrsim$ GFN1–xTB $>$ GFN–FF, which is consistent with the quality of atomic charges and dipole moments produced by these methods. With regards to frequencies, a well-known problem are missing anharmonicities arising from the harmonic approximation as outlined in Chapter III. To correct for this model deficiency, a new correction scheme based on modification of atomic masses is introduced.

Chapters 6 and 7 present the calculation of acid dissociation constants in water (pK_a values) from dissociation free energies. A protocol is employed that combines computations of the total energy at DFT level, thermal contributions G_{trv}^{T} at SQM level and solvation free energies from continuum solvation models. The resulting acid dissociation free energies can be fitted to experimental pK_a with the so-called free energy relationship (FER), most commonly applied in a linear form (LFER). An initial attempt of LFER-based pK_a predictions was made with this protocol in a collaborative study with the Novartis AG in context of the SAMPL6 blind challenge.[45,46] The combination of high level double hybrid DFT calculations with ro-vibrational free energy contributions from GFN1–xTB and COSMO-RS implicit solvation[218,454] herein provided the best results of all SAMPL6 contestants. In a second study (Chapter 7) the pK_a protocol is revised and generalized. Here, higher order FER are used to provide more flexibility with regards to a single input parameter (the dissociation free energy). Several different DFT methods are tested in comparison with over three hundred experimental pK_a values. Rather independently of the employed functional it is found that low errors below 1 pK_a unit can be achieved. The special purpose GFN*n*–xTB methods, on the other hand, were not designed to yield good absolute energies and consequently good heterolytic dissociation free energies. Upon correction of the latter with an empirical energy term, pK_a values calculated entirely at the GFN level are shown to be of use at least for qualitative interpretations or initial screening and

are broadly applicable to arbitrarily composed systems.

For all three chapters in this part, CREST is used to sample the low-energy chemical space. While for the IR spectra calculations this involves only the screening of conformers, for pK_a calculations also initial (de-)protonation sites and tautomers are generated. Significant influence of the molecular conformations are observed, *e.g.* for pK_a values of drug like molecules. The final objective for all studies presented here is to formulate efficient QM workflows that do not rely on proprietary software and provide reasonable accuracy even on a SQM level.

5. Comprehensive Assessment of GFN Tight-Binding and Composite DFT Methods for Calculating Gas-Phase IR Spectra

Philipp Pracht,[*] David F. Grant,[†] and Stefan Grimme[*]

Received 24th of August 2020, Published online 14th of October 2020

Reprinted (adapted) with permission from[‡]
Pracht, P.; Grant, D. F.; Grimme, S. *J. Chem. Theory Comput.* **2020**, *16*, 7044–7060.

DOI 10.1021/acs.jctc.0c00877

Own manuscript contribution

- Performing all calculations
- Fitting atomic mass scaling factors
- Writing the `newspecmatch` code
- Interpretation of the computed data
- Writing the manuscript

[*]Mulliken Center for Theoretical Chemistry, Institut für Physikalische und Theoretische Chemie, Rheinische Friedrich-Wilhelms-Universität Bonn, Beringstraße 4, 53115 Bonn, Germany
[†]Department of Pharmaceutical Sciences, University of Connecticut, Storrs, Connecticut 06268, United States
[‡]Reproduced with permission from the American Chemical Society.

Abstract

Vibrational spectroscopy is a valuable and widely used analytical tool for the characterization of chemical substances. We investigate the performance of semiempirical quantum mechanical GFN tight-binding and force field methods for the calculation of gas-phase infrared spectra in comparison to experiment and low-cost (B3LYP-3c) density functional theory. A data set of 7247 experimental references was used to evaluate method performance based on automatic spectra comparison. Various quantitative spectral similarity measures were employed for the comparison between theory and experiment and for determining empirical scaling factors. It is shown that the scaling of atomic masses provides an accurate yet simple alternative to standard global frequency scaling in DFT and semiempirical calculations. Furthermore, the method performance for 58 exemplary transition metal complexes was investigated. The efficient DFT composite method B3LYP-3c, that was introduced in the course of this work, was found to be excellently suited for general IR spectra calculations. The GFN1- and GFN2–xTB tight-binding methods clearly outperformed the PMx competitors. Conformational changes were investigated for a subset of the data and are found to have a mediocre strong influence on the simulated spectra suggesting that the corresponding elaborate sampling steps may be neglected in automated compound identification workflows.

5.1. Introduction

Rotational-vibrational spectroscopy is one of the most common analytical tools for the characterization of chemical substances.[455] The analysis of experimental infra-red (IR) or Raman spectra gives detailed insight into molecular structure and can be used, *e.g.*, for analytical compound identification purposes.[44,456] Theoretical, mainly quantum chemistry calculations of vibrational spectra including assignment of the involved normal modes have a long and very successful history.[457,458] Vibrational spectra can be calculated by various classical atomistic and quantum chemical (QM) methods, of which density functional theory (DFT) is the most common today. Still, even with modern theoretical and technological advances, the calculation of vibrational frequencies can be prohibitively expensive for moderately sized (100-200 atoms) molecules. Accurate and fast calculation of IR spectra and the related thermostatistical free energy corrections at a reduced theoretical level (see *e.g.* Ref. 204) is therefore desirable. Progress has been made in this field with classical force fields (FF), but their errors are much larger and less systematic than at a QM level.[459,460] Semiempirical quantum chemical methods (SQM) can provide a good alternative because they are considered to bridge the gap between FFs and QM, both in terms of computational cost and accuracy. In particular methods of the recently developed GFN family[36] are promising candidates since they are designed to yield good Geometries, Frequencies, and Noncovalent interaction energies.

Vibrational frequencies are commonly calculated within the harmonic approximation and depend on the second derivatives of the energy with respect to atomic displacements and ad-

ditionally on the atomic masses. The comparison of harmonic theoretical with experimental (fundamental) frequencies leads to systematic errors also in higher rung DFT treatments.[381,461] A typical countermeasure is to linearly scale the harmonic vibrational frequencies (HVF) obtained from QM (or FF) calculations, where different levels of theory (*e.g.*, DFT functionals, basis sets) require varying scaling factors.[381–383,457] In the case of hybrid functional DFT the HVF error can be attributed mainly to the employed Fock exchange, which consistently overestimates frequency values. For FF or SQM, the errors are less systematic and global frequency scaling works less well than in (hybrid) DFT so that other means to improve the accuracy have been investigated.[462–464] Introducing more adjustable parameters provides a higher degree of empirical freedom, but also many more data points are required for a robust fit. Multiparameter frequency scaling methods have been proposed in the past and yield more accurate results than simple linear and uniform global scaling.[386,463,465–470] More recently, machine learning approaches have successfully been combined with molecular dynamics and QM calculations for the computation of IR spectra and anharmonic frequency corrections.[471,472]

Due to the intrinsic relation of molecular vibrational motions to the involved atomic masses, the latter could also be used as a parameter in calculations. Here, we explore the scaling of atomic masses as an alternative to the standard global scaling of the HVF. Atomic mass scaling has been introduced by Irikura[473] for diatomics and is related to the approach of Pulay and coworkers, who proposed an internal coordinate specific scaling of force constant matrix elements instead of global frequency scaling.[385,474]

In this work, we used a data set of 7247 experimental gas-phase IR spectra (obtained from the NIST chemical database[434]) to evaluate the performance of SQM methods of the GFN family (GFN1-, and GFN2–xTB)[38,39] for IR spectra calculations. In addition, the performance of the recently introduced, non-electronic variant GFN–FF is investigated as a representative of an advanced classical FF.[80] This involves the introduction of a similarity measure for the quantitative comparison between experimental and theoretical spectra. Furthermore, a low-cost composite DFT method based on the B3LYP density functional[55] similar to the the well established PBEh-3c[158] composite method is proposed and tested. For recent studies with our GGA-based low-cost method B97-3c[180] for IR spectra calculations see Refs. 475,476.

In a recent study,[464] Henschel *et al.* pursued a similar approach to evaluate the performance of several FFs for the calculation of IR spectra. In their work two different metrics (Pearson correlation coefficient, Spearman rank correlation) were used for the evaluation. For comparability the same spectral similarity scores were used here, plus two metrics (match score, Euclidian norm) we originally based this work on. In the literature, other measures have been proposed and adapted for various purposes.[47,477] The different similarity scores have different advantages and disadvantages but it has been suggested[464] that they can be used complementarily. Still it remains an open question whether all of the similarity measure mentioned above are equally suitable for the quantitative comparison of IR spectra. The long term goal of this work is to provide the theoretical basis for automated high-throughput workflows in analytical chemistry (unknown compound identification) applications.

5.2. Theory

5.2.1. Calculation of IR Spectra

By a Taylor expansion of a general molecular potential energy surface, vibrational motions can be related to force constants. Force constants are the second derivatives of the total energy with respect to nuclear displacements.[4] In *ab initio* (or semiempirical) calculations vibrational frequencies and the corresponding eigenfunctions (normal modes) are most commonly obtained within the harmonic approximation, where the force constants are contained in the mass-weighted Hessian matrix

$$\mathbf{F}_{ij}^{(m)} = \frac{1}{\sqrt{m_i m_j}} \left(\frac{\partial^2 E}{\partial R_i \partial R_j} \right) \ . \tag{5.1}$$

In Eq. 5.1 E is the total molecular energy, $R_{i/j}$ is the displacement in Cartesian space and $m_{i/j}$ are the masses of nuclei i and j respectively. The second derivatives may be obtained analytically, or numerically from the first-order derivative, *i.e.*, the nuclear gradient. Diagonalization of the matrix $\mathbf{F}^{(m)}$ according to

$$\mathbf{F}^{(m)}\mathbf{Q} = \epsilon\mathbf{Q} \ , \tag{5.2}$$

yields the diagonal matrix ϵ that contains the eigenvalues of $\mathbf{F}^{(m)}$ and $\mathbf{Q}$ are the normal modes in Cartesian space. The vibrational frequencies ν are given by

$$\nu_p = \frac{1}{2\pi}\sqrt{\epsilon_p} \tag{5.3}$$

for mode p.

The IR absorption intensity for each fundamental transition is associated with the change of the molecular dipole moment μ along the normal mode coordinate q_i[478] with the leading term $\frac{\partial\mu}{\partial q_i}$. Very accurate calculations of intensities require knowledge also of higher order terms,[479] however, they are approximately (the so-called double harmonic approximation) proportional to the squared derivative $\left(\frac{\partial\mu}{\partial q_i}\right)^2$[480] which is is used in this work. Since intensities depend on the quality of the calculated dipole moments their description is rather sensitive to the employed level of theory (*e.g.*, the quality of the AO basis set) and several approaches for the improved calculation of IR intensities exist.[481–483] In particular, the mode localization and intensity tracking approaches pioneered by Reiher and coworkers[484–487] seem to be capable tools for the improved calculation and interpretation of vibrational spectra. However, the fine details of IR spectra calculations are not the subject of the present work. Instead, we focus here on a reasonably accurate and straightforward description of the more typical case of a low-resolution / large molecule.

5.2.2. Comparing IR Spectra

In general, a molecule with n nuclei has $3n-6$ normal modes ($3n-5$ for linear molecules) which in principle can be compared with an experimental reference IR spectrum. For the latter, fundamental frequencies need to be extracted and assigned to the respective theoretical modes. This can become tedious for medium sized molecules where modes in the *fingerprint* regime, $<$ 1500 cm^{-1} often have small intensities and strongly overlap due to their line width. To evaluate larger compounds and a large number of spectra, it is therefore neither sufficient nor efficient to compare only selected fundamental frequencies. Instead, metrics for a quantitative comparison of two (entire) spectra have to be developed and tested.

Experimental spectra typically consist of a set of k equidistant frequency data points with respective intensities. In order to be comparable the spectra need to be normalized. Different procedures are possible, but in this work all spectra will be normalized according to

$$\sqrt{\sum_{i}^{k} u_i \mathrm{d}u} \stackrel{!}{=} 1 \ , \tag{5.4}$$

where u can be considered as a k-dimensional vector, u_i is the intensity of the i-th point in the spectrum and $\mathrm{d}u$ is the distance between the points u_i and u_{i+1} (here set to 1.0 cm^{-1}), *i.e.*, the normalization is done via a Riemann integration. The square-root dependence in Eq. 5.4 was chosen to increase the weight of small intensities relative to larger intensities. Normalizations to single signals, *e.g.*, the largest peak of the spectrum, should be avoided because relative intensities and frequencies are strongly dependent on the theoretical level and hence it cannot always be ensured that the same peak is selected. For theoretical spectra, the frequencies and intensities of the vibrational modes are available as isolated signals ('*stick spectrum*') and first have to be expanded to the same spectral domain as the experimental data. This is achieved by employing a Lorentzian line shape function for each mode

$$\phi_p(\nu) = I_p \left(1 + \frac{\nu_p - \nu}{0.5w}\right)^{-1} \ , \tag{5.5}$$

where ν_p is the position (calculated frequency) of the mode p, I_p is its intensity and w is the full width at half maximum (FWHM). Typical values employed for the FWHM in theoretical studies range from 20 to 40 cm^{-1}, whereas the average line width in experimental spectra was determined as 24 cm^{-1}.[488] The simulated spectrum is then simply given by the sum of all the Lorentzian functions over the modes

$$\Phi_{norm}(\nu) = I_{norm} \sum_{p}^{N_p} \phi_p(\nu) \ , \tag{5.6}$$

with the normalization constant I_{norm}. Analogously to Eq. 5.4, the spectrum is normalized by $\sqrt{\int \Phi_{norm} \mathrm{d}\nu} \stackrel{!}{=} 1$ and in this form is directly (point wise) comparable to experimental data.

Hence, in order to match experimental and theoretical spectra, a vector v with the same number of points as u is constructed from Φ_{norm}.

In this study four different spectral similarity measures have been investigated. The first is a simple match score (r_{match}),

$$r_{match} = \frac{\left(\sum_i^k u_i v_i\right)^2}{\left(\sum_i^k u_i^2\right)\left(\sum_i^k v_i^2\right)} \ , \tag{5.7}$$

where u and v are the k-dimensional vectors obtained for the two compared spectra. The r_{match} corresponds to a Cauchy-Schwarz inequality in $\mathbb{R}^k$ dimensional Euclidian space which essentially is a simplified overlap. r_{match} values range from $0 \leq r_{match} \leq 1$, where unity denotes a perfect match.

The second measure used is the Euclidean norm (r_{euclid}),

$$r_{euclid} = \left(1.0 + \frac{\sum_i^k (u_i - v_i)^2}{\sum_i^k (v_i)^2}\right)^{-1} \ . \tag{5.8}$$

In the literature two additional measures have been employed for IR and Raman spectra comparisons.[489–492] One is the Pearson correlation coefficient ($r_{pearson}$), which is similar to the r_{match},

$$r_{pearson} = \frac{\sum_i^k (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_i^k (u_i - \bar{u})^2}\sqrt{\sum_i^k (v_i - \bar{v})^2}} \ , \tag{5.9}$$

with the mean values $\bar{u}$ and $\bar{v}$ for u and v. Both the r_{match} and $r_{pearson}$ are linear correlation measures that are based on the Cauchy-Schwarz inequality. The other (nonlinear) measure that has often been employed is the Spearman rank correlation coefficient ($r_{spearman}$)

$$r_{spearman} = 1.0 - \frac{6\sum_i^k (rg(u_i) - rg(v_i))^2}{k\,(k^2 - 1)} \ , \tag{5.10}$$

where $rg(u_i)$ and $rg(v_i)$ are the respective ranks of u_i and v_i.

Note that if fundamental frequencies can be clearly identified in the experimental spectra, one is able to create a synthetic (noise-less) spectrum similar to the theoretical spectrum using Eq. 5.6 for the comparison. This has been tested in the course of our work. However, automated identification of the peaks in experimental spectra is challenging for molecules with many normal modes and prone to errors due to experimental noise. Therefore, using the measured experimental data points as a reference seems to be a more robust and general approach. If required, additional values can be estimated by linear interpolation between points u_i and u_{i+1}.

5.3. Computational Details

Calculations at the GFN–FF, GFN1–xTB, and GFN2–xTB level were performed with version 6.3.2 of the `xtb` code which was also used as a wrapper for numerical Hessian calculations with PM6-D3H4,[66,327] PM6-D3H4X,[326] and PM7,[67] using the `MOPAC` program (version 19.179L). DFT calculations at the B3LYP-3c level were conducted with the Turbomole program package (version 7.4.1).[493] The term B3LYP-3c stands for a standard self-consistent B3LYP-D3(BJ)$^{\mathrm{ATM}}$-gCP/def2-mSVP[55,152,158,172,173] calculation, *i.e.*, the method employs the well-known B3LYP exchange-correlation functional with a small (adapted) double-zeta Ahlrichs type Gaussian AO basis set, the standard D3 dispersion correction with three-body term added and a geometrical correction for BSSE (gCP). This composite approach has been benchmarked in the course of this work on subsets of the GMTKN55 meta-database[109] and found to perform similar for thermochemistry and non-covalent interactions as the PBEh-3c parent method. A similar composite scheme based on the B3LYP functional has already been proposed for computations of molecular thermochemistry.[494] As opposed to PBEh-3c, however, the functional (GGA) part has not been modified in B3LYP-3c. D3(BJ)$^{\mathrm{ATM}}$ and gCP are additive corrections that are implemented in the Turbomole program packages (version 7.4 and higher) and can be requested with the `$gcp dft/sv(p)` and `$disp3 -bj -abc` keywords. Similar usage options exist, *e.g.*, for the ORCA program package.[495] The composition of the def2-mSVP basis set is discussed in Ref. 158.

The DFT calculations in Turbomole were conducted with a *m4* grid, using the RI approximation and matching auxiliary basis functions. The `xtb`, Turbomole, and ORCA program packages allow a modification of atomic masses via the input files. For the metal-containing complexes in section 5.4.4 reference HVF were calculated at the B3LYP-D4/def2-TZVPP[174,338] level using the RIJCOSX approximation[496] (with keywords: *gridx7*, *nofinalgridx*) throughout. These calculations were performed using the ORCA program (version 4.2).[495] All structures were optimized using the `xtb` program as a driver for the ANCOPT optimizer with *normal* convergence criteria. The various similarity measures were calculated with a small standalone code called `newspecmatch` written in `Fortran`. For the FWHM a value of 30 cm^{-1} was chosen. From Ref. 464 it is known that optimization of the FWHM with respect to an experimental-theoretical best match yields larger values (>60 cm^{-1}). This is physically not plausible and seems to be an artifact of errors in the theoretical frequencies. A width of 30 to 35 cm^{-1} is physically realistic but still provides some leeway for slightly larger frequency deviations between the spectra.

6556 experimental gas-phase IR spectra and corresponding molecule structures were obtained from the NIST database,[434] containing combinations of the elements H, C, N, O, F, Cl, and Br. We will refer to this simply as the 'HCNO' set. Ionic structures were excluded. Furthermore, 538 sulfur-, 100 phosphorus-, and 53 silicon-containing structures and their associated gas-phase IR spectra were also obtained from the NIST. The molecules are mainly "organic" but contain various, sometimes rather complex inorganic parts. Because very few experimental spectra were available for transition metal complexes, HVF were calculated for 58 species taken from the TMG145 database,[324] and the high B3LYP-D4/def2-TZVPP level taken as reference for the

tested SQM and FF methods.

5.4. Results and Discussion

5.4.1. General Performance and Global Scaling of Frequencies

It is standard practice in computational chemistry to scale the calculated HVF by a method dependent parameter, that is typically determined by minimization of the root-mean-square deviation (RMSD) between experimental fundamental frequencies and theoretical HVF.[381,383] Special scale factors exist also for thermochemical applications.[382,497] A fitting strategy based on the maximization of the $r_{pearson}$ and $r_{spearman}$ was reported in Ref. 464. In contrast to QM or SQM methods, force field methods already include (by fitting to experimental data) some empirical adjustment and global scaling of the frequencies has very little impact on overall performance.[464] The tested GFN–FF employs a default global scaling factor of 1.03 which was adjusted to match reference B97-3c computed HVF in the course of its development.

The SQM methods aim at bridging the gap between FFs and *ab initio* methods, both qualitatively and in terms of computational cost. Due to their underlying quantum chemical nature and the use of (minimal) AO basis sets, dipole moment estimations are expected to be better than in FFs but worse than with DFT. The quality of the force constants in SQM on the other hand is strongly dependent on the methods parametrization and will mainly determine the quality of the HVF. As noted above, methods of the GFN family were developed to provide good performance for the calculation of frequencies and therefore are a natural choice for this study. The monopole based GFN1- and multipole extended GFN2–xTB tight-binding schemes, and a non-electronic variant called GFN–FF, are tested. For a broader perspective, results will also be shown for some PMx methods (PM6-D3H4,[327] PM6-D3H4X,[326] PM7[67]) because they are another important class of SQM methods. These results will, however, not be discussed in detail.

A point of reference for all further comparisons is, besides the experimental IR spectra, the performance of the B3LYP-3c method. B3LYP is often cited among the best performing DFT methods for the calculation of vibrational frequencies.[381,382] This was confirmed for a few test cases in comparison to PBEh-3c,[158] B97-3c,[180] as well as PBE0[337] functionals in the course of this project. Because B3LYP is a non-local, global hybrid functional, the calculations can become quite expensive which is counterbalanced here by the use of the pre-defined small def2-mSVP basis. Dipole moments, IR intensities, and frequencies on a medium accuracy level are rather insusceptible to the basis set size as long as proper polarization functions are present.[498–501] Hence, B3LYP-3c is expected to yield reasonable results for typical organic molecules, but for more complicated electronic structures the level of theory should be improved. Whether B3LYP-3c really represents an efficient standard DFT level for IR spectra calculations of large systems is one of the questions of the present work.

The IR spectra for the 6556 compounds in the HCNO set were calculated at all levels of

theory. Respective geometry optimizations were consistently started from the structures in the NIST database and in this first evaluation, conformational effects were not considered, *i.e.*, the structures taken correspond to some random conformation in the case of flexible systems (see Section 5.4.3).

The distribution of similarity measures (Eqs. 5.7-5.10) between the unscaled theoretical and the experimental IR spectra over the benchmark set are shown in Fig. 5.1.

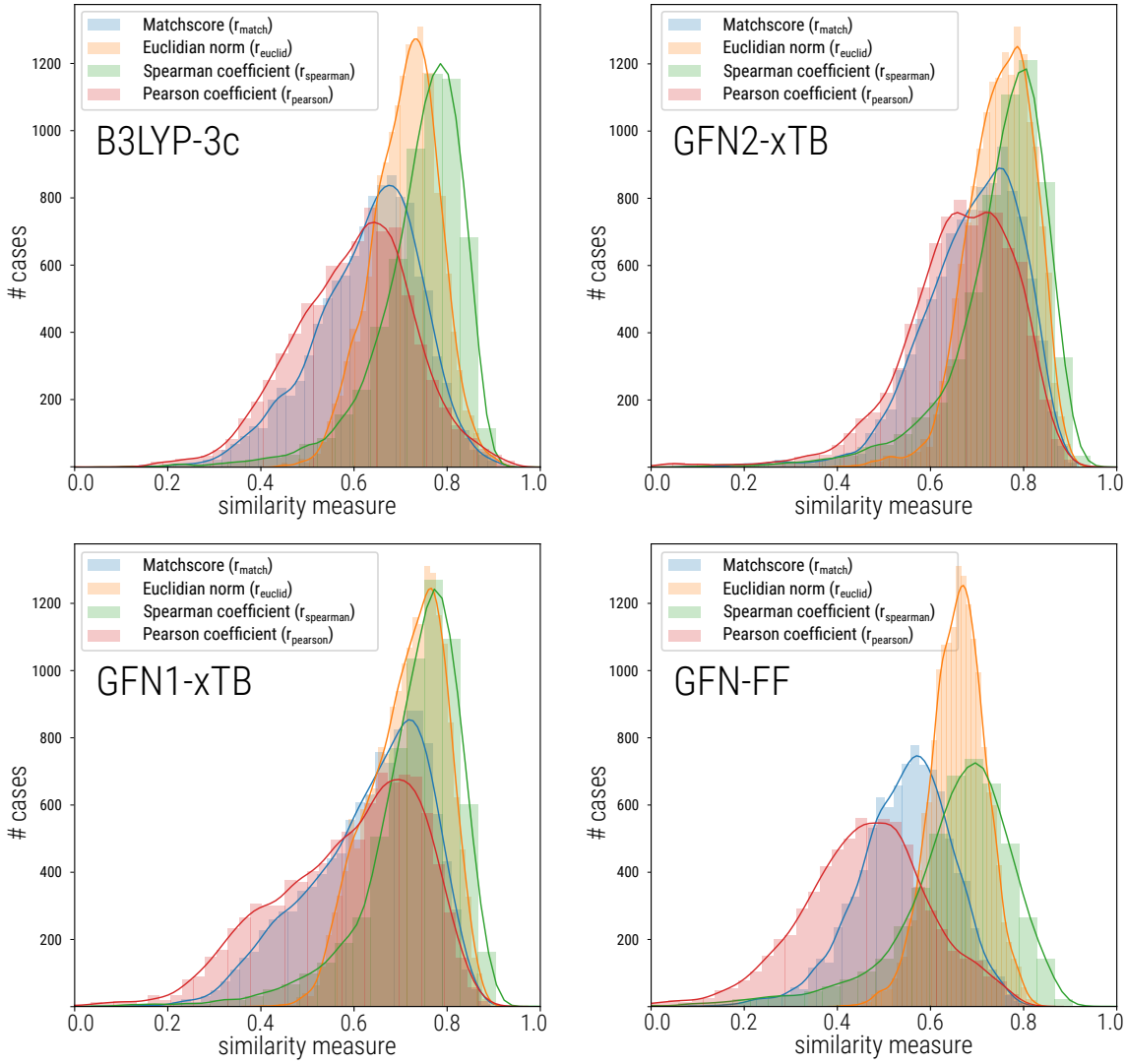

Figure 5.1.: Similarity measure distribution (6556 cases) for unscaled theoretical spectra at the B3LYP-3c, GFN1–xTB, GFN2–xTB and GFN–FF level. Bars were binned (40 bins) for better visualization.

Several observations can be made for the unscaled spectra comparison. The r_{match} and $r_{pearson}$ yield similar distributions at all levels of theory, with the r_{match} yielding slightly higher scores. This is expected since both scores have the same mathematical origin. As already noted in the literature the $r_{spearman}$ can display also non-linear correlations and yields higher scores than the r_{match} or $r_{pearson}$ on average. The distribution of the $r_{spearman}$ scores is more narrow than those

of r_{match} and $r_{spearman}$, except at the FF level. This could either indicate a general non-linear correlation between experimental and theoretical spectra, or an inability of the $r_{spearman}$ to reflect finer characteristics. The r_{euclid} also gives relatively high values for the correlation and shows a very narrow score distribution at all levels of theory. Of the four investigated metrics the r_{euclid} seems to be the least sensitive to the spectral features but this issue has not been investigated in detail. In all further comparisons we will mainly discuss results obtained with the r_{match}.

In terms of performance GFN2–xTB gives the best (average) qualitative correspondence to the experimental spectra, followed by GFN1–xTB, B3LYP-3c and lastly GFN–FF. The three different PMx methods perform very similar, but overall worse than the GFN methods. The average scores are given in Table 5.1. At first glance it may be surprising that the DFT method performs worse than both tight-binding variants. The simplest reason for this observation is the systematic deviation (overestimation) of the HVF with hybrid DFT methods. Intensities in the IR spectra are well reproduced by B3LYP-3c, but in particular the higher frequency modes are not described well by the harmonic approximation. A typical correction for this behavior is the linear scaling of frequencies by a global factor (global frequency scaling, GFS). For the 6556 systems and all methods, molecule specific frequency scaling (MSFS) factors, *i.e.*, an optimum scaling factor for each case, were determined by maximizing the r_{match} in a steepest descent optimization approach. The corresponding similarity measure distributions are shown in Fig. 5.2 and the average scores are given in Table 5.1.

Table 5.1.: Average metrics r_{match}, r_{euclid}, $r_{spearman}$, and $r_{pearson}$ for the 6556 unscaled and ideally (MSFS) scaled IR spectra calculated at all tested levels of theory. For better visibility standard deviations have been omitted from the table and can be found in the electronic supporting information.

	unscaled				molecule specific scaling			
method	r_{match}	r_{euclid}	$r_{spearman}$	$r_{pearson}$	r_{match}	r_{euclid}	$r_{spearman}$	$r_{pearson}$
B3LYP-3c	0.628	0.710	0.739	0.597	0.864	0.880	0.810	0.873
GFN1–xTB	0.632	0.714	0.727	0.587	0.727	0.775	0.752	0.710
GFN2–xTB	0.690	0.751	0.748	0.661	0.765	0.802	0.759	0.750
GFN–FF	0.545	0.659	0.650	0.459	0.597	0.691	0.666	0.528
PM6-D3H4	0.518	0.644	0.591	0.424	0.640	0.713	0.640	0.584
PM6-D3H4X	0.518	0.644	0.591	0.424	0.634	0.713	0.641	0.584
PM7	0.476	0.620	0.570	0.366	0.634	0.713	0.641	0.582

The r_{match} and $r_{pearson}$ distributions in Fig. 5.2 are much more narrow compared to Fig. 5.1 for all methods but the force field. As expected, the largest improvements by scaling are obtained for B3LYP-3c. The average r_{match} increases by 0.236, with an average scaling factor of 0.970±0.011. This is consistent with GFS factors reported for B3LYP in the literature.[381–383,502,503] For the GFN1- and GFN2–xTB methods smaller improvements are observed for linear, molecule specific scaling of the frequencies.

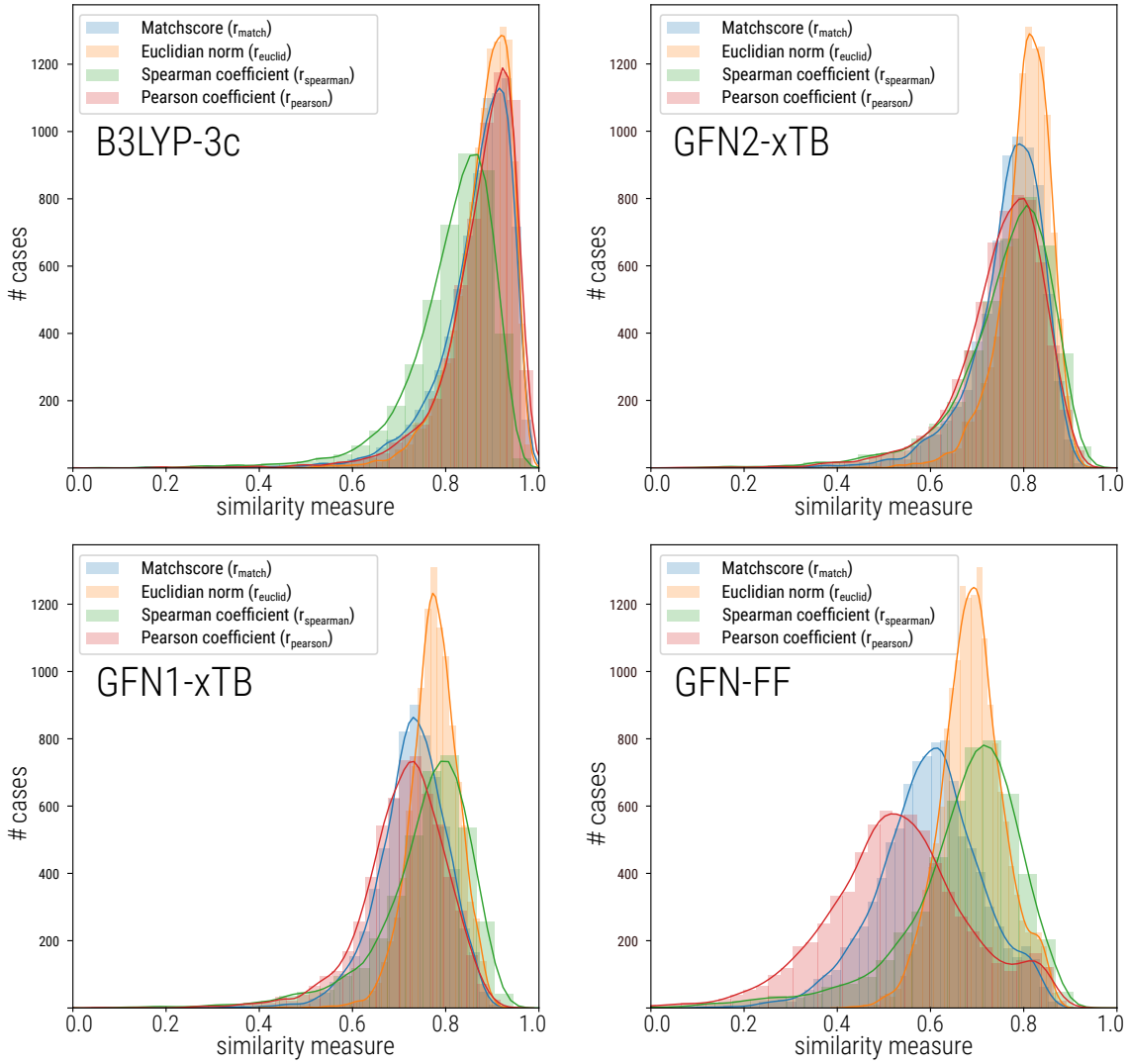

Figure 5.2.: Similarity measure distribution (6556 cases) for spectra at the B3LYP-3c, GFN1–xTB, GFN2–xTB and GFN–FF level. The calculated frequencies were uniformly, but specifically scaled for each molecule in order to maximize the r_{match} to experiment. Bars were binned (40 bins) for better visualization. Values of unity for the r_{match} result from graphic interpolation by the plot program.

The average scaling factor is 0.982 and 0.990 with standard deviations of 0.019 and 0.017, respectively. GFN–FF shows the least improvement by MSFS scaling. Among the investigated GFN methods, the factor of GFN–FF is the one closest to unity and further has the largest standard deviation, *i.e.*, the error is more molecule specific and randomly distributed. In case of the PMx methods this effect is even more pronounced than for the force field. Using MSFS the average similarity measures among the PMx methods is again very similar and the performance is in between GFN–FF and GFN1–xTB. Determined GFS factors are around 1.019 and show a very large standard deviation of 0.04–0.05, which essentially obviates the value of global frequency scaling for the PMx methods. This seems to be consistent with the literature, where

non-linear scaling schemes are used in combination with PM6 or PM7 because the linear scaling factors tend to be large.[463,504,505]

Table 5.2.: Frequency scaling factors for the investigated theoretical methods determined as the average of the MSFS factors and the respective similarity measures obtained by application of these factors as global frequency scaling parameters.

method	ν_{scal}	SD(ν_{scal})	r_{match}	r_{euclid}	$r_{spearman}$	$r_{pearson}$
B3LYP-3c	0.970	0.011	0.862	0.879	0.869	0.819
GFN1–xTB	0.982	0.019	0.686	0.747	0.660	0.754
GFN2–xTB	0.990	0.017	0.723	0.773	0.702	0.757
GFN–FF	0.993	0.020	0.564	0.670	0.434	0.659

Based on the absolute similarity measure values provided in this section a clear qualitative trend can be seen. DFT on average has the highest correspondence to experimental gas-phase data while GFN–FF shows the worst performance. The SQM methods GFN1–xTB and GFN2–xTB are in between the FF and B3LYP-3c, but overall closer to the DFT performance. The initial performance expectations of the SQM methods thus were validated. Although the IR spectra at the GFN–FF level are less accurate than at the SQM level, the calculated frequencies still might be useful to obtain thermostatistical corrections.[204] Visual inspection of a few typical cases reveals that a major part of the observed error with GFN–FF stems from bad intensities, which is understandable considering the relatively simple description of the electrostatics in the FF. Furthermore, the GFN1- and GFN2–xTB tight-binding methods clearly outperform the PMx competitors.

For all investigated levels of theory the different metrics provide consistent descriptions of the similarity, *i.e.*, there is no performance reordering between different theoretical methods if another similarity measure is used in Table 5.2. Furthermore the absolute values (and the respective standard deviations, see electronic supporting information) of r_{match} and $r_{pearson}$, as well as r_{euclid} and $r_{spearman}$ are often very similar. This indicates that any of the similarity measures could be used for spectra comparisons, or that they can be used complementary.

The r_{match} is a quantitative similarity measure, but since errors from mismatching frequencies and errors from wrong intensities are often inseparable, the scores should be interpreted with care, and on a qualitative rather than quantitative basis. For r_{match} around 0.6 or smaller, the largest deviations result from systematic frequency shifts for unscaled DFT. In case of SQM and FF methods errors in the fingerprint frequency range are common and intensities are often not well described. Spectra with r_{match} between 0.7 and 0.8 already match well to the experiment and in particular intensities are much better reproduced, but some smaller peaks might still be missing or shifted. Spectra with a r_{match} above 0.9 have predictive quality and some of the remaining deviations are often only a result of noise in the experimental data. A few examples for the comparison of theoretical and experimental spectra are shown in Fig. 5.3 to provide a qualitative impression for different r_{match} values.

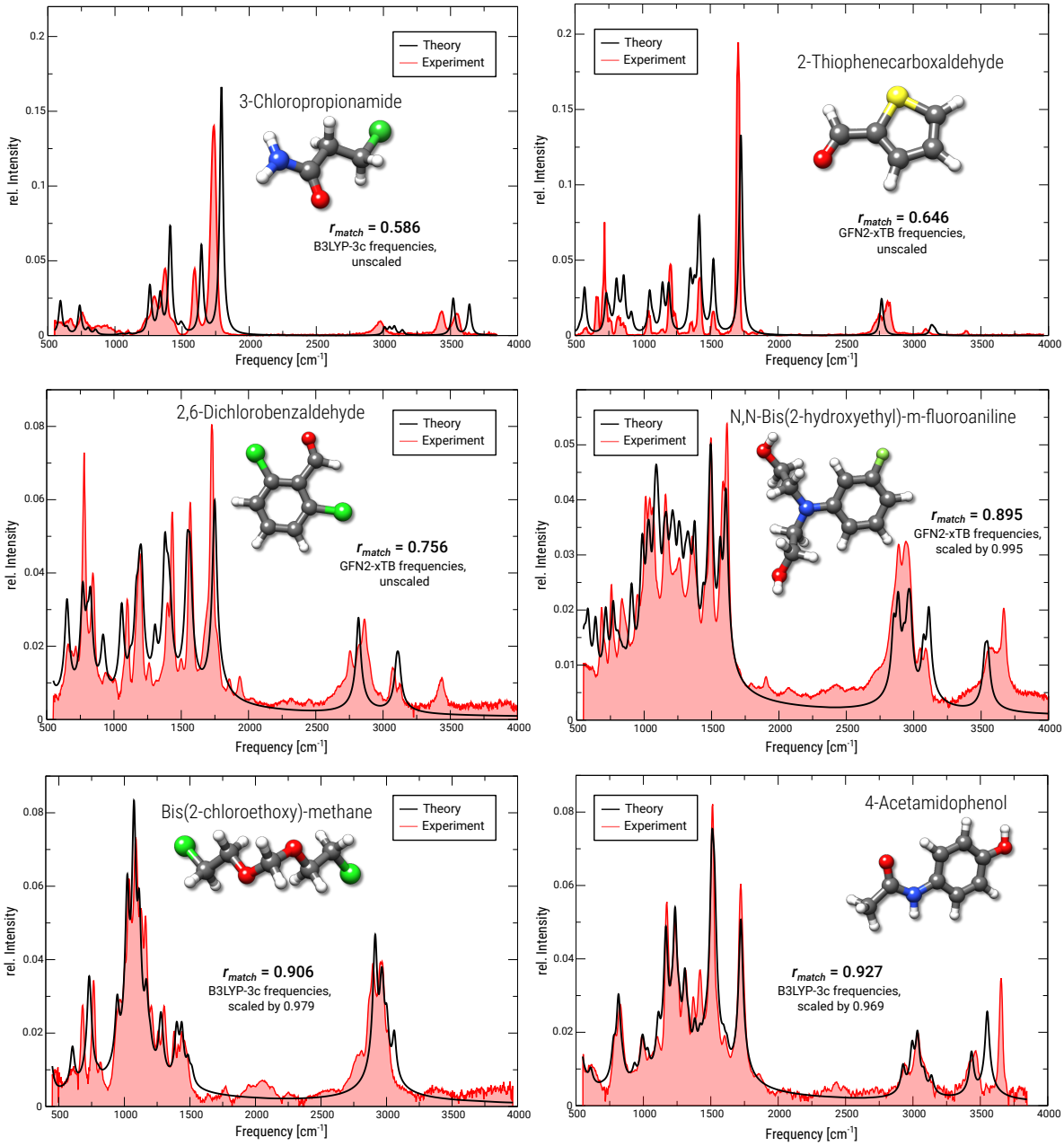

Figure 5.3.: Comparison between the theoretical (black) and the experimental (red) spectra for six exemplary molecules, at GFN2–xTB or B3LYP-3c level.

At all r_{match} ranges, noise in the experimental data can be a problem. For better comparability one could either try to smooth-out the experimental data by standard polynomial procedures (*e.g.*, Savitzky-Golay filtering[506]), or automatically identify fundamental frequencies from the spectra. Both strategies pose their own problems and can also depend on the amount of noise. Based on our experience, where applicable, a direct comparison to the experimental data is to be preferred.

The aim of this study is to find theoretical methods that provide maximum similarity scores

when comparing a calculated spectrum with its experimental counterpart. However, with regard to possible applications in the field of automated unknown compound identification, it is not fully clear whether a candidate ranking based on a maximized spectral similarity is the optimum search strategy. The reason for this is that comparisons of theory and experiment are then made between different molecules, opposed to the already known structure in this study. The application of similarity measures for compound discrimination under realistic conditions is currently investigated in our labs and will be discussed elsewhere.

5.4.2. Scaling of Atomic Masses

In this section, another easily applicable scheme to improve calculated spectra is explored. ¿From Eq. 5.1–5.3 it is obvious that modification of the atomic masses $m_{i/j}$ will have a (non-linear) effect on the frequencies. As noted in the introduction this is known in context of isotope effects in IR spectra.[455,507] In a recent study Irikura[473] proposed a scaling of reduced masses for diatomic molecules as a possible alternative to linear frequency scaling factors ν_{scal}. The modification of the masses could be interpreted as an atom specific way to mimic effects of anharmonicities. In an extension to this idea we investigate the effect of element mass scaling (EMS) factors, which should provide more flexibility than scaling of the frequencies directly. This has never been tested thoroughly for a large data set of polyatomic molecules. This approach is related to the work of Pulay *et al.*, who proposed individual scaling of the internal coordinate force matrix elements.[385,468,474,508] In contrast to this scheme, however, the EMS does not require additional internal coordinate setup or any other modification to the frequency computation. All modern DFT and SQM methods are designed as atomistic models, *i.e.*, basis functions are defined for different atoms, and SQM parametrizations are normally element dependent. Thus, an element-dependent EMS parameter is physically plausible and technically feasible. Scaling factors β_i for the atomic masses m_i non-linearly enter the Hessian matrix $\mathbf{F}^{(m)}$ in Eq. 5.1 by the power of minus one half. Because the normal modes are affected depending on the involved atom types, the non-linear mass scaling should be able to improve normal modes that include movements of different atom types, *e.g.*, C-O or N-N stretches are treated differently. For simplicity (and easy input handling) instead of introducing an element dependent scaling parameter β_i into Eq. 5.1, the masses can directly be treated as a variable parameter ($m'_i = m_i\beta_i$). Many computational chemistry program packages allow users to read in atomic masses via the input, and hence code modifications are not necessary.

The masses m'_i for H, C, N, O, F, Cl, Br were fitted directly to the experimental data by maximization of the r_{match}. A single global fit was performed for these elements using a Levenberg-Marquardt algorithm.[421,422] The 6556 spectra were randomly split (≈50%/50%) into a training and a evaluation set, which is common practice for cross-validation in data sciences.[509] The resulting optimized masses m'_i are shown in Table 5.3.

Note that the masses m'_i should not be interpreted as actual atomic masses, but can only be understood by their effect on the force constants. Since the masses are in the denominator of the prefactor in Eq. 5.1, the usually obtained larger mass m'_i repairs too large force constants

Table 5.3.: Optimized atomic masses for the elements H, C, N, O, F, Cl, and Br (in a.u.). m_i refers to standard atomic masses, modified values m_i' for the GFN methods and B3LYP-3c were determined by fitting to experimental IR spectra.

Atom type	reference m_i	B3LYP-3c m_i'	GFN1–xTB m_i'	GFN2–xTB m_i'	GFN–FF m_i'
H	1.0079	1.07823	1.0801	1.0518	1.0607
C	12.0110	12.7059	11.7831	11.7844	11.6786
N	14.0067	15.1762	14.1462	12.8194	11.8136
O	15.9994	17.4074	18.8858	17.5056	15.7004
F	18.9984	15.7403	13.5474	11.8874	9.3633
Cl	35.4530	23.1230	52.3449	14.0078	14.1648
Br	79.9040	79.9040[a]	60.3144	45.4038	43.1998

[a]Modification of mass did not improve the fit.

(missing anharmonicity). The GFN methods show similar trends for the adjusted masses H to F, which suggests systematic errors of the force constants within the method family. During the fitting procedure it was observed that a modification of the hydrogen mass has the largest influence on the frequencies. This is expected since it is the most common element in the data set. The heavier elements Cl and Br show the largest deviations from their true masses. However, their overall effect on the frequencies is small and drastic changes of the masses are necessary to obtain any influence at all. For Cl and Br GFN1–xTB is the only method that does not follow the general trend observed for these elements in the other methods. This could be related to the halogen bond correction terms employed in GFN1–xTB.

The average similarity measures calculated with the fitted masses are shown in Table 5.4. Following good cross-validation practices, these scores were obtained only for the test set, but the results are nearly identical to the similarity measures of the fitting set.

Table 5.4.: Averaged similarity measures between experimental and theoretical IR spectra obtained from fitted-mass spectra calculations.

method	r_{match}	r_{euclid}	$r_{spearman}$	$r_{pearson}$
B3LYP-3c	0.865	0.881	0.813	0.873
GFN1–xTB	0.721	0.771	0.747	0.699
GFN2–xTB	0.750	0.793	0.756	0.731
GFN–FF	0.589	0.686	0.678	0.516

With the EMS, average similarity measures are only slightly worse than the scores obtained by MSFS in Table 5.1. Keep in mind that the actual comparison has to be made with the values provided in Table 5.2. MSFS spectra represent some upper limit for the correspondence to the experiment, but in practice spectra are normally globally scaled by the same fixed factor ν_{scal}. Compared to these values the mass-scaling approach yields better results throughout. For

B3LYP-3c the EMS scheme even slightly outperforms MSFS, but the difference between the two schemes is small. The reason for this behavior in DFT is that errors are much more systematic compared to SQM. A few examples for mass scaled spectra in comparison with unscaled and the experimental spectra for GFN2–xTB, GFN–FF and B3LYP-3c are shown in Fig. 5.4.

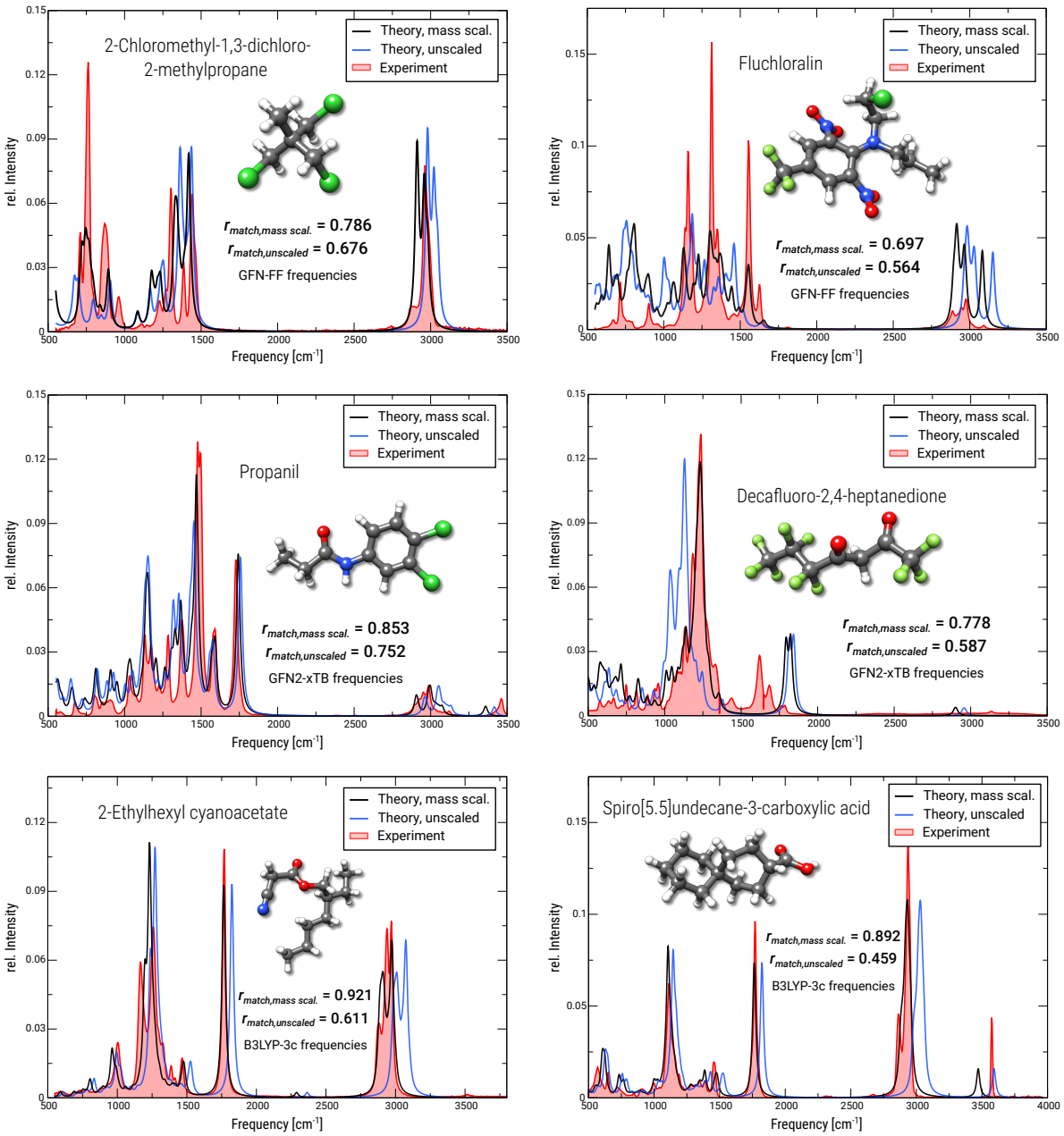

Figure 5.4.: Comparison between the theoretical mass scaled (black), theoretical unscaled (blue) and the experimental (red) spectra for six exemplary molecules, at GFN2–xTB, GFN–FF or B3LYP-3c level.

Performances are summarized for all four methods as box plots in Fig. 5.5. As can be seen, at all levels of theory there is a clear tendency suggesting that the mass-scaling approach provides

better spectra than global frequency scaling.

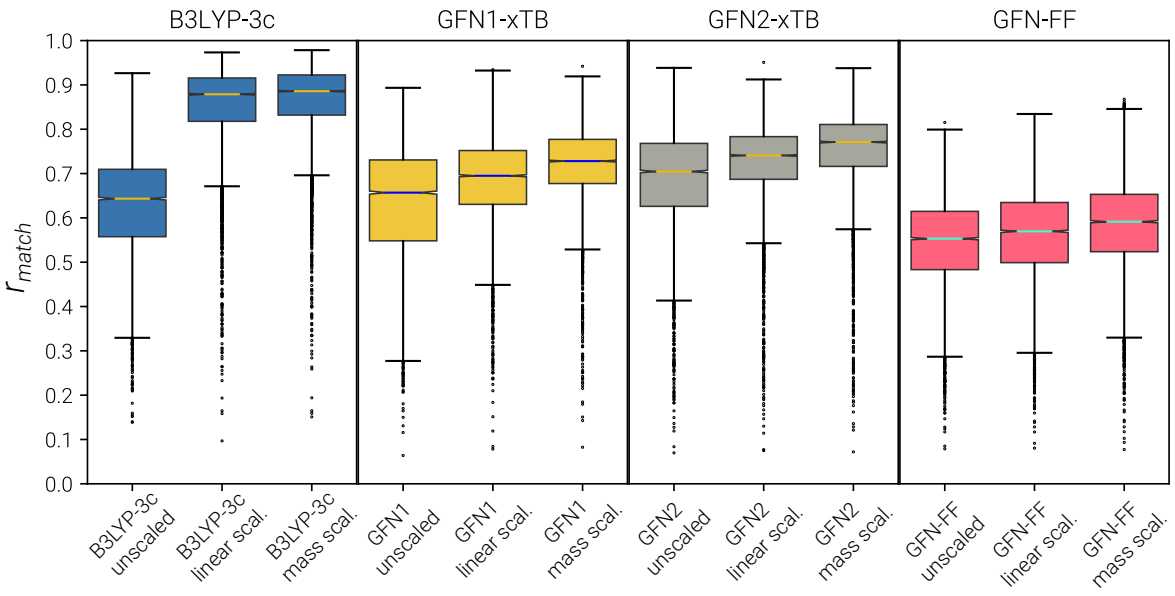

Figure 5.5.: Similarity measures (r_{match}) for the 6556 spectra comparisons at the B3LYP-3c, GFN1–xTB, GFN2–xTB and GFN–FF levels of theory, visualized as box plots.

We tested the transferability of the mass scaling approach for B3LYP-3c and GFN2–xTB for molecules containing other, heavier atoms. The idea is that once optimum atomic masses are determined for a set of elements, other effective element masses can easily be determined in separate fits. This was investigated for the main group elements sulfur, phosphorus and silicon. Experimental gas phase IR spectra were again obtained from the NIST database, totaling 538 spectra for sulfur containing compounds, 100 for phosphorus, and 53 for silicon. The masses of Si, P, S were modified on top of the previously defined masses m_i' of H, C, N, O, F, Cl, Br from Table 5.3, and are given in Table 5.5.

In general, the procedure of fitting only the mass of a single element separate from predetermined scaled masses of the other elements improves the overall result. At the B3LYP-3c level the average r_{match} for the silicon, phosphorus and sulfur sets is again better than the unscaled or globally scaled results. The improvement of the mass scaling in comparison with the GFS for B3LYP-3c is larger here than for the HCNO set above. Note, however, that the same $\nu_{scal} = 0.970$ was used here and no new linear scaling factor was determined for the silicon, phosphorus, and sulfur containing molecules and that a specific linear scaling factor might improve the $r_{match,lin}$. Surprisingly, at the GFN2–xTB level the mass scaling does not provide a better overall performance than the linear frequency scaling. The main issue here likely is that the fit is stuck in a local minimum due to predetermined masses of the other elements. This hypothesis was briefly tested with a fit for GFN2–xTB on the 538 sulfur containing structures, in which *all* element masses were adjusted. Significant changes in the HCNO masses can be observed this way, but the average match score ($r_{match,mass}$) for the sulfur subset increases to 0.677. However, the average similarity measure for the 6556 structure set decreases to 0.731

Table 5.5.: Optimum masses for the elements Si, P, and S and average match scores for the sets. $r_{match,unscal.}$ is the average similarity measure without any modification of the frequencies, $r_{match,lin}$ is the average similarity measure with a fixed ν_{scal} applied. $r_{match,mass}$ refers to similarity measures obtained with the fitted masses given in this table and Table 5.3.

Atom type		reference	B3LYP-3c	GFN2–xTB
Si	m_i' [a.u.]	28.0855	31.1671	26.1354
	$r_{match,unscal}$		0.650	0.668
	$r_{match,lin}$		0.795	0.686
	$r_{match,mass}$		0.802	0.681
P	m_i' [a.u.]	30.9738	26.0246	25.2719
	$r_{match,unscal}$		0.748	0.734
	$r_{match,lin}$		0.791	0.740
	$r_{match,mass}$		0.822	0.742
S	m_i' [a.u.]	32.0600	26.0670	23.7071
	$r_{match,unscal}$		0.576	0.612
	$r_{match,lin}$		0.782	0.640
	$r_{match,mass}$		0.806	0.637

using these masses. The performance is still slightly better than that seen for a global ν_{scal}, but overall worse than the result for ideally scaled masses (cf. Table 5.4). This apparently confirms the assumption of a local minimum in the fit. As a consequence, when using scaled masses some errors can be expected for modes with large contributions of non-mass scaled elements. A larger global adjustment that includes the other elements would likely solve this problem. This was not pursued here because too few experimental reference data points were available for the heavier elements.

5.4.3. Conformational Dependence

The molecular conformation can have a significant influence on the IR spectrum[510] but this is rarely investigated comprehensively in the literature. Spectra can be affected due to steric effects or intramolecular non-covalent interactions and in our experience, different hydrogen bonding patterns in particular can have a strong influence. Intensities can be influenced as well if dipole moment changes are large. An example is presented in Fig. 5.6. The 9-aminononanoic acid zwitterion was chosen as an extreme case to demonstrate the effect of different conformations on the dipole moment and IR spectrum. No experimental spectrum is available for this molecule.

For a folded conformation a much smaller dipole moment is expected than for a structure with a large spatial charge separation. This is confirmed at the GFN2–xTB[GBSA(H_2O)] level where the folded structure is the global minimum energy conformer. As can be seen in Fig. 5.6, most vibrational frequencies are not affected by the conformation. Significant shifts are only observed for the signal at around 3200 cm^{-1}, which corresponds to a N-H stretch, and the signal

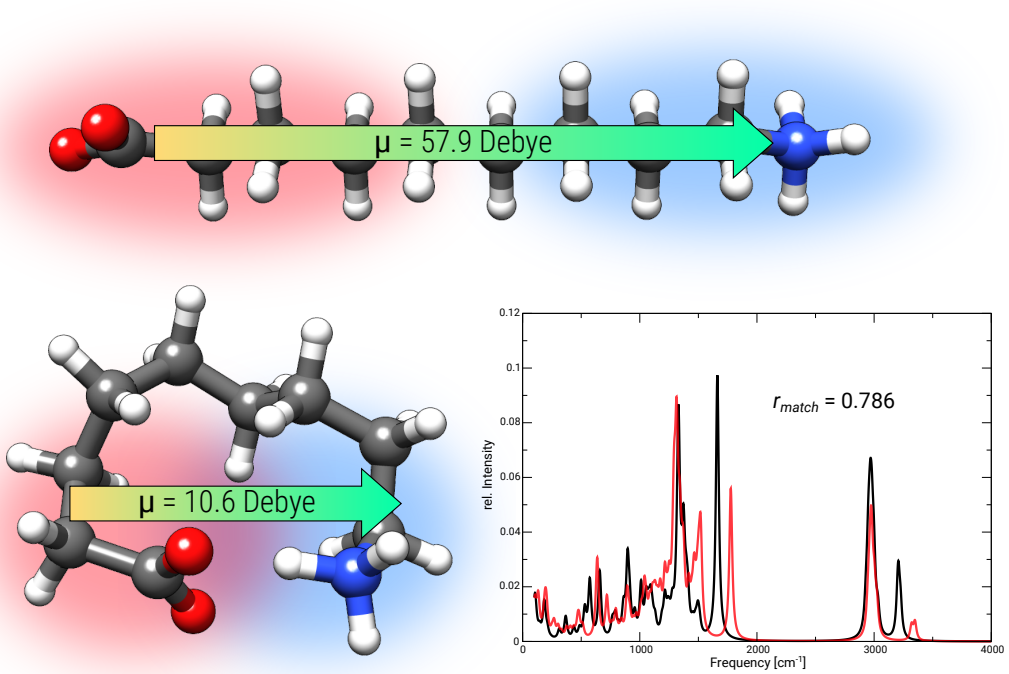

Figure 5.6.: Two conformations of the 9-aminononanoic acid zwitterion and their respective calculated IR spectra. Orientation of the dipole vector is indicated by the green arrow. The dipole moments and frequencies were calculated at the GFN2–xTB[GBSA(H_2O)] level. The spectrum plotted in black corresponds to the linear conformer, the red spectrum refers to the folded conformer.

at 1800-1900 cm^{-1}, corresponding to a C=O vibration. Note that both functional groups are connected by intramolecular hydrogen bonding in the folded conformer. Larger differences can also be seen for the intensities, overall leading to a r_{match} between the two spectra of only 0.786.

Ideally one would conduct a conformational search before calculating the IR spectrum and use the lowest energy conformer or compute Boltzmann population weighted spectra. To obtain a more conclusive impression of the conformational effect, we selected a subset of 554 flexible molecules from the 6556 set, and compared the unscaled GFN2–xTB IR spectrum of the lowest conformer with that just computed for the input structure (as obtained from the NIST database). The subset structures were selected based on the flexibility score ξ_f proposed in Ref. 33. Conformational searches were conducted with the iMTD-GC workflow as implemented in the `crest` code[33] at the GFN2–xTB level. All 554 molecules have flexibility scores of $0.8 < \xi_f < 0.9$ commonly obtained for relatively floppy structures such as poly-peptides or medium sized branched alkanes.

Conformational effects are investigated by a similarity measure based comparison of the spectra before and after the conformational sampling, *i.e.*, two theoretical spectra are compared here. For molecules with a similarity measure close to unity, the conformation either does not affect the spectrum, or the input geometry already was (close to) the minimum energy conformer.

The average r_{match} between the two spectra at GFN2 level over all 554 cases is 0.965, *i.e.*, on average a deviation of about 3.5 % was observed. An example is shown in Fig. 5.7.

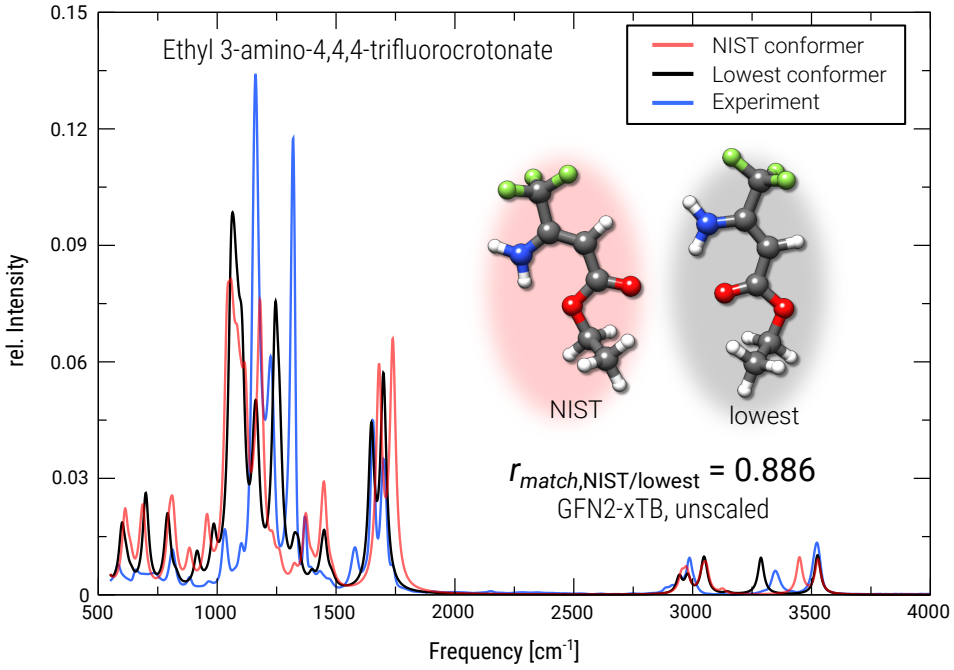

Figure 5.7.: Comparison between the theoretical IR spectra (GFN2–xTB) of two conformations of ethyl 3-amino-4,4,4-trifluorocrotonate. The conformer (and spectrum) marked in red corresponds to the structure as obtained from the NIST database. The conformer marked in gray (black line) was obtained through conformational sampling. The experimental spectrum is plotted in blue.

For this molecule significant differences between the two IR spectra are observed, both regarding frequencies and intensities. With a r_{match} of 0.886, the conformational effect is considered to be quite large. Furthermore, the lowest energy conformation found matches better with the experiment (r_{match}=0.746) than the structure from the NIST database (r_{match}=0.699). This results mainly from better matching signals in the 1000 to 1500 cm^{-1} and 3000 to 3600 cm^{-1} regions. Note that for this example only two dihedral angles differ significantly between the two conformations and the effect on the simulated vibrational spectrum is already large. This indicates that general conclusions based solely on inspection of the structure are difficult to draw but that at a high accuracy level conformational changes should be considered when calculating IR spectra for all but the most rigid structures. According to our limited experience, even moderately flexible structures with approximately $0.5 > \xi_f > 0.3$ can still undergo considerable conformational changes. In fact, several low energy conformations will typically contribute to the experimental spectrum according to their respective Boltzmann population at given temperature and should be considered for very accurate predictions.[471,510] In automatic workflows, IR spectra averaging could be adapted similar to previously discussed averaging of calculated NMR spectra.[40]

5.4.4. IR Spectra of Transition Metal Compounds

A major advantage of the GFN methods over other semiempirical schemes is their robustness with regards to the possible chemical composition of the investigated molecules. The methods are consistently parameterized up to radon (Z=86) and do not require any manual structure preparation step. In this last section we briefly investigate the performance of GFN1-, GFN2–xTB and GFN–FF for the calculation of IR spectra of transition metal (TM) containing compounds. The benchmark set is composed of 58 structures taken from the recently published TMG145 benchmark set.[324] Only those structures were selected from TMG145 that are known to qualitatively maintain the correct geometry upon optimization at the GFN levels and for which we were able to obtain reasonable structures and frequencies at the B3LYP-D4/def2-TZVPP level. Because no gas-phase IR spectra are available for these molecules we compare directly to DFT as reference. The larger basis set was used because transition metal complexes usually involve more challenging electronic structures than the mostly organic systems from the previous sections. Since the difference between the mass scaling and linear frequency scaling is small for B3LYP and we do not have a reference for the masses m_i' of the transition metal atoms, the DFT frequencies were scaled by a global factor of 0.967, which was taken from the literature.[383] The r_{match} for all complexes are shown as bar plots for GFN1-, GFN2–xTB and GFN–FF in Fig. 5.8.

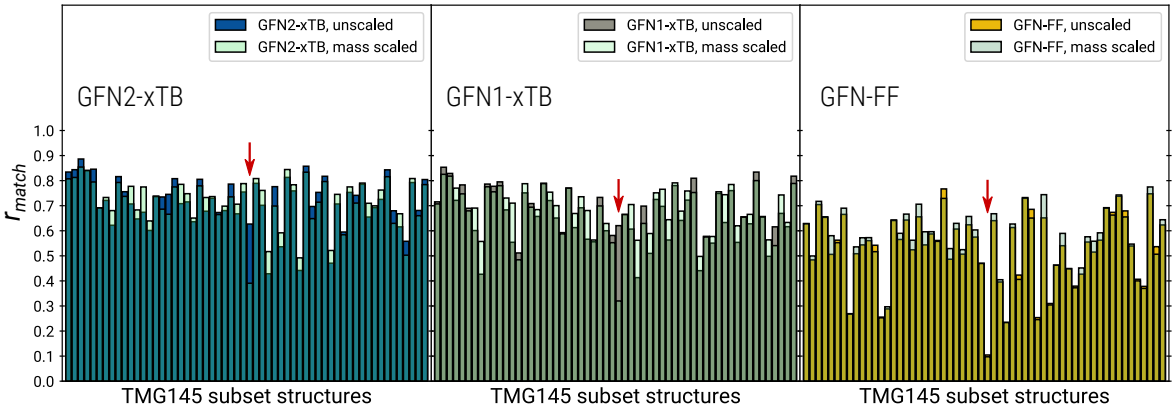

Figure 5.8.: Bar plots for the r_{match} of all the 58 transition metal compounds investigated at the GFN1–, GFN2–xTB and GFN–FF level. The reference IR spectra were calculated at the B3LYP-D4/def2-TZVPP level. The light-colored bars denote the r_{match} obtained by using mass scaling. The system indicated by the red arrow is discussed in the text below.

For this set and taking B3LYP as a reference, GFN2–xTB (avg. $r_{match,unscal} = 0.714$, $r_{match,mass} = 0.716$) performs better than GFN1–xTB (avg. $r_{match,unscal} = 0.662$, $r_{match,mass} = 0.682$), and both SQM methods perform better than GFN–FF (avg. $r_{match,unscal} = 0.532$, $r_{match,mass} = 0.544$) . For all three methods, the mass scaling improves the performance only slightly compared to the unscaled result. One case (indicated by the red arrow in Fig. 5.8) is

very problematic for all the GFN methods. The corresponding structure is a mercury complex, in which the metal is bound in between two dicarba-*closo*-dodecaboranyl clusters.[511] The IR spectrum of this structure is not very complicated, but is dominated by a very strong signal at 2600 cm^{-1} corresponding to many B-H vibrations (see Fig. 5.9). Because the hydrogen mass is fitted and the mass of boron is not, this signal is too strongly shifted in the mass scaling approach which has a large impact on the r_{match}. Frequencies in the fingerprint region are barely affected by the mass scaling. However, the intensities are not well described at the SQM level which is the main problem in many cases with low r_{match} values. At B3LYP-3c level a similary

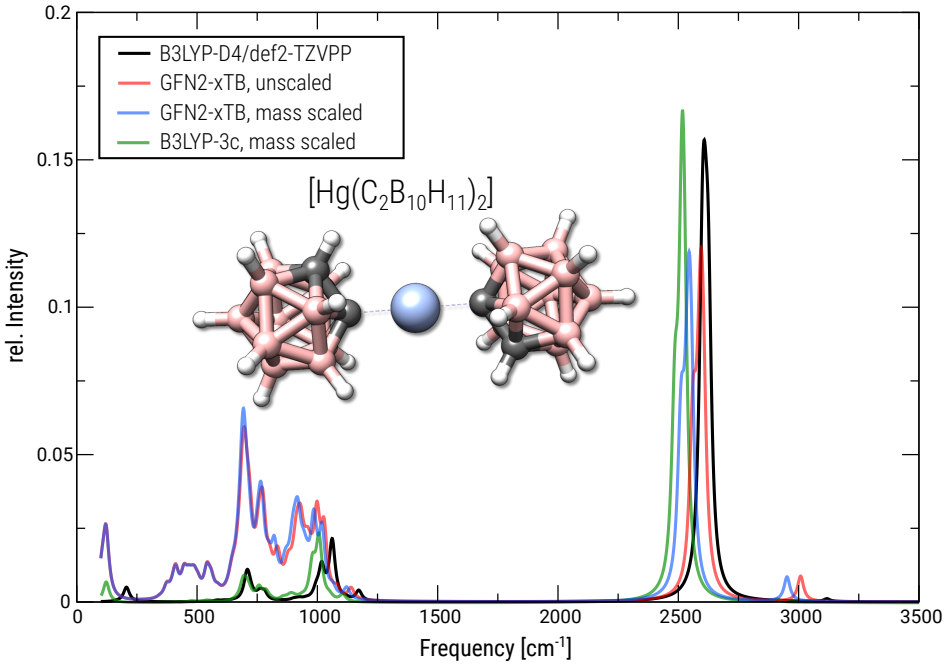

Figure 5.9.: Theoretical IR spectra for the $[Hg(C_2B_{10}H_{11})_2]$ complex.

frequency shift is observed due to the mass scaling, but intensities are reproduced very well. In fact, if neither the calculated B3LYP-D4/def2-TZVPP reference nor the B3LYP-3c spectrum is scaled, the r_{match} will be very close to unity (0.972). This excellent agreement is a further indication for the viability of B3LYP-3c.

Many other IR spectra in this benchmark are reasonably well reproduced by the GFN methods, with most of the errors resulting from the fingerprint frequency region. The main factor determining the similarity to the reference here is probably the description of the electrostatic interactions. Of the three GFN methods GFN2–xTB provides the most sophisticated description also including up to atomic quadrupoles[39] while GFN1–xTB and the force field rely only on atomic charges. For GFN–FF they are determined classically while in GFN1–xTB they are based on the total valence electron density. This usually leads to slightly better intensities and r_{match} at the GFN2–xTB level. As a typical example, a ruthenium based catalyst[512] is shown in Fig. 5.10. The intensities computed at the GFN2–xTB level more closely resemble the

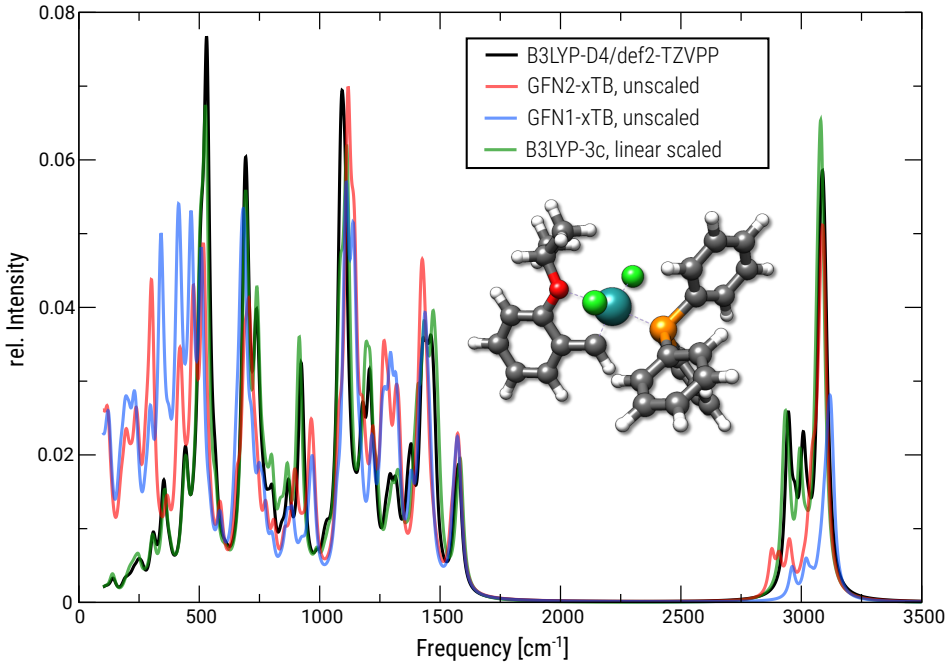

Figure 5.10.: Theoretical IR spectra for the $Cl_2Ru(=CH\text{-}o\text{-}OMeC_{10}H_6)(PCy_3)$ complex.

B3LYP reference in comparison with GFN1–xTB. Upon scaling with the same GFS factor as the reference, B3LYP-3c yields a very good spectrum and in particular the frequencies are almost perfectly reproduced. Apparently, B3LYP-3c seems to be applicable also for electronically more complicated structures in the TMG145 subset.

5.5. Conclusion

We investigated the performance of the semiempirical GFN tight-binding and force field methods for the simulation of IR spectra in direct comparison with experimental data. The performance of a new low-cost DFT method termed B3LYP-3c, which was introduced and tested in the course of this study, was investigated as well. A large benchmark set of 7247 gas-phase IR spectra was compiled with molecules containing the elements H, C, N, O, the halogens and the heavier elements silicon, phosphorus and sulfur. Electronically more complicated transition metal-containing structures and the effect of conformational changes on the IR spectra were also investigated.

For the comparison between theory and experiment several similarity measures (spectral overlap metrics) were investigated in order to avoid manual identification and assignment of fundamental frequencies from the spectra. We find that all of the tested measures can be used to sufficiently represent the similarity between two IR spectra. One similarity measure was then used to determine global frequency scaling factors and to explore a new atomic mass scaling approach which provides more flexibility than the linear scaling.

Our results show on a sufficiently large sample size that the low-cost DFT (B3LYP-3c) com-

puted harmonic vibrational spectra can be brought into good agreement with experimental spectra upon linear scaling of the frequencies. The small but well balanced double-zeta AO basis set employed provides sufficiently accurate intensities leading to overall excellently simulated IR spectra. For the semiempirical methods the scaling of frequencies is less effective and their overall performance is worse than for DFT. In general, GFN2–xTB provides the best and still useful results at a semiempirical QM level, followed by GFN1–xTB. The GFN–FF mainly suffers from relatively large errors in the intensities.

Similar trends were also observed for a benchmark of 58 transition metal complexes, which are only slightly worse than the (mostly) organic compounds at the GFN levels. Furthermore, it was found that molecular conformation can play an important role for IR spectra calculations in some cases. For medium flexible molecules, the conformation can influence the molecular dipole moments and vibrational modes might be differently coupled, leading to differences in both intensities and frequencies.

All methods suffer from inherent errors for the potential energy surface and the employed harmonic approximation. For the GFN methods and B3LYP-3c an atomic mass scaling correction approach was successfully employed as an alternative to standard linear frequency scaling. For the semiempirical methods, mass scaling yields more consistent and overall better results, but also at the DFT level the performance is slightly better than with a global frequency scaling factor. Attention has to be paid, however, for molecules that contain elements without a fitted mass scaling parameter. This problem was observed for the transition metal complexes and the silicon, phosphorus, and sulfur subsets.

Based on the presented results, we recommend the usage of semiempirical QM methods as a cost efficient and reasonably accurate tool to study vibrational spectra in initial steps of large scale compound identification workflows where thousands of DFT Hessian calculations are prohibitive in terms of computational effort. In particular the GFN2–xTB method seems to be promising for this task due to a sophisticated description of the electrostatic energy terms and an excellent cost/accuracy trade-off. If higher accuracy is required, low-cost DFT or even higher level *ab initio* methods can be used for smaller sets of compounds. Furthermore, for flexible molecules it is important to investigate conformational ensembles.

Empirical adjustment of the atomic masses provides a robust alternative to linear frequency scaling and is particularly promising at the SQM level. Frequencies might be further improved by combining the linear and mass scaling approaches, or by including anharmonicities explicitly. Pairwise atomic mass scaling factors would probably provide the highest flexibility, but would require modification of the QM codes and complicates the fitting procedure. First advances have also recently been made to obtain anharmonicity corrections *via* machine learning.[472] A combination of the mass scaling approach with machine learning could be potentially promising. However, these techniques only improve vibrational frequencies and in fact, larger deviations from experiment often results from mismatching intensities. A better description of molecular dipole moments and their derivatives is particularly problematic for low-level SQM methods with their small AO basis sets and in particular FFs due to the classical monopole models. Another

problem is the availability and quality of experimental reference data. For metal-containing substances very few gas-phase IR spectra are available. Some experimental spectra also show significant noise, which can produce artifacts in the similarity measures. A possible strategy for future work to address these problems might be the fit to very high level theoretical data which, however, will limit the size and number of the molecules considered.

Acknowledgments

This work was supported by the DFG in the framework of the "Gottfried Wilhelm Leibniz-Prize" awarded to S.G. and US National Institutes of Health Grant GM087714 awarded to D.F.G.

Supporting Information

Some additional revised results can be found in Appendix A4. Detailed results from this chapter are only available as electronic supporting information from https://pubs.acs.org/doi/10.1021/acs.jctc.0c00877. The `newspecmatch` source code and input structures can be obtained from GitHub under https://github.com/grimme-lab/newspecmatch.

6. High Accuracy Quantum-Chemistry-Based Calculation and Blind Prediction of Macroscopic pKa Values in the Context of the SAMPL6 Challenge

Philipp Pracht,[*] Rainer Wilcken,[†] Anikó Udvarhelyi,[‡] Stephane Rodde,[†] and Stefan Grimme[*]

This article is part of the J. Comput.–Aided Mol. Des. SAMPL6 special issue

Received 30th of May 2018, Published online 23rd of August 2018

Reprinted (adapted) with permission from[§]
Pracht, P.; Wilcken, R.; Udvarhelyi, A.; Rodde, S.; Grimme, S. *J. Comput.–Aided Mol. Des.* **2018**, *32*, 1139–1149.

DOI 10.1007/s10822-018-0145-7

Own manuscript contribution

- Performing all calculations of submission xvxzd
- Interpretation of the computed data
- Writing the manuscript

[*]Mulliken Center for Theoretical Chemistry, Institut für Physikalische und Theoretische Chemie, Rheinische Friedrich-Wilhelms-Universität Bonn, Beringstraße 4, 53115 Bonn, Germany
[†]Novartis Institutes for Biomedical Research, CH-4002 Basel, Switzerland
[‡]Novartis Pharma AG, Technical Research and Development, CH-4002 Basel, Switzerland
[§]Reproduced with permission from the Springer Nature Switzerland AG: Springer Journal of Computer-Aided Molecular Design.

Abstract

Recent advances in the development of low-cost quantum chemical methods have made the prediction of conformational preferences and physicochemical properties of medium-sized drug-like molecules routinely feasible, with significant potential to advance drug discovery. In the context of the SAMPL6 challenge, macroscopic pK_a values were blindly predicted for a set of 24 of such molecules. In this paper we present two similar quantum chemical based approaches based on the high accuracy calculation of standard reaction free energies and the subsequent determination of those pK_a values via a linear free energy relationship (LFER). Both approaches use extensive conformational sampling and apply hybrid and double-hybrid density functional theory (DFT) with continuum solvation to calculate free energies. The blindly calculated macroscopic pK_a values were in excellent agreement with the experiment.

6.1. Introduction

A significant number of drugs on the market today contain ionizable functional groups.[513] Owing to the influence of ionization state on a range of ADME ("absorption, distribution, metabolism, excretion") properties from solubility to permeability and blood-brain-barrier penetration, pK_a values are routinely determined in the pharmaceutical industry alongside other physicochemical properties like logD. New methods that can accurately predict pK_a values for drug-like[514] molecules in water but also in non-aqueous solvents and solvent mixtures have great utility across the industry. It is therefore not surprising that historic SAMPL challenges (SAMPL0–SAMPL5)[515–521] have featured the prediction of physicochemical properties such as solvation free energies and distribution coefficients alongside prediction of host-guest complex affinities. In the current SAMPL6 blind test[514] macroscopic pK_a values should be calculated for a set of 24 medium sized, drug-like molecules. While a microscopic pK_a is specific for each functional group of a molecule and refers to the deprotonation at this position, the macroscopic pK_a is defined by the dissociation constant of deprotonation regardless of from which functional group the proton dissociates.[522–524] Hence, the macroscopic pK_a is directly related to the standard reaction free energy for the general loss of a proton. In the following we apply three computational schemes for the quantum chemical calculation of macroscopic pK_a values of the 24 SAMPL6 molecules, which correspond to the Type III submissions "xvxzd", "yqkga" and "8xt50". Due to the independence of the macroscopic pK_a from the deprotonation position it is necessary to use the Boltzmann average of all so-called microstates, *i.e.*, the differently protonated and deprotonated subspecies of each molecule. Different conformations and tautomers might also contribute to the standard reaction free energy. Therefore, our approaches are based on extensive sampling of the chemical ensemble, consisting of the different conformers for all tautomers and microstates. To obtain high accuracy of the standard reaction free energies we apply hybrid and double-hybrid DFT, including entropic corrections to the free energy in the rigid-rotor-harmonic-oscillator (RRHO) approximation and corrections to the free solvation energy calculated with continuum

solvation models for water.

6.1.1. Theoretical Details

Boltzmann populations for the ranking of different conformations are calculated from the free energies according to

$$p_i = \frac{\mathrm{e}^{-\epsilon_i/k_BT}}{\sum_{j=1}^{N} \mathrm{e}^{-\epsilon_j/k_BT}} \ , \tag{6.1}$$

where p_i is the population of state i, k_BT is the Boltzmann constant times temperature and ϵ_i is the energy of state i, which is calculated as the free energy G according to

$$G = E_{el} + G^T_{RRHO} + \delta G^T_{solv}(X) \tag{6.2}$$

and consists of the electronic energy E_{el}, an entropic and thermostatistical contribution calculated in the modified rigid-rotor-harmonic-oscillator approximation[42] G^T_{RRHO} and a solvation contribution $\delta G^T_{solv}(X)$ for the solvent X.
The macroscopic pK_a value is directly related to the dissociation free energy ΔG_{diss} attributed to the loss of a proton. This can be expressed as the chemical equation

$$\mathrm{A}^n\mathrm{H}_{solv} + \mathrm{B}^m_{solv} \rightarrow \mathrm{A}^{n-1}_{solv} + \mathrm{B}^{m+1}\mathrm{H}_{solv} \tag{6.3}$$

where the proton dissociates from species A, reducing the molecular charge from n to n-1. The proton is then absorbed by B, which could for example be a solvent molecule with the molecular charge m. The dissociation free energy can then be calculated in a thermodynamic cycle[525–527] from the individual free energies

$$\Delta G_{diss} = (G^{\mathrm{A}^{n-1}} - G^{\mathrm{A}^n\mathrm{H}}) + (G^{\mathrm{B}^{m+1}\mathrm{H}} - G^{\mathrm{B}^m}). \tag{6.4}$$

To obtain the ΔG_{diss} for a macroscopic pK_a, the correctly averaged and Boltzmann weighted free energies G have to be used in Eq. 6.4. pK_a values can be calculated from ΔG_{diss} using the linear free energy relationship (LFER)[252]

$$\mathrm{p}K_a = c_1 \frac{\Delta G_{diss}}{\ln(10)RT} + c_0, \tag{6.5}$$

where RT is the ideal gas constant times the temperature and the parameters c_1 and c_0 are fitted to experimental data.

6.2. Methodology

6.2.1. Fully Quantum Chemical Calculation of Macroscopic pKa Values (submission xvxzd)

Since macroscopic pK_a values are related to the dissociation free energy ΔG_{diss} via the LFER, it is possible to obtain pK_a values from quantum chemical calculations. It is also crucial to use only a single, correctly weighted ΔG_{diss}, since deprotonation at different positions could contribute to the macroscopic pK_a. Therefore all low energy conformers for all low energy tautomers for different microstates have to be determined and weighted according to Eq. 6.1.
For the first approach (submission xvxzd) in this work we use the semiempirical tight binding method GFN–xTB[38] in combination with high level quantum chemical calculations for the prediction of macroscopic pK_a values. The LFER parameters c_1 and c_0 are fitted at the same level of theory to a set of 59 small molecules, taken from related pK_a studies.[252,528]

Generation of Conformers, Tautomers and Protonation States

The procedure is based on a correct averaging of the free energy for all relevant neutral and ionic structures. Hence, determining correct conformations, tautomers and ionic microstates is the major prerequisite for the success of the pK_a calculation. The general workflow is outlined in Fig. 6.1.

Starting from a provided SMILES identifier string, three-dimensional structures for the 24 SAMPL6 molecules were created in the neutral state. From these structures, an initial conformational search was conducted using the recently published MF-MD-GC//GFN–xTB workflow.[40] Most notable in this MF-MD-GC procedure is the mode following (MF) approach, in which new conformations are obtained from the minima on 1-dimensional potential energy surfaces of the molecules normal modes (NM). This approach is physically motivated and generates conformations directly on the semiempirical GFN–xTB level, which also enables the search to be conducted with an implicit solvation model (generalized Born solvent accessible surface area, GBSA) for water to obtain solution geometries. The conformational search yields an initial conformer ensemble (CE) that is used in the automated searches of prototropic tautomers and the determination of the protonation and deprotonation sites.
Protonation sites are determined by calculating the localized molecular orbitals (LMOs), adding a proton at the positions of π-LMO and lone pair centers and then screening over all the newly generated protomers using the GFN–xTB method.[38,220] Deprotonation sites are generated with a similar screening procedure, but instead of using the LMOs one simply has to remove the protons at different positions to generate anions. Both procedures are automated and can use an input CE to enhance the respective search space. The determination of tautomers is done by a sequence of automated protonation and deprotonation as described above and yields all the permuted prototropic structures, as well as possible ring-chain tautomers. This procedure is also automated and will be published in detail elsewhere. From the newly generated tautomers the

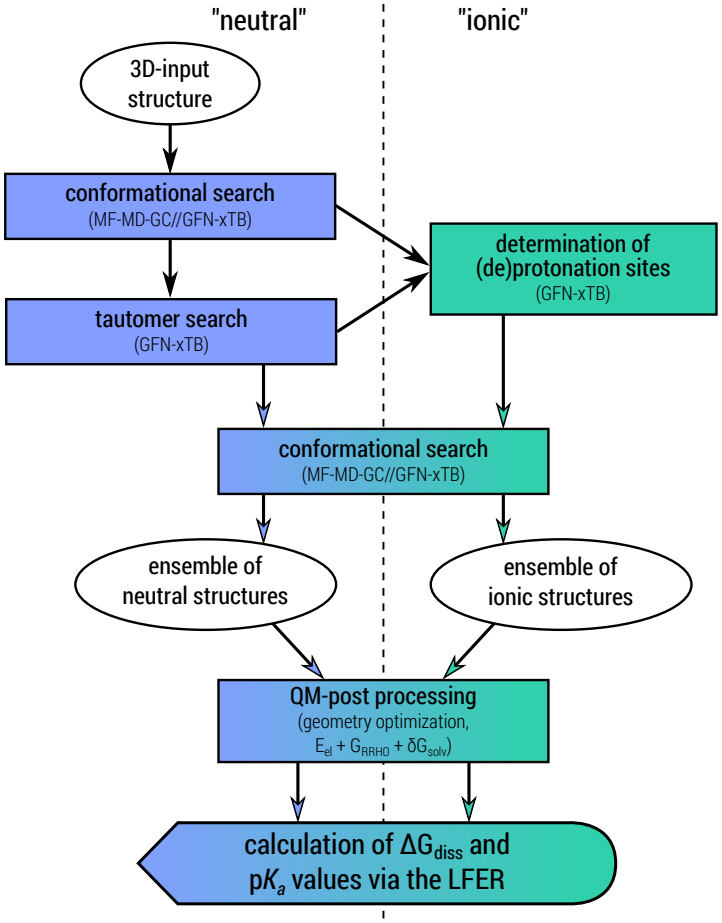

Figure 6.1.: Workflow from an initial 3D-input structure to the macroscopic pK_a value, involving several automated GFN–xTB procedures.

protonation/deprotonation site search is conducted again. Since only representative molecules for the ionic and tautomeric species are created in the procedures, another conformational search with the MF-MD-GC//GFN–xTB workflow has to be conducted for every one of these structures.

One obtains ensembles consisting of conformers for several tautomers, *i.e.*, one such ensemble for the neutral, one for the singly protonated and one ensemble for the singly deprotonated state. Energy thresholds have to be applied in each of the different automated procedures to limit the number of generated structures in each ensemble to the energetically most accessible species. For ionic microstates and tautomers this threshold was set to 10 kcal mol^{-1} and for conformers to 6 kcal mol^{-1} as default. For each of these three ensembles a high level total free energy has to be calculated using the Boltzmann weighted total free energies of each conformer, tautomer and protomer calculated according to Eq. 6.2.

Free energies were obtained within an multilevel ansatz that was already successfully applied in the calculation of spin-spin-coupled nuclear magnetic resonance spectra (NMR) (see Ref. 40):

- GFN–xTB geometries are reoptimized at the PBEh-3c[158] level of theory, including

DCOSMO-RS[340], the self-consistent implementation of the COSMO-RS[218,529] implicit solvation model for water.

- Total energies E_{el} are calculated with the DSD-BLYP-D3(BJ)/def2-TZVPD double-hybrid density functional.[148,172,173,530]
- Solvation contributions $\delta G_{solv}^{298\mathrm{K}}$ are calculated using the COSMO-RS(fine) continuum solvation model for water.[218,529]
- Entropic contributions $G_{RRHO}^{298\mathrm{K}}$ are calculated in the rigid-rotor-harmonic-oscillator approximation using GFN–xTB(GBSA).[38]

The Boltzmann weighted free energies are then used for the calculation of ΔG_{diss} and the macroscopic pK_a via the LFER.

Fit to Experimental Data and pKa Calculation

For the calculation of macroscopic pK_a values with the LFER it is necessary to determine the parameters c_1 and c_0. In common practice both values are fitted to experimental pK_a values and calculated dissociation energies. It is crucial to be aware that such a fit is not universally applicable to any standard reaction free energy calculation. The fit reflects the level of theory at which the standard reaction free energies for the reference molecules were calculated and thus must only be used for calculations of the exact same level.
We fitted the LFER to a set of 59 small organic and inorganic molecules taken from related literature[252,528] where the level of theory corresponds to the one described above. The fit set is mainly composed out of Klamt's dataset for pK_a calculation[252] and therefore exclusively contains small acids. We chose this dataset to have a better comparison between our results and the COSMOtherm pK_a calculations (as used in submissions yqkga and 8xt50), since these calculations are based on a LFER fit with the very same set of molecules. The fit is shown in Fig. 6.2.

From Fig. 6.2 a clear linear correlation between the experimental pK_a and the calculated dissociation free energy can be seen. The corresponding R^2 value is 0.896. The LFER parameters were determined from the linear fit as $c_1 = 0.5665$ and $c_0 = -1.1473$.
Several outliers can be observed for smaller pK_a values that correspond mostly to small inorganic acids. Another outlier for which we have no detailed explanation so far is for the dimethadione molecule, which has a standard reaction free energy of 19.5 kcal mol^{-1}, but only shows a experimental pK_a of 6.10. The fit set of molecules including the calculated standard reaction free energies is listed in Appendix A5.

Since pK_a values depend on the pH value, it is possible to observe more than one macroscopic pK_a, depending on the protonation state of the molecule. Therefore dissociation free energies

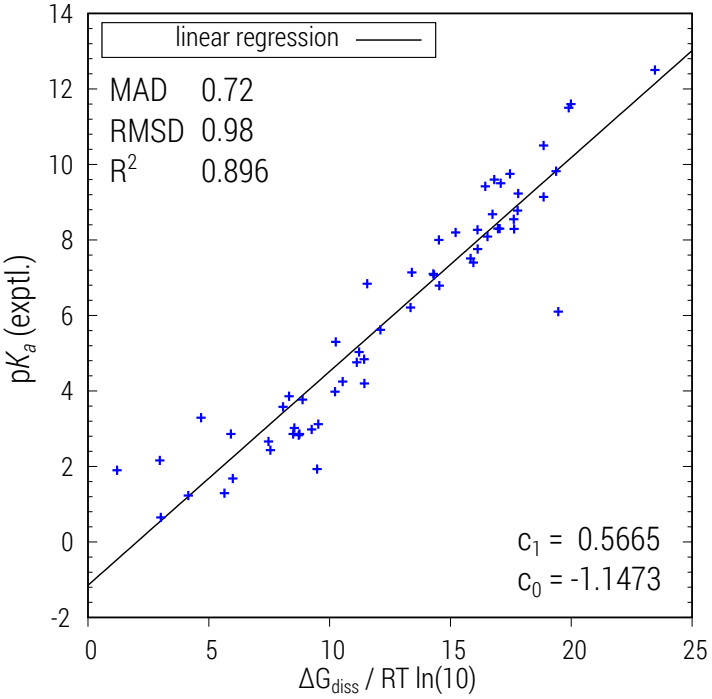

Figure 6.2.: Fit to experimental data to obtain the LFER parameters c_1 and c_0.

were calculated for two different types of reactions:

$$\mathrm{AH}_{solv} + (\mathrm{H_2O})_{4,solv} \rightarrow \mathrm{A}^{-1}_{solv} + \mathrm{H_3O^+}(\mathrm{H_2O})_{3,solv} \tag{6.6}$$

$$\mathrm{AH}^{+}_{solv} + (\mathrm{H_2O})_{4,solv} \rightarrow \mathrm{A}_{solv} + \mathrm{H_3O^+}(\mathrm{H_2O})_{3,solv} \tag{6.7}$$

In the first case the neutral structure is the starting point, while in the second case the input is the protonated species AH^{+}_{solv}, which would typically be expected for smaller pH values. For each molecule, two macroscopic pK_a values are obtained in this way. Doubly protonated and deprotonated microstates were not included in this study, but can be handled with the same computational protocol. Since all calculations were performed in scope of a blind challenge, including multiply charged ions would have required the additional application of our protocol to every doubly protonated and deprotonated microstate of all 24 molecules, which naturally leads to huge computational effort. By omitting these calculations we tried to limit the computational cost. Also, we expected large standard reaction free energies for the multiply charged ions and therefore only few of the molecules to show additional macroscopic pK_a values in the range of pK_a 2 to pK_a 12. It is evident from the results of our other submissions(yqkga and 8xt50), that this is true since only two out of 24 molecules (SM14 and SM18) turned out to have relevant doubly charged species for which pK_a values could be measured.
As can be seen in the chemical equations 6.6 and 6.7, a small water tetramer was chosen as the proton adsorbant. In principle it is also possible to use other reference molecules or just a single

water monomer, or even calculating the "raw" deprotonation without any baseline, as long as these calculations are also consistently performed throughout the fit. The small water cluster was chosen since it better reflects the experimental setup for the proton state in solution than an isolated water molecule.

Computational Details

All GFN–xTB calculations were conducted with version `xtb 4.9.4` code. Automated procedures for conformer, tautomer and protonation site search were implemented in standalone Fortran codes. The PBEh-3c geometry optimizations including the self-consistent DCOSMO-RS implementation were conducted with the *TURBOMOLE 7.2* program package[433]. The calculation of the solvation free energy contributions within the COSMO-RS framework were performed with *COSMOtherm*, using the 2014 *fine*-parametrization. DSD-BLYP-D3(BJ) calculations were performed with the *ORCA 4.0.1* program package[531].

6.2.2. ReSCoSS Workflow and COSMOtherm pKa Calculations (Submissions yqkga and 8xt50)

Determining Relevant Microspecies

Starting from the neutral form of each of the 24 SAMPL6 pK_a molecules, we generated plausible tautomer structures manually and repeated the tautomer analysis for singly and where reasonable doubly charged species. In order to allow a meaningful comparison of relative tautomer energetics, we used the recently developed ReSCoSS (short for "Relevant Solution Conformer Sampling and Selection") workflow which is discussed in detail elsewhere.[532]
Briefly, starting from 2D structures, 3D conversion was done using the *CORINA* software[533] and a full conformational search was carried out using Schrodinger MacroModel[534] with the Monte Carlo multiple minimum (MCMM) method,[535,536] the OPLS2005 force field,[74,76,537] including the GBSA implicit solvation model for water, while the all-atom RMSD threshold set to 0.75 Å and the potential energy cutoff increased to 30 kJ/mol. The geometries of all conformers for each microspecies (protomers of tautomers) were then optimized using the GFN–xTB method also employing a GBSA continuum solvation model for water[38]. Following a BP86/TZVP/COSMO[125,127,538–540] single point calculation, the conformations were clustered by shape diversity and within each cluster, any conformer corresponding to the lowest-energy conformation in either of 10 different COSMO-RS solvents (water, DMSO, hexane, octanol, methanol, propan-one, ammonia, acetonitrile and vacuum) was deemed relevant and retained. Single point electronic energies were then computed at the PW6B95-D3(BJ)/TZVPD[530,541] level and combined $G^{298\mathrm{K}}_{RRHO}$ at the GFN–xTB[GBSA(water)] level and solvation free energies at the COSMO-RS/FINE17 level. The total free energy of each conformer was obtained from a summation of these three terms according to Eq. 6.2. Only microstates where at least one conformer was within 10 kcal mol^{-1} compared to the minimum-energy microstate conformer were retained.

Full Quantum Chemical Optimizations of Relevant Conformer Sets

For all microstates carried forward from the first step, all chosen conformers were then fully optimized at the PBE-D3(BJ)/def2-TZVP/COSMO[128,172,173,338,540] level using the *TURBOMOLE 7.2* program package.[433] Total free energies according to Eq. 6.2 were then calculated at the DSD-BLYP-D3(BJ)/def2-TZVPD//PBE-D3(BJ)/def2-TZVP/COSMO level, including G^{298K}_{RRHO}[GFN–xTB(GBSA)] and δG^{298K}_{solv}[COSMO-RS(FINE17/TZVPD)] contributions.

Calculation of Macroscopic pKa Values

For submissions yqkga and 8xt50 we used COSMOtherm's own LFER fit, COSMOtherm,pK_a[252] as implemented in *COSMOtherm17*. Our two submissions differ in how COSMOtherm was applied to calculate the pK_a values. In the first submission (yqkga), we selected all conformers which had at least 5 % Boltzmann weights at the DSD-BLYP-D3/def2-TZVPD+RRHO(GFN–xTB)+COSMO-RS(fine) level for each microspecies and applied COSMOtherm pK_a directly to those sets of conformers. This led to an internal re-weighting of the conformers within COSMOtherm at the so-called *fine* level which employs a standard GGA functional[125,127] without dispersion corrections (BP86/TZVPD), but since the selection of the input conformations was done according to the Boltzmann weights from the higher-level QM method, the calculated pK_a values should still be influenced by the conformer selection strategy outlined before. In the second submission (8xt50), we computed conformationally aware pK_a values with COSMOtherm at the conformer level using the conformer Boltzmann weights at the DSD-BLYP-D3/def2-TZVPD+RRHO(GFN–xTB)+COSMO-RS(fine) level and the equations from Bochevarov *et al.*[542] to obtain final pK_a values.

6.3. Results and Discussion

6.3.1. Results of Submission xvxzd

The SAMPL6 molecule set[514] consists of 24 medium sized drug like molecules, of which most show several different protonation/deprotonation sites. Most of the 24 molecules can also be expected to have several populated conformations at ambient temperature in water. Hence the described computational protocol was used to generate, optimize and weight the different structures for each of the blind-test molecules and then calculate macroscopic pK_a values from their free energies. In the following we refrain from a detailed discussion of the chemical ensembles, but give some information in Appendix A5. The calculated pK_a values are presented in Table 6.1.

Overall the mean absolute deviation (MAD) and root-mean-square deviation (RMSD) have the values of 0.579 and 0.680 respectively, and the determination coefficient R^2 is 0.937. With this result, the best overall agreement between theory and experiment in scope of the SAMPL6 challenge was achieved. The correlation plot is shown in Fig. 6.3.

Table 6.1.: Calculated macroscopic pK_a values in comparison with experimental data.[45,46]. The two missing values (submission xvxzd) for molecules SM14 and SM18 stem from the doubly charged ions, which were neglected for pK_a calculation.

molecule	pK_a(xvxzd)	pK_a(yqkga)	pK_a(8xt50)	pK_a(exptl.)
SM01	10.14	9.69	9.69	9.53
SM02	4.93	6.26	6.27	5.03
SM03	7.52	6.92	7.18	7.02
SM04	5.17	7.27	6.84	6.02
SM05	4.36	4.85	4.58	4.59
SM06	3.41, 11.25	4.45, 12.51	4.72, 13.17	3.03, 11.74
SM07	5.43	7.43	6.94	6.08
SM08	5.72	5.33	4.67	4.22
SM09	5.07	6.78	7.00	5.37
SM10	8.27	7.97	10.32	9.02
SM11	3.95	4.08	3.87	3.89
SM12	5.04	6.48	6.55	5.28
SM13	5.14	7.10	7.82	5.77
SM14	(–), 4.62	3.38, 5.17	3.38, 5.17	2.58, 5.30
SM15	4.18, 9.35	4.44, 9.19	4.44, 9.19	4.70, 8.94
SM16	4.99, 9.59	5.98, 11.59	6.05, 12.27	5.37, 10.65
SM17	2.07	3.78	3.41	3.16
SM18	2.08, 8.52, (–)	2.14, 7.53, 9.26	2.12, 8.74, 10.54	2.15, 9.58, 11.02
SM19	8.56	9.14	12.55	9.56
SM20	6.52	4.92	5.89	5.70
SM21	3.86	4.17	4.24	3.86
SM22	3.09, 6.90	0.06, 7.32	1.17, 6.44	2.40, 7.43
SM23	4.52	4.79	4.54	4.52
SM24	2.61	2.46	2.24	2.60
MAD	0.58	0.80	0.81	—
RMSD	0.68	1.01	1.07	—

The smallest deviation from the experiment can be observed for molecule SM24, where the pK_a value was overestimated only by 0.01 units. The largest disagreement between theory and experiment was observed for molecule SM08 with 1.5 pK_a units. Out of 29 predicted macroscopic pK_a values, 13 are within the $\leq$0.5 pK_a confidence interval, 12 are within 0.5 and $\leq$1.0 pK_a confidence and only 4 predictions show a deviation >1.0 pK_a units, which demonstrates the predictive power of our composite quantum chemical approach.

Interestingly, the MAD and RMSD values for the SAMPL6 set are slightly smaller than the ones we achieved on our calibration set. This is even more surprising when considering that the fit molecules are mostly small acids, while in the blind test typical drug like molecules were given. As for the time of writing we have no explanation for this behavior, but apparently the LFER approach is depending much more on the accuracy of ΔG_{diss} than it depends on the

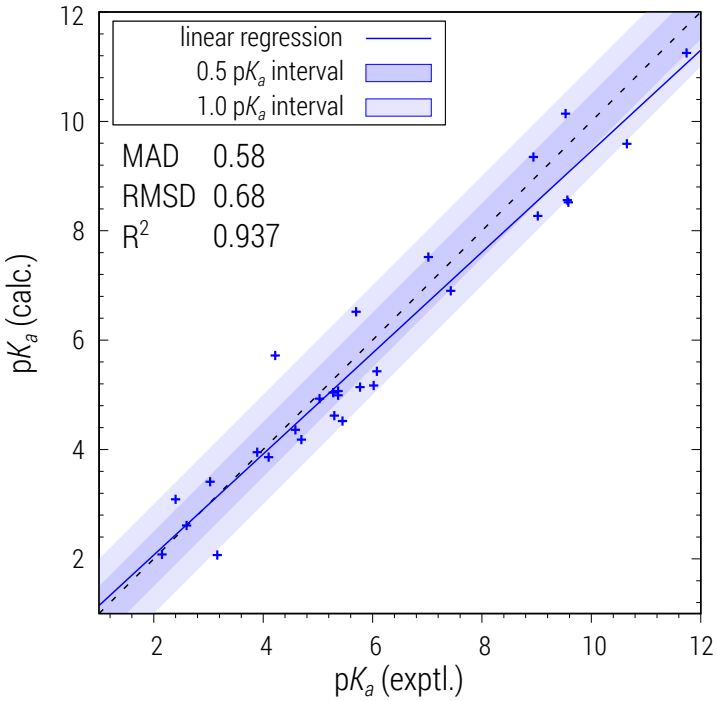

Figure 6.3.: Correlation plot for the calculated (submission xvxzd) and experimental macroscopic pK_a values of the SAMPL6 challenge molecules.

size and diversity of the reference molecules. Using the LFER to determine pK_a values from dissociation energies is an established approach that has already been published several times before using various levels of theory.[248–250,252,526,543–545] In contrast to most (but not all[249,542]) of these studies we tried to include a complete chemical ensemble for each molecule, consisting of conformers, tautomers and protomers obtained from quantum chemical calculations. In our opinion the inclusion and right averaging of species in those chemical ensembles and pairing this with high level quantum chemical calculations are the major reasons for the success of our approach. The QM calculation of such ensembles also makes it possible to take into account temperature via frequency calculations and solvation effects via an implicit solvent model, which would be neglected by chemoinformatic approaches, for example.

Conformations, tautomers and protomers were obtained at the GFN–xTB level of theory, including implicit solvation via the GBSA model. Structure generation and screening at this level was shown to be a well working procedure, which can be used as a starting point for higher ranking DFT calculations. Hence, the desired accuracy of the pK_a values is an interplay between a good chemical structure ensemble and the high level DFT accuracy. Since the energetic difference between different conformations can reach several kcal mol^{-1}, a flawed conformational ensemble would quickly show up in strongly deviating macroscopic pK_a, barring unfortunate error compensation. Inaccurate conformational energies would have the same effect, which is why a high level of theory should be used. DSD-BLYP-D3, for example, was chosen because this density functional showed the best performance for conformational energies on the GMTKN55

benchmark set[109]. However, concerning the calculation of pK_a values via the LFER, the influence of the DFT level of theory is comparably small. Since the data is fitted to experimental pK_a values, useful results can be obtained with a variety of functionals, of which we tested the PBEh-3c, BP86[125,127] and PW6B95[541] functionals.
In literature[252,525] the constant c_0 has been reported to have the theoretical value $c_0 = -\log[H_2O]$ $= -1.74$, but depending on the chosen method or density functional large deviations from this value are obtained in the fitting procedure. Even values with the opposite sign have been published.[252] The best match with the value we encountered so far is the result for our fit, where $c_0 = -1.1473$. It should be mentioned, however, the constant c_0 is an ill-defined expression (since it is not possible to take the logarithm of a unit) and thus has no physical meaning apart from defining a correction to the arbitrary baseline within the LFER. Therefore c_0 should not be used as a measure for the correlation between level of theory and the quality of the LFER fit.

The inclusion of the RRHO term and the entropic contribution at finite temperature has only a small effect on the calculated pK_a values, which is why it has been completely neglected in some studies.[252,543] The main component of the RRHO contribution to a proton dissociation free energy (or proton affinity) is the loss of zero-point vibrational energy of the cut X-H bond. This value, however, is similar (but of opposite sign) when the bond is formed with the conjugated base and hence the overall effect is negligible in a LFER treatment. However, we were able to see a slight overall improvement by including it. For example, omitting the RRHO term leads to an MAD of 0.77 pK_a units and and RMSD of 1.10 pK_a units on the used fit set, while the c_0 parameter increased to -0.5847. Higher level RRHO calculations might even further improve the results, but were not tested due to the immense computational effort.

A larger error source is attributed to the accuracy of the solvation free energies, calculated with the continuum solvation models. Omitting the solvation terms in the conformer search leads to very different conformations, and hence must never be neglected for the pK_a calculation. Since the mere presence of a electrostatic screening continuum influences the conformation, even simple models as the GBSA are sufficient for the geometry generation, *i.e.*, to generate good starting conformations representing a CE in solution. Concerning the free solvation energies however, we noticed in the early stages of the project that even the standard COS-MO-RS parametrization (compared to the *fine* parametrization) yields unreasonable free total energies. Therefore we have used only the COSMO-RS *fine* parametrized model for the final calculations, the SMD model[546–548] showed comparable performance however and could alternatively be used, which was also tested in the early stages of this work. The accuracy of solvation free energies is still a limiting factor and in our opinion gives the largest room for improvements.

6.3.2. Results of Submissions yqkga and 8xt50

Over the whole dataset our two submissions yqkga and 8xt50 achieved rankings #4 and #6 respectively, representing the best QM-based predictions behind the winning submission xvxzd but not quite matching the two best QSPR-based methods. The calculated macroscopic pK_a

values are shown in Table 6.1 and the respective correlation plots are shown in Fig. 6.4 and 6.5.

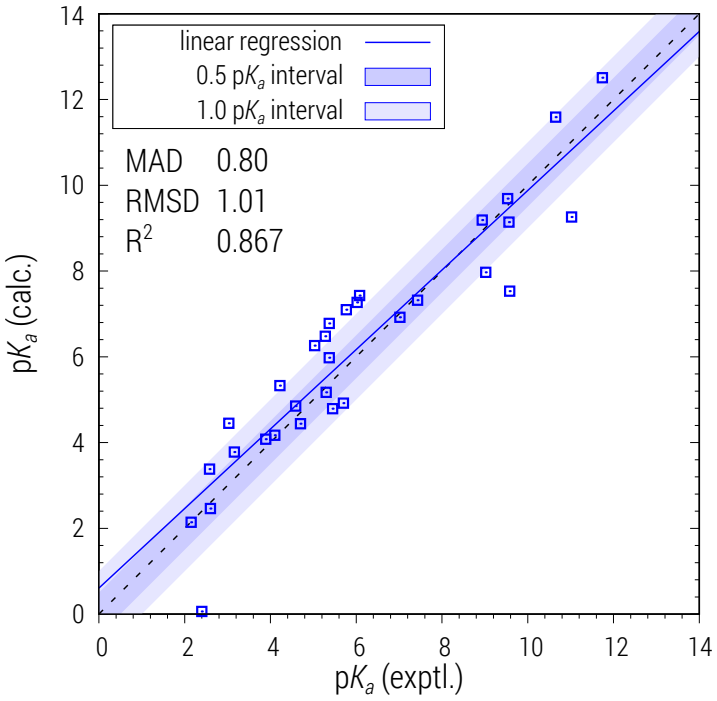

Figure 6.4.: Correlation plot for the calculated (submission yqkga) and experimental macroscopic pK_a values of the SAMPL6 challenge molecules.

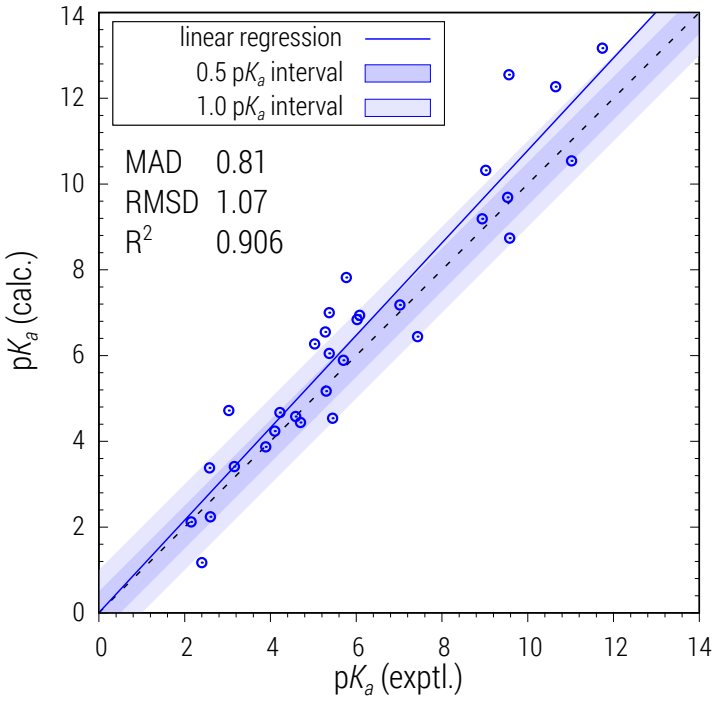

Figure 6.5.: Correlation plot for the calculated (submission 8xt50) and experimental macroscopic pK_a values of the SAMPL6 challenge molecules.

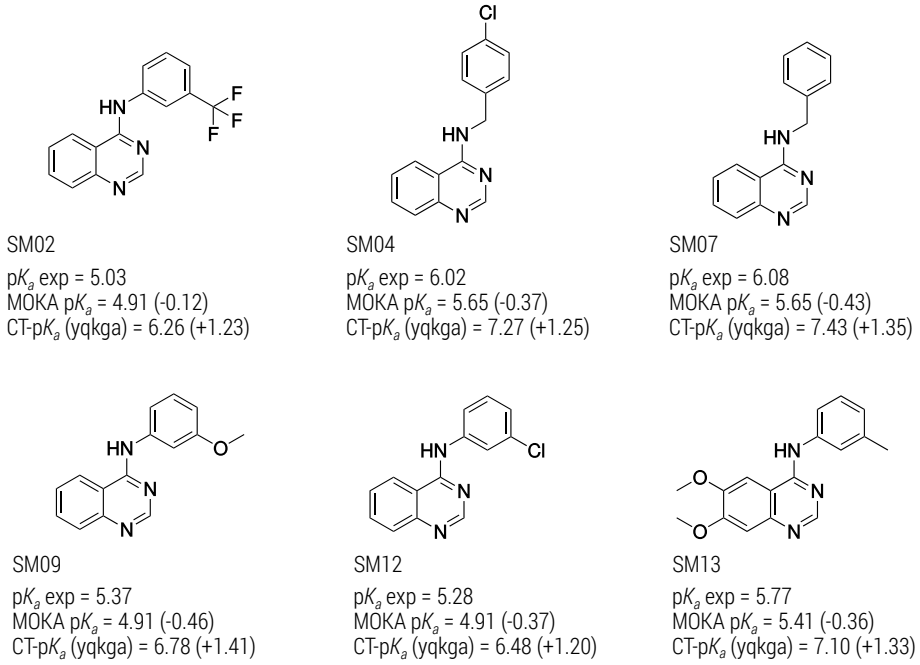

Figure 6.6.: Comparison of an empirical method (MoKa) and QM-based pK_a predictions (COSMOtherm) for six related relatively simple molecules from SAMPL6. The molecules are predicted by both methods to be protonated at the ring nitrogen opposite the amino group resulting in 4-aminoquinazolin-1-ium species.

Analysing the 24 molecules from the SAMPL6 pK_a challenge in more detail, it is apparent that there are several subgroups of highly similar molecules. For instance, six out of 24 molecules – SM02, SM04, SM07, SM09, SM12 and SM13 – share the same aminoquinazoline scaffold which is very common in drug-like molecules and should be well parameterized in the QSPR codes due to availability of experimental data. Indeed, retrospective analysis with MoKa 2.5.4[549] reveals that these are well captured (MAD = 0.35 units) while our COSMOtherm pK_a values from submission yqkga seem to be systematically off (MAD = 1.30 pK_a units for this subset; Fig. 6.6). On the other hand, our procedure does identify the correct protonation site on the scaffold as determined experimentally by NMR[45,46] – protonation leading to the 4-aminoquinazolin-1-ium species – and as the ionization constants for these molecules are much better described in our other submission xvxzd, the most likely cause for this systematic deviation is the COSMOtherm LFER fit. We believe that a re-fit of the COSMOtherm pK_a LFER based on higher-level QM calculations – for instance, using hybrid functionals such as PBE0[337,550] or PW6B95 with a large basis set and, most crucially, including D3 dispersion – could significantly improve its performance.

SM18 is arguably the most interesting and complex of the 24 molecules, both with regards to its conformational flexibility and its possible tautomerism. The Quinazolinone moiety alone could exist as Quinazolin-4(1H)-one, Quinazo-lin-4(3H)-one or Quinazoline-4-ol tautomer in the neutral state, and there is additional potential for tautomerism in the thiazolylamide moiety.

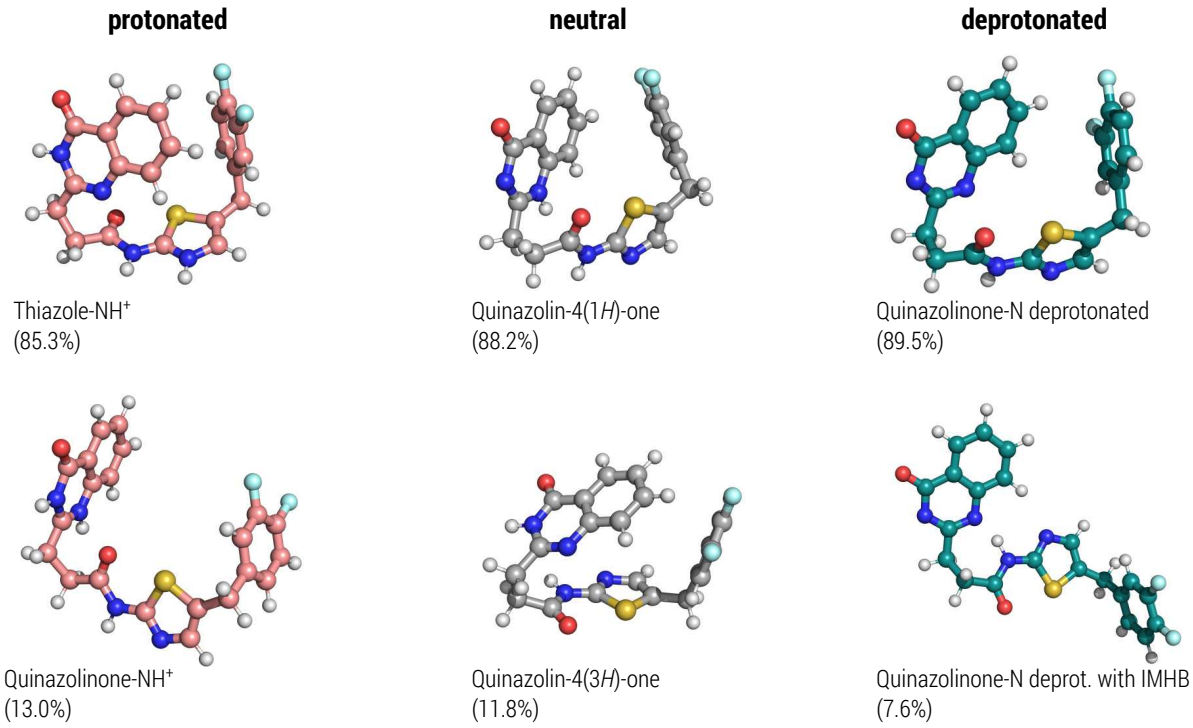

Figure 6.7.: 3D conformations of the dominant conformer in each relevant microspecies predicted by our DSD-BLYP-D3/def2-TZVPD+RRHO(GFN–xTB)+COSMO-RS(fine) scheme based on the ReSCoSS workflow. Percentages indicate the sum of the Boltzmann weights for all related species (*i.e.*, all conformations of that microspecies), but only the energetically favoured conformation is depicted

We enumerated a total of 16 microstates spanning formal charges -2 to +2 for this molecule. Through our workflow we predicted that only two tautomers are relevant in the neutral state, namely the two Quinazolinone forms, with the Quinazolin-4(1H)-one form dominating over the (3H) form. We predicted that the lowest pK_a (exp: 2.15; 8xt50: 2.12) belongs to the protonation of the molecule which should occur predominantly (85.3 %) at the thiazole moiety; that the second pK_a (exp: 9.58, 8xt50: 8.74) represents the first deprotonation step taking place at the Quinazolinone moiety with a second deprotonation step to the doubly negatively charged species (exp: 11.02, 8xt50: 10.54) at the remaining amide moiety. Fig. 6.7 illustrates the conformations of the main species predicted to be relevant in the neutral and singly ionized states of SM18.

6.4. Conclusion

Ionizable groups are very common motifs in drugs[513] owing to their significant influence on solubility, permeability, and biological activity of a molecule. Therefore it is not surprising that significant efforts have been made in the development of pK_a prediction programs, usually with focus on fast structure-based QSPR codes that can predict aqueous pK_a in a matter of seconds. We have previously reported on our internal efforts to re-train the MoKa pK_a model based on internal data.[251,549,551,552]

QM-based pK_a predictions via LFER, while undoubtedly much slower than empirical QSPR, allow a very different perspective on the exact molecular structures involved – in terms of being expected to more or less accurately predict tautomers, protonation sites, and conformations in addition to working better for novel chemistry and mixed or non-aqueous solvents.

As is evident from our results in the SAMPL6 challenge, the fully quantum-chemically based calculation of macroscopic pK_a values for drug-like molecules is a competitive and general approach which works well for systems of medium size. It allows detailed insight into tautomeric states, conformers and protomers and should be easily generalized to non-aqueous solvents, making it attractive for application within the pharmaceutical industry.

Both the combined ReSCoSS+COSMOtherm and the GFN–xTB based approach show excellent MAD and RMSD, which led to ranks 1 (submission xvxzd), 4 (submission yqkga) and 6 (submission 8xt50) in the blind test. What these two approaches have in common is the focus on the best possible chemical ensemble consisting of the different conformations for different neutral, protonated and deprotonated forms of a molecule. By selecting efficient quantum chemical methods, an acceptable computation time could be achieved so that a macroscopic pK_a value for a molecule could be calculated within approximately one day on a 28 CPU node. This is just a crude estimate, since the computation time depends on the number of different structures included in the chemical ensemble, which can strongly vary depending on the molecule. Since the computational bottleneck for all methods presented here is the geometry optimization of the molecule conformations, the approaches are most likely limited to medium-sized systems with about 50-100 atoms, at least at this high level of theory.

Possible improvements could be made in terms of the solvation free energies, for which continuum models seem to be not accurate enough. In terms of computing time, however, implicit solvation models are still the only routinely applicable option for calculating free solvation energies, but explicit solvation models are an option for future studies.

Finally, the novel GFN2–xTB method[39] currently in development offers the possibility for future improvements. Compared to the GFN–xTB predecessor method, GFN2–xTB contains a number of theoretical innovations, the most important of which is a multipole expansion of the electrostatic term and the use of the newly developed D4 dispersion correction.[178] First tests show a general improvement of GFN2–xTB compared to GFN–xTB results at practically the same computational speed, even and especially for relative tautomeric and conformational energies. Thus further improvements in the calculation of macroscopic pK_a values could be achieved

in the future by replacing GFN–xTB by GFN2–xTB in the presented workflow.

Acknowledgements

This work was supported by the DFG in the framework of the "Gottfried Wilhelm Leibniz-Prize" awarded to S.G.

Supporting Information

Some additional supporting information can be found in Appendix A5.

7. Efficient Quantum-Chemical Calculations of Acid Dissociation Constants from Free Energy Relationships

Philipp Pracht* and Stefan Grimme*

Received 16th of April 2021, Published online 18th of June 2021

Reprinted (adapted) with permission from†
Pracht, P.; Grimme, S. *J. Phys. Chem. A* **2021**, *125*, 5681–5692.

DOI 10.1021/acs.jpca.1c03463

Own manuscript contribution

- Further development of the CREST code
- Performing and supervising the computations
- Interpretation of the computed data
- Writing the manuscript

*Mulliken Center for Theoretical Chemistry, Institut für Physikalische und Theoretische Chemie, Rheinische Friedrich-Wilhelms-Universität Bonn, Beringstraße 4, 53115 Bonn, Germany
†Reproduced with permission from the American Chemical Society.

Abstract

The calculation of acid dissociation constants (pK_a) is an important task in computational chemistry and chemoinformatics. Theoretically and with minimal empiricism, this is possible from computed acid dissociation free energies *via* so-called linear free energy relationships. In this study some modifications are introduced to the latter, providing a straight-forward, broadly applicable protocol with adjustable degree of sophistication for quantum chemistry based calculations of pK_a in water. It targets a wide pK_a range (about 70 units) and medium-sized, flexible molecules. Herein, a focus is set on the recently published r^{2}SCAN-3c and related efficient composite density functionals and the semiempirical GFN2–xTB method including a newly introduced energy correction for heterolytic dissociation, both in combination with implicit solvation models. The performance is evaluated in comparison with experimental data, showing mean errors often smaller than a targeted 1 pK_a unit accuracy. Larger deviations are observed only upon inclusion of challenging highly negative (<-5) or positive (>15) pK_a values . Among all those tested, it is found that B97-3c is the best performing functional, although rather independently of the DFT method used, low root-mean-square errors of 0.8 to 1.0 pK_a units for typical drugs are obtained. For optimal performance, it is recommended to employ DFT functional specific free energy relationship parameters. Additionally, a significant conformational dependence of the pK_a values is revealed and quantified for some non-rigid drug molecules.

7.1. Introduction

Acid dissociation constants (pK_a) are among the most featured physicochemical properties in the literature.[251,553,554] The significance of pK_a values is owing to the importance of ionization states of drug molecules under physiological conditions and their direct influence on a range of other properties. In the pharmaceutical industry an understanding about a molecule's ionization behavior is critical for an accurate prediction and characterization, *e.g.*, of solubility, permeability, and an associated range of ADME (absorption, distribution, metabolism, excretion) properties.[251] This process can be aided by computational methods, which enable the *a priori* calculation of pK_a values without the need for costly and time consuming experimental titration studies.

Over the last few decades a wide variety of computational models have become available for pK_a calculations. The two arguably most common approaches are based on quantum mechanical (QM) free energy calculations using linear free-energy relationships (LFER)[553,554] and, more empirically quantitative structure-activity relationship (QSAR) and machine learning (ML) models,[555–557] which derive the pK_a from chemoinformatic rules and large collections of reference data. The latter approaches have an obvious advantage with regards to their computational cost and also enable the screening of pK_a values in large databases. However, many studies show that the achievable accuracy of QSAR and ML models strongly depends on their ability to recognize ionization sites and thus on the used test and training data sets.[245,557–559] QM based (LFER) models on the other hand, despite the much higher computational cost provide a high

degree of generality in the computational modeling and allow for more detailed investigations, *e.g.*, of conformational or stereochemical effects.[292,542,560,561] This furthermore includes the applicability to highly negative (<-5) or positive (>15) pK_a values, which is out of the usual range for highly parameterized QSAR/ML models.

Many different flavors of LFER schemes have been proposed over the years that differ with regards to their computational setup or parametrizations for different functional groups. However, any of these methods center on the accurate description of the dissociation free energy of a proton in solution (usually water) and, at least in principle, could be described by any wave function theory (WFT), density functional theory (DFT) or even semiempirical quantum chemical (SQM) method. The respective QM data may also be used as descriptors for QSAR modeling, *e.g.*, in the form of HOMO-LUMO gaps[562,563] or atomic charges.[564–566] One major source of error for computed pK_a values is the description of solvation effects, which are commonly treated by implicit solvation models[252,543,553,560,567–571], explicit solvation[572,573] or a combination thereof.[574–577] Nonetheless, predictions of the pK_a based purely on QM have been proven to work very well,[554,560] for example in comparison with experimental data in the SAMPL6[45,46] blind challenge. Here, LFER results based on QM provided by our group and others[221] made the best predictions compared to the experiment with errors smaller than 1 pK_a unit which is commonly considered as 'chemical accuracy'. As noted above, the main drawback of LFER/QM based calculations is the comparatively large computational cost required for the calculation of high quality dissociation free energies at DFT level. However, in recent studies, for example by Jensen *et al.*[249,560,566], it was shown that reasonable accuracy and errors of only 1–2 pK_a units can still be achieved using computationally much cheaper SQM methods. Overall, the *in silico* calculation of pK_a values is one of the most studied subjects in computational chemistry and chemoinformatics. Hence it is no surprise that a large number of pK_a prediction tools are commercially available, based either on the LFER or QSAR models (*e.g.*, COSMOtherm[252,543], ACD/pka[578], ChemAxon[579], SimulationPlus[580], Schrödinger[244,542,581], and Optibrium[557]). In contrast, no specialized software is required for QM free energy calculations and very efficient protocols exist that can automate major parts of these computations.[22] Furthermore, QM calculations are, at least in principle, also possible for molecules containing inorganic elements, for which no reference (training) data is available for QSAR/ML based models.

In this study we present the calculation of pK_a values by means of the recently introduced r^{2}SCAN-3c composite functional[182] using COSMO-RS implicit solvation[218,454] and by the GFN2–xTB SQM method[36,39] including the ALPB implicit solvation model.[22] A protocol for the routine calculation of the pK_a from dissociation free energies is discussed that involves free energy computations only for two species, *i.e.*, the acid and its conjugate base. In the following section a short review of the theory is given and a higher-order free energy relationship (FER) is introduced. Furthermore, conformational effects and a correction for GFN2–xTB dissociation energies are briefly discussed. The latter is first evaluated in Sec. 7.4.1 for a set of 171 small molecule pK_a values in comparison with r^{2}SCAN-3c/COSMO-RS. Afterwards, the performance of several theoretical methods is shown for different chemical functional groups in

Sec. 7.4.3, followed by challenging pK_a calculations for flexible molecules in Sec. 7.4.4. Finally, in Sec. 7.4.5 conformational effects are investigated.

7.2. Theory

The pK_a is the logarithm of the equilibrium constant K_a for the dissociation reaction of an acidic molecule HA in a given solvent (commonly water) according to the chemical equation

$$\mathrm{HA}^n + \mathrm{H_2O} \xrightarrow{K_a} \mathrm{A}^{n-1} + \mathrm{H_3O^+} \ , \tag{7.1}$$

where n is the molecular charge of the acid. For simplicity, it is assumed that $n = 0$ in this section. The reaction free energy of this dissociation, ΔG_{diss}, is related to the equilibrium constant via the Gibbs free energy relation $\Delta G = -RT \ln K_a$, which is applied to pK_a calculations in the form of a linear free energy relationship (LFER)

$$\mathrm{pK}_a = c_1 \frac{\Delta G_{diss}}{RT \ln 10} + c_0 \ . \tag{7.2}$$

In the literature many works based on the LFER are found,[542,543,553,560] where ΔG_{diss} is obtained from thermodynamic cycles and c_1 and c_0 are adjustable parameters, chosen as to best reproduce experimental reference pK_a values. Ideally, c_0 is unity if ΔG_{diss} correctly reflects the experimental conditions, but systematic errors of the theoretical methods may be compensated by adjustment of this parameter. Because according to Eq. 7.1 the dissociation free energy can be defined as

$$\Delta G_{diss} = \big(G(\mathrm{A^-}) - G(\mathrm{HA})\big) + \underbrace{\big(G(\mathrm{H_3O^+}) - G(\mathrm{H_2O})\big)}_{const.} \ , \tag{7.3}$$

the solvent, which is acting as the base, has a constant contribution to ΔG_{diss} and hence its effect can be entirely absorbed into a_1. This suggests employing $\Delta G'_{diss} = G(\mathrm{A^-}) - G(\mathrm{HA})$ directly instead of ΔG_{diss}.

The LFER assumes a linear relation for the entire range of HA dissociation energies. However, non-linear contributions, *e.g.*, in the solvation terms may appear for very high or very low pK_a values. Hence, as a simple alternative we propose a more flexible, n-th order polynomial FER given by

$$\mathrm{pK}_a = c_0 + c_1 \Gamma_{diss} + c_2 \left(\Gamma_{diss}\right)^2 + ... + c_n \left(\Gamma_{diss}\right)^n \ , \tag{7.4}$$

with the empirical fit parameters c_0 to c_n, and the reduced free energy variable $\Gamma_{diss} = \frac{\Delta G'_{diss}}{RT \ln 10}$. Higher-order free FER are known in physical chemistry, *e.g.*, from the Marcus theory, which employs a quadratic free energy relationship (QFER) formulation to describe electron transfers.[582] As for the LFER, the free energy contribution of the solvent acting as a base in the higher-order FER is assumed to be constant and absorbed into the fit parameters, *i.e.*, $\Delta G'_{diss}$ is used. For the final working equation this leads to only two required free energies G(HA) and $G(\mathrm{A^-})$, which can be routinely obtained from standard QM calculations in implicit solvation

according to

$$G = E_{el} + G^{T}_{trv} + \delta G^{T}_{solv}(S) \ . \tag{7.5}$$

Here, E_{el} is the electronic energy calculated by any QM or SQM method, G^{T}_{trv} is the ro-vibrational free energy contribution at finite temperature T calculated in a modified rigid-rotor harmonic-oscillation (mRRHO) approximation,[42,186,420] and $\delta G^{T}_{solv}(S)$ is the implicit solvation free energy for the solvent S, including volume work. Conformational effects may be included by a population averaged free energy $\overline{G} = \sum_i p_i G_i$, where the sum is taken over the entire conformational ensemble. The thermal populations p_i at absolute temperature T are given by

$$p_i = \frac{e^{-G_i\beta}}{\sum e^{-G_i\beta}} \ , \tag{7.6}$$

where $\beta = \frac{1}{RT}$, R is the gas constant, and G_i is the free energy (Eq. 7.5) of the equilibrium structure of conformer i. The extension to different tautomeric states can be very comprehensive by taking into account all the possible reactions of the tautomers of HA to all the tautomers of A^- (see, *e.g.*, the work of Bochevarov *et al.*[542]). For simplicity, additional tautomers are not considered in this study, and only the most relevant protonation sites for each molecule are employed. However, motivated by the fact that there is only one observable ΔG_{diss} for the reaction, $\overline{G}$ may simply be calculated by including all conformational ensembles for all tautomers. In general, the individual conformers and tautomers of the ensemble can be seen as free energy "levels" for the acid or its conjugate base. Hence, the ensembles define the pK_a value obtained from the respective Boltzmann population averaged free energies $\overline{G}^A$ and $\overline{G}^B$ as shown in Fig. 7.1. Furthermore, in order to improve the accuracy, separately computed conformational free energies[22,186,216] $G^{A/B}_{conf}$ could be added for the acid and the base. However, for the cause of efficient computations as subject of this study, this term and its influence on the pK_a will be omitted here and discussed elsewhere. In contrast, if only single, random conformations are chosen from the ensembles for the acid and the base, the calculated pK_a can lie anywhere between the limits given by the minimum and maximum possible dissociation free energies.

The calculation of pK_a values from free energies with either the LFER or higher-order FER is straightforward and works across a range of theoretical methods. Even with comparatively inexpensive SQM methods reasonably accurate results are obtained.[249,557,560] Because errors for chemical reaction energies are typically larger for SQM methods than for higher level (DFT) methods,[36,109] it seems appropriate to introduce an energy correction term specifically for acidic dissociation reactions. This holds specifically for the here considered GFN methods which have not been parameterized originally for thermochemical applications. The here newly proposed energy correction E^{TB}_{mod} depends on the chemical topology (via bond orders) and atomic charges in the acid/base species and is given by

$$E^{\mathrm{TB}}_{mod} = \varepsilon(X) + k_1 \mathbf{\Delta BO}_{ab}(X) + k_2 q_a(X) + k_3 q_b(X) + k_4 q_a(H) \ . \tag{7.7}$$

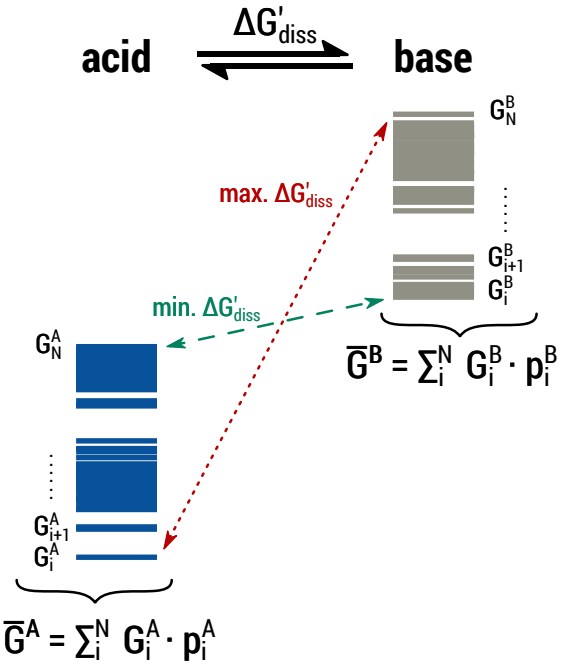

Figure 7.1.: Schematic representation of (free)energy levels for conformational ensembles of an acid and its conjugate base. The observable pK_a value is calculated from the population averages $\overline{G}^A$ and $\overline{G}^B$. The minimum and maximum possible dissociation free energies are visualized by the dashed lines.

In Eq. 7.7, X is the reactive atom from which the acidic proton is dissociated, $\varepsilon(X)$ is an element dependent energy shift, $\Delta\mathbf{BO}_{ab}(X)$ is the bond order difference of X between the acid and the deprotonated molecule, and q_a and q_b are the atomic charges of X and the proton in the acid or the base as shown schematically in Fig. 7.2. The four parameters k_{1-4} and the

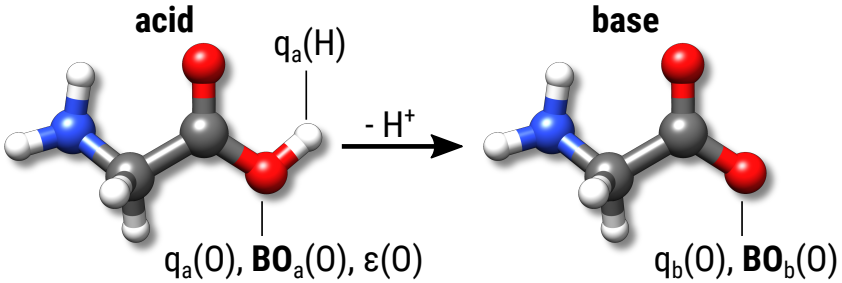

Figure 7.2.: Structural description of the parameters q_a, q_b, $\mathbf{BO}_a$, $\mathbf{BO}_b$, and ε for the example of glycine with X=oxygen in Eq. 7.7.

energy shift $\varepsilon(X)$ are fitted so that $\Delta E_{el}^{\mathrm{TB}} + E_{mod}^{\mathrm{TB}}$ at GFN2–xTB/ALPB(H_2O) level reproduces DFT r^{2}SCAN-3c/COSMO-RS(H_2O)//GFN2–xTB/ALPB(H_2O) dissociation energies (see below). The r^{2}SCAN-3c composite functional has been chosen because it is one of the generally most accurate, robust, and efficient DFT methods available and is our group default for general chemistry applications.

7.3. Computational Details

Calculations at the GFN2–xTB/ALPB(H_2O) level were conducted with version 6.4.0 of the `xtb` program. Higher-level calculations were mainly carried out at the r^{2}SCAN-3c/COSMO-RS(H_2O) level.[182] In Sec. 7.4.3 calculations were additionally performed at the B97-3c,[180] B97-D/def2-TZVPP,[171] PBE0-D3/def2-TZVPP,[172,173,337,338] PW6B95-D3/def2-TZVPP[541] and ωB97X-V/ def2-TZVPP[583] levels of theory, all including COSMO-RS(H_2O) using the 2019 "normal" parametrization. All COSMO-RS calculations use the density and basis set corresponding to the respective DFT level. For simplicity, in the following discussions the levels of theory are often just referred to by their underlying DFT functional or SQM method, although implicit solvation is always used. DFT calculations were done with Turbomole 7.5.1[432,493] and COSMO-RS implicit solvation contributions were calculated with COSMOtherm19.[529,584] Geometry optimizations were conducted at the r^{2}SCAN-3c/DCOSMO-RS[340] level, using the `xtb` program as a driver. Free energy contributions G_{trv}^{T} from a rigid-rotor harmonic-oscillator (RRHO) treatment were always calculated at the GFN2–xTB/ALPB(H_2O) level, using the recently introduced single-point-Hessian (SPH) approach[420] to avoid changes to the geometry and the appearance of imaginary modes. For the initial determination of E_{mod}^{TB} parameters in Sec. 7.4.1 r^{2}SCAN-3c/COSMO-RS(H_2O) energies are calculated on GFN2–xTB/ALPB(H_2O) geometries. All other DFT calculations were performed using the r^{2}SCAN-3c/DCOSMO-RS geometries. The calculation of pK_a values and related routines were implemented in a development version of the CREST program,[33,423] which was also used for generating molecular conformations at the GFN2–xTB/ALPB level. For post-processing of the final conformational ensembles and calculation of DFT energies the `censo` script[22,585] was used (version 1.0.5, `part1` and `part2`, default settings). The three simple steps required for the calculation of pK_a values are given by:

1. Starting structures for the acid and the base are determined and conformational ensembles are computed for both independently using CREST.
2. $\overline{G}^{A}$ and $\overline{G}^{B}$ at the DFT level are calculated using `censo` script[585] from the two GFN2–xTB/ALPB ensembles.
3. From the resulting $\Delta G'_{diss}$ the pK_a is determined using Eq. 7.4.

For further automation aspects and technical details, the reader is referred to Ref. 22 and Appendix A6.

7.3.1. Benchmark Sets

Four benchmark sets were used in this study. For the fit of GFN2–xTB dissociation energy corrections, element dependent parameters in Eq. 7.7 were determined for the elements C, N, O, F, Si, P, S, and Cl as possible deprotonation sites, using a set of 171 small molecules. A subset of this fit set consisting of 82 small molecules (elements HCNO only) was then used for the

determination of the parameters in Eq. 7.4, with reference pK_a values spanning the full experimentally known range from -24 (protonated benzene) to 50 (ethane). Because of this pK_a range we refer to the set as PKA74 in the following. In a cross-validation procedure FER parameters are also determined for a second, larger set, but with a smaller pK_a range. This set was composed by Thapa and Raghavachari[577] and contains 224 experimental pK_a for small molecules, distributed on 12 different functional groups (aliphatic alcohols R-OH, phenols Ph-OH, thiols R-SH, thiophenols Ph-SH, carboxylic acids R-COOH, benzoic acids Ph-COOH, primary R-NH_2, secondary R2-NH and tertiary amines R3-N, anilines Ph-NH_2, nitrogen containing hetereocycles Ar-N, and carbon acids R-C-H). In the following we refer to this set as TR224. For an independent validation for the final pK_a procedure, the second and third sets contain values for larger, drug like molecules, and also enable the study of conformational effects. One of these sets was introduced in the SAMPL6 challenge and contains 24 molecules with 31 experimental pK_a values.[45,46] The other was first introduced by Eckert and Klamt[543] and later revised by Jensen *et al.*[560] and contains 53 pK_a values for 48 molecules. Details are given in Appendix A6. For the sake of computational efficiency, only a single protonation site of each molecule as described in the literature was considered in this study and no additional tautomers were investigated. The respective structures were either taken from the literature[45,543,557,560,577] as Cartesian coordinates directly or converted from SMILES codes using OpenBabel.[586] However, we note that, if desired, protonation sites and tautomers could easily be screened[33,220] and included within the same computational framework.

7.4. Results

7.4.1. Free Energy Relationships and Corrected Dissociation Energies

Motivated by systematically underestimated and overall inaccurate heterolytic dissociation energies at the GFN2–xTB/ALPB(H_2O) level (see Appendix A6), an element dependent energy correction term E_{mod}^{TB} was developed and fitted to r^{2}SCAN-3c/COSMO-RS reference dissociation energies. Thermostatistic contributions G_{trv} were neglected here because they are always calculated at the same GFN2–xTB level of theory in this study. Hence, only the respective reaction energies $\Delta E'_{diss}$ were adjusted in element-wise fits for which the respective mean absolute deviation (MAD), root-mean-square deviation (RMSD) and standard deviation (SD) are shown in Tab. 7.1.

The table clearly shows the large errors of uncorrected GFN2–xTB/ALPB(H_2O) compared to DFT. The $\Delta E'_{diss}$ values at the SQM level are underestimated by roughly a factor of two, which translates into errors of about 150 kcal mol^{-1}(*cf.* Tab. A6.2 in Appendix A6). This systematic deviation is attributed to a wrong energy description of the ionic species in the dissociation, as result of the not energy-parameterized TB-Hamiltonian as well as not self-interaction error free second- and third-order electrostatic terms.[39] Similar errors have been observed in the past and can also be seen, *e.g.*, for some subsets of the well-known GMTKN55 benchmark set[109] as well

Table 7.1.: Error measures for the fit of GFN2–xTB/ALPB(H_2O) $\Delta E'_{diss}$ energies and respective corrected energies in comparison with r^{2}SCAN-3c/COSMO-RS//GFN2–xTB/ALPB (denoted as $\Delta\Delta E_{DFT/GFN2}$). Statistics for the C, N, O set are shown separately because they make up for roughly half of the benchmark set and are most important for the studied pK_a values. All values in E_h and kcal mol^{-1}.

		$\Delta\Delta E_{DFT/GFN2}$		$\Delta\Delta E_{DFT/GFN2} + E^{\mathrm{TB}}_{mod}$	
		E_h	kcal mol^{-1}	E_h	kcal mol^{-1}
C,N,O	MAD	0.242	152.00	0.006	4.07
	RMSD	0.243	152.47	0.009	5.34
	SD	0.019	12.02	0.008	5.27
all	MAD	0.242	151.84	0.008	5.00
	RMSD	0.242	152.37	0.012	7.39
	SD	0.020	12.69	0.012	7.24

as in electrochemical applications[185]. In comparison, differences between the implicit solvation models ALPB and COSMO-RS are much smaller.[22] The developed dissociation energy correction is able to reduce the error by the earlier-mentioned order of magnitude and yield MADs of only 4.1 and 5.0 kcal mol^{-1} for the CNO and full fit set, respectively.

Methods providing accurate dissociation free energies should enable reasonable pK_a calculations using a similar set of FER parameters. Therefore, in this study an initial comparison between DFT and GFN2–xTB is given for pK_a values obtained from the same reference FER parameters determined at the r^{2}SCAN-3c/COSMO-RS level. As already mentioned, the motivation for using a higher-order free energy relationship instead of the well-known LFER is a better flexibility of the fit with respect to the same input quantity ($\Delta G'_{diss}$). Different orders of FER (linear LFER, quadratic QFER, and cubic CFER) were tested for the PKA74 set and are shown in Tab. 7.2. The shown values demonstrate the expected trend of decreasing errors with increasing polynomial order. Naturally, a larger number of empirical parameters provides a more detailed adjustment of the fit, although over-fitting has to be avoided. In the discussed FER case a data point–to–parameter ratio of ten (or more) to one is expected to yield reasonable fits. To estimate the best FER order without over-fitting, statistical measures such as the Bayesian information criterion (BIC) are used, which has to be minimized (see Appendix A6). Here, this is the case for the CFER, fourth- and fifth-order polynomial FER, but the CFER was chosen to avoid extremely large FER prefactors c_0 to c_3. In fact, cross-checking on the TR224 and drug benchmarks sets revealed typical over-fitting problems for polynomial FERs larger than third-order, despite the seemingly better statistical performance for the PKA74 set.

The statistical deviations of the calculated pK_a values (MAD) at the DFT level seem to be higher than for comparable studies in the literature. This is attributed to the considered large range of pK_a values over more than 70 units (from -24 to 50) while the vast majority of QSAR training data sets include the more typical range from -5 to 15 units.[555] For the small/large value regions of the pK_a scale there are often no well defined functional groups (which need to

Table 7.2.: Free energy relationship (FER) parameters obtained from polynomial regression of $\frac{\Delta G'_{diss}}{RT \ln 10}$ and experimental pK_a values for the PKA74 set. The reference level r^{2}SCAN-3c/COSMO-RS is abbreviated as r^{2}SCAN-3c and GFN2–xTB+E_{mod}^{TB}/ALPB is abbreviated as GFN2–xTB. Statistical error measures are given in pK_a units. CFER* denotes a fit to GFN2–xTB/ALPB free energies instead of DFT ones.

		FER parameters				statistics [pK_a]	
	method	c_0	c_1	c_2	c_3	MAD	RMSD
LFER	r^{2}SCAN-3c	-177.715641	0.930626	—	—	2.06	2.67
	GFN2–xTB					3.04	4.05
QFER	r^{2}SCAN-3c	-68.715113	-0.134833	0.002590	—	1.94	2.58
	GFN2–xTB					2.89	3.96
CFER	r^{2}SCAN-3c	-1511.889979	21.110068	-0.101200	0.000168	1.86	2.42
	GFN2–xTB					2.84	3.86
CFER*	r^{2}SCAN-3c	-1855.025277	26.075982	-0.124964	0.000206	1.96	2.64
	GFN2–xTB					2.79	3.68

be identified by QSAR schemes) involved in the acid dissociation reaction, but QM free energies can easily be calculated and FER based methods offer an intrinsic advantage. However, extreme pK_a values may lead to significant errors even in QM methods because of the more complicated electronic structures involved, often associated with unusual solvation effects. A parity plot for pK_a calculated with the CFER of Tab. 7.2 is shown in Fig. 7.3.

For the r^{2}SCAN-3c/COSMO-RS reference level a good correlation of $R^2 = 0.96$ and a MAD of 1.86 pK_a units is obtained while GFN2–xTB/ALPB with the same CFER parameters yields a slightly lower correlation of $R^2 = 0.92$ and higher MAD of 2.84 pK_a units. As previously noted, the same FER fit should ideally work with any theoretical method able to compute dissociation free energies. This was tested for GFN2–xTB/ALPB by a separate re-fit (CFER* in Tab. 7.2, *i.e.*, CFER parameters were adjusted by fitting on GFN2–xTB/ALPB energies) which results in only minor changes (of about 0.1 pK_a units) of the statistical measures, although the FER parameters are noticeably different. Hence, remaining errors are attributed to methodical shortcomings of the different levels of theory, *i.e.*, inaccurate dissociation free energies, instead to the quality of the FER fit. Seemingly, the same FER fit may therefore be used in conjunction with different levels of DFT or SQM as long as the absolute free dissociation energies are on the same level. Because the performance for the r^{2}SCAN-3c reference is slightly better, in the following sections the CFER fit shown in Tab. 7.2 is taken as a standard. Whether this is an adequate choice was investigated in comparison with other DFT methods.

7.4.2. Method and Reference Data Dependence

To put results from the previous section into perspective, additional CFER fits for other DFT methods were determined on the PKA74 set. Additionally, the influence of the selection of reference data is investigated in this section. For this purpose CFER fits were performed on the TR224 set introduced by Thapa and Raghavachari.[577] The TR224 set contains more reference

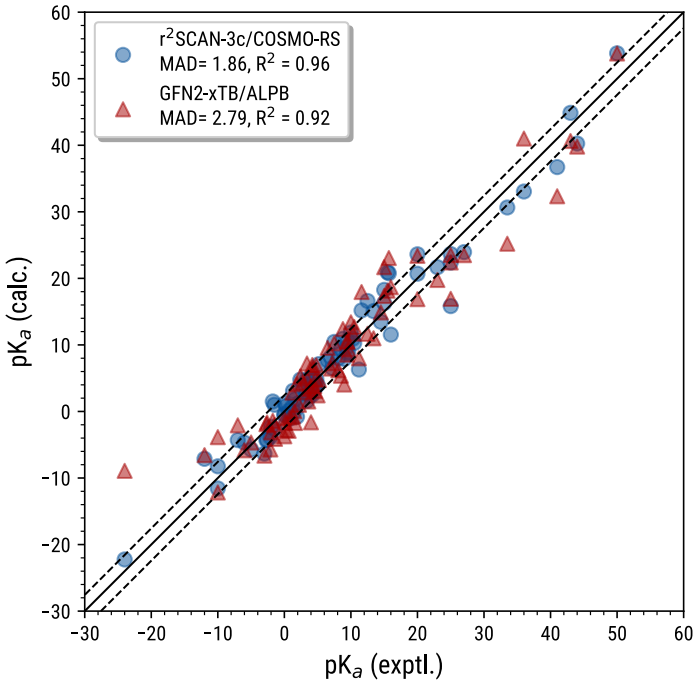

Figure 7.3.: Parity plot for the r^{2}SCAN-3c/COSMO-RS pK_a value (blue dots) in comparison with experimental values. GFN2–xTB+E_{mod}^{TB}/ALPB data (red triangles) were calculated using the same regression CFER. The dashed lines correspond to a SD of 2.43 pK_a units obtained for r^{2}SCAN-3c.

data points than the PKA74 set but only represents a smaller range of about 20 pK_a units, similar to typical QSAR training data. With two sets of CFER parameters for two fit sets and one "reference" CFER determined at the r^{2}SCAN-3c level (Tab. 7.2), the performance for each level of theory can be cross validated. Statistical data is given in Tab. 7.3 and the respective CFER parameters can be found in Appendix A6.

Interesting observations can be made from the presented data. Despite the assumption in the previous section, no universal or "default" CFER should be applied for all levels of theory. The respective MADs and RMSDs are strongly increased in comparison to CFER fits that were determined for each individual method. Obviously, CFER fits show the best performance for each method on the benchmark set for which they were determined (*i.e.*, PKA74 or TR224). This is expected because the polynomial regression minimizes the errors. However, while CFER parameters determined on PKA74 seem to be applicable also to pK_a predictions of the TR224 set and produce errors of similar magnitude, the *vice versa* approach shows an opposing trend. In fact, CFER parameters determined for the TR224 set only yielded very large RMSDs of up to 10.8 pK_a units at DFT level when applied to the PKA74 set. On the other hand, within the smaller pK_a range of TR224, the respectively fitted parameters showed superior performance and produced MADs within the target 1 pK_a unit range.

A probable interpretation is that the large pK_a range of PKA74 can be seen as a CFER

Table 7.3.: Statistical values for pK_a calculated at various levels of theory. CFER parameters were either determined for the PKA74 or TR224 set and evaluated with the respective other set. The "default" CFER refers to r^{2}SCAN-3c parameters from Tab. 7.2. All DFT calculations employ COSMO-RS(H_2O) implicit solvation. TZ is an abbreviation for the def2-TZVPP basis set. GFN2–xTB includes the energy correction E_{mod}^{TB} and employs ALPB(H_2O) implicit solvation. All values are provided in pK_a units.

	"default" CFER		CFER adjusted on PKA74		CFER adjusted on TR224	
method	MAD	RMSD	MAD	RMSD	MAD	RMSD
			PKA74			
GFN2–xTB	2.84	3.86	2.79	3.68	5.54	11.59
r^{2}SCAN-3c	—	—	1.86	2.42	4.84	10.10
B97-3c	2.58	3.23	1.81	2.30	3.64	6.69
B97-D/TZ	3.48	3.96	1.70	2.28	5.10	10.79
PBE0-D3/TZ	1.93	2.54	1.90	2.50	5.28	10.55
PW6B95-D3/TZ	2.47	3.10	1.67	2.26	4.80	10.03
ωB97X-V/TZ	2.52	3.14	1.64	2.30	2.32	3.91
			TR224			
GFN2–xTB	1.98	2.47	1.75	2.16	1.38	1.81
r^{2}SCAN-3c	—	—	1.43	1.89	0.93	1.22
B97-3c	2.67	2.97	1.19	1.52	0.70	1.04
B97-D/TZ	3.77	4.04	1.22	1.65	0.79	1.03
PBE0-D3/TZ	2.61	3.10	2.43	2.89	0.97	1.22
PW6B95-D3/TZ	2.64	3.14	1.29	1.66	0.93	1.18
ωB97X-V/TZ	2.47	3.02	1.11	1.50	0.87	1.12

parameter constraint for very complicated systems that still model a proper dependence on the dissociation free energies. In this case, fits on TR224 would be "fine-tuning" for the smaller pK_a range, but could fail for the extreme values. Note that this is reflected in the empirical factors c_0 to c_3, which differ significantly (in magnitude and pre-sign) between the two fits for all methods, with exception for the range-separated hybrid ωB97X-V. Since ωB97X-V/def2-TZVPP is the highest level method tested within this study, the reason for its good performance are very well behaved and practically charge self-interaction free dissociation free energies, especially for challenging anionic systems. However, at least for the limited number of methods employed in this study, there is no significant correlation between the DFT rung and the quality of pK_a calculations. Rather independently of the employed functional, all DFT methods yielded qualitatively and quantitatively similar results. All three employed hybrid functionals involved significantly higher computational cost than the r^{2}SCAN-3c method (factor of 50 and more) but in comparison showed only minor improvements for the computed pK_a values (< 0.2 pK_a units). Surprisingly, the otherwise very robust and commonly applied PBE0 functional even shows the worst performance of all evaluated DFT methods. In view of the higher computing efficiency, it makes thus sense to employ a cheaper yet accurate "all-purpose" composite functional such

as r^{2}SCAN-3c. Interestingly at about the same computational effort, the B97-D/def2-TZVPP GGA functional and its composite counterpart B97-3c outperformed all other methods, at least for the TR224 set. However, this good performance must be attributed mainly to beneficial compensation of the self-interaction error (SIE) that is present in GGA functionals, and this will be discussed briefly in the next subsection. r^{2}SCAN-3c suffers slightly less from SIE and should provide a more physical description of dissociation free energies.[182] Furthermore, outstanding performance of r^{2}SCAN-3c was observed for conformational energies, which is relevant for the treatment of ensembles of non-rigid molecules (see Sec. 7.4.4).

GFN2–xTB as the only SQM method in the evaluation provides the most cost efficient calculations of all discussed theory levels. Nonetheless, predictions of pK_a values at GFN2–xTB+E_{mod}^{TB}/ALPB level are only slightly worse than the DFT results for both the TR224 and PKA74 sets. Considering that the method was not designed for thermochemistry and a comparatively simple ALPB implicit solvation model is used, this is encouraging. Trends of pK_a values will likely be qualitatively correct at the GFN2–xTB level and hence may provide the opportunity for fast pre-screening applications. Note, however, that the E_{mod}^{TB} term is required here. SQM might also be used for supportive thermostatistical calculations (G_{trv}^{T}, as has been applied here), to provide chemoinformatic descriptors[557,566], initial conformational sampling[33], or calculation of the conformational entropy.[186]

All discussed results clearly demonstrated a good correlation between dissociation free energies and experimental pK_a values. The main conclusion here is that, given a method specific CFER fit, any of the employed methods may be used for pK_a calculations with sufficient accuracy, indicating the physical plausibility of the presented approach.

7.4.3. Functional Group pKa Values

To further decompose the errors, the functional group dependent performance is discussed. MAD values with respect to experimental data for the 12 chemical groups included in the TR224 set at the GFN2–xTB+E_{mod}^{TB} and r^{2}SCAN-3c level are shown as a radial plot in Fig. 7.4. Because the investigated molecules were mostly rigid, only a small influence of the molecular conformation was expected and single structures for acid/base were considered. Non-rigid molecules are discussed in the following section.

From the radial chart the largest errors at the r^{2}SCAN-3c/COSMO-RS reference level can be seen for alcohols and thiols. This is qualitatively consistent with results for other DFT or QM methods observed in Ref. 577. Errors at the GFN2–xTB/ALPB level were larger on average and in particular for the R-C-H, R2-NH, R-NH$_2$, and R/Ph-COOH subsets. However, the overall maximum deviations of GFN2–xTB for TR224 are not larger than those of DFT. The CFER parameters determined from PKA74 seemingly worked quite well in these cases. A minimization of errors by polynomial regression of the TR224 set reveals slightly mismatching slopes and shifts of the pK_a values compared to the fit set. This is evident form Fig. 7.4, as large errors for some functional groups were minimized (mainly R-OH), while others (R-C-H) are increased by the set-specific regression (*e.g.*, compare the red area and black outline for r^{2}SCAN-3c/COSMO-

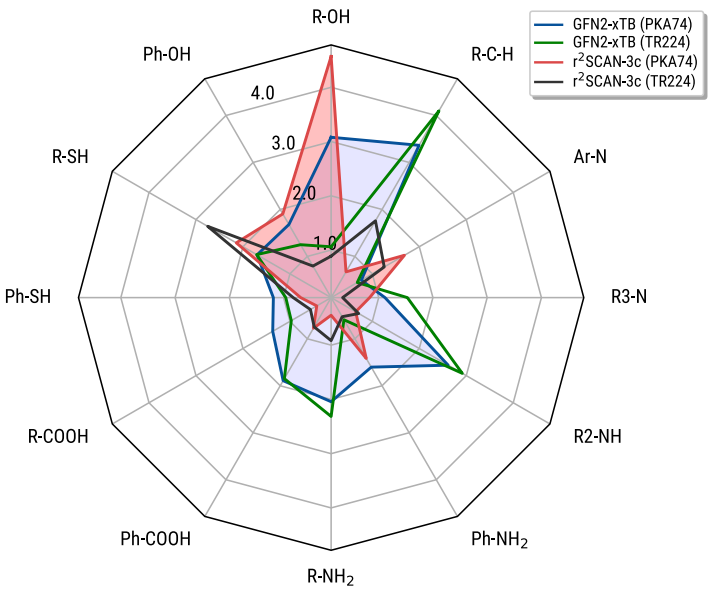

Figure 7.4.: Radial plot for MADs of GFN2–xTB+E_{mod}^{TB}/ALPB and r^{2}SCAN-3c/COSMO-RS pK_a values for different chemical groups included in the TR224 set. The benchmark set on which CFER parameters have been determined is given in parentheses behind the levels of theory.

RS). The differences were mainly caused by the functional groups with very high and very low acid dissociation constants, which can be seen from the color-coded parity plot of experimental and calculated (CFER from PKA74, r^{2}SCAN-3c/COSMO-RS) pK_a values in Fig. 7.5a.

Here, pK_a values for R-OH, Ph-OH, anilines, and R-NH$_2$ evidently require a much steeper slope, while other functional groups within the 5–10 pK_a range are in relatively good agreement with the experiment. As can be seen from Fig. 7.5b, calculated pK_a for phenols and alcohols agree especially well with the experiment at the B97-3c level. As mentioned earlier, this is presumably an effect of error compensation. One possible explanation is that the SIE will lower the energy of the anionic bases relative to the energy of the conjugate acids for these molecules and hence will less strongly overestimate the pK_a values, *e.g.*, compared to the r^{2}SCAN-3c level. High-level ωB97X-V calculations do not profit from similar error compensation, but expectedly should yield the most consistent dissociation energies of all tested methods. Hence, its performance is in between r^{2}SCAN-3c and B97-3c. Values for thiols seem to be systematically underestimated (*i.e.*, shifted) at all levels of theory and are seemingly only very little affected by the CFER re-fit (*cf.* Fig. 7.4). Thiols furthermore comprise the only outlier in the set. The respective molecule, 3-mercaptopropanoic acid, denotes a thiol pK_a according to the reference literature.[576,577] Here, carboxylic acid dissociates first, but neither the corresponding pK_a nor the one calculated for the thiol, nor the double deprotonation yields a value corresponding to the reference. The reason for this mismatch is not clear, but note that another reference value of 4.34 can be found in the literature[587] that deviates by only 0.1 pK_a units from our calculated value using the TR224 CFER parameters. Overall, all trends and observations from

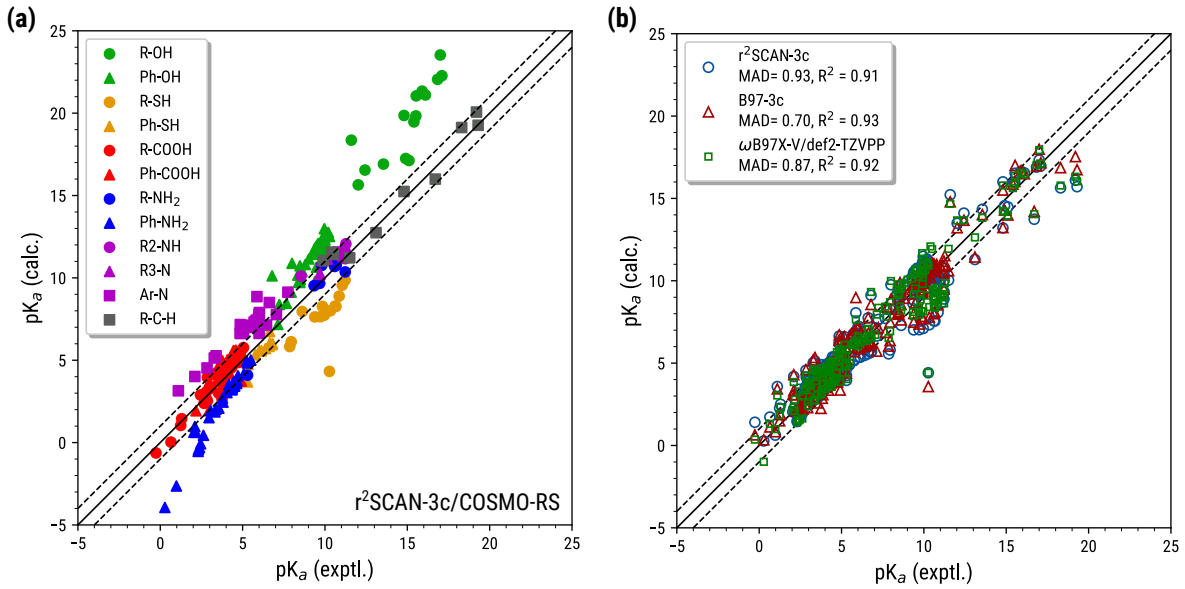

Figure 7.5.: Parity plots for pK_a values calculated at DFT level. **(a)** Values calculated at the r^{2}SCAN-3c/COSMO-RS level of theory using the CFER parameters from PKA74, colored by functional group. **(b)** Comparison of r^{2}SCAN-3c/COSMO-RS pK_a values with B97-3c/COSMO-RS and ωB97X-V/def2-TZVPP/COSMO-RS, all with CFER data fitted on TR224. The dashed lines correspond to a target error range of 1 pK_a unit.

the literature[577] could be reproduced, and all methods tested here provide the same qualitative results. Mismatching slopes can be easily repaired by the CFER re-fit, which is visible from the correlation plots in Fig. 7.5b. The few remaining errors, *e.g.*, the earlier-mentioned shift for thiols, seem to be systematic and depend on the theoretical description provided by the DFT method, rather than on the CFER quality. Thapa and Raghavachari have shown that remaining errors for this set can efficiently be reduced by including 1–3 explicit water molecules into the calculation. In this way, low MADs of about 0.45 pK_a units at the CBS-CQB3 level of theory were obtained.[577] Similar observations have been made in Refs. 574–576. In the present study we have limited ourselves to the use of implicit solvation models and (semi-)automated workflows for the calculation of $\Delta G'_{diss}$, but the automatic inclusion of solvent molecules as an extension of the presented method is currently under investigation in our lab.

7.4.4. Flexible Drug Molecules

After having discussed different levels of theory and CFER fits in the previous sections, the protocol is tested for pK_a calculations of larger, non-rigid molecules. The benchmark sets compiled by Jensen *et al.*[560] and the SAMPL6 set[45,46] will be used. Because experimental pK_a values for these molecules more closely resemble a range similar to the TR224 set, the respective CFER parameters were employed. Furthermore, dissociation free energies were calculated from Boltzmann averaged dissociation free energies for the acid and base to include conformational

effects. A more detailed investigation of conformational effects will be given in the following section.

As in previous sections, the GFN2–xTB/ALPB and r^{2}SCAN-3c/COSMO-RS methods are tested here, as well as B97-3c/COSMO-RS because it was the best performing method in Sec. 7.4.3. The corresponding error measures are shown in Tab. 7.4 in comparison with data from the literature. For all FER methods discussed in this work, good agreement with exper-

Table 7.4.: Error measures (MAD, RMSD, in pK_a units) and dimensionless R^2 for r^{2}SCAN-3c/COSMO-RS, GFN2–xTB+E_{mod}^{TB}/ALPB and B97-3c/COSMO-RS in comparison with other methods from the literature. The CFER parameters were adjusted on the TR224 set.

	this work, CFER			literature		
	GFN2–xTB	r^{2}SCAN-3c	B97-3c	other DFT	Epik[e]	ACDlabs[f]
drug set (Jensen *et al.*)						
MAD	1.18	0.71	0.63	0.53[a,b]	0.43[a]	0.39[a]
RMSD	1.43	0.84	0.86	0.69[a,b]	0.72[a]	0.64[a]
R^2	0.71	0.88	0.88	0.92[a,b]	0.91[a]	0.93[a]
SAMPL6						
MAD	1.74	0.77	0.85	0.58[c]	0.78[d]	0.55[d]
RMSD	2.68	0.89	0.97	0.68[c]	0.95[d]	0.77[d]
R^2	0.14	0.89	0.87	0.94[c]	0.88[d]	0.92[d]

[a] Ref. 560 [b] Ref. 543, COSMOtherm [c] Ref. 221, DSD-BLYP-D3/def2-TZVPD/COSMO-RS
[d] Refs. 45,46 [e] Ref. 244 [f] Ref. 578

imental data is observed. Both r^{2}SCAN-3c and B97-3c show MADs, and in case of the drug benchmark set also RMSDs, below the target 1 pK_a error. Note that errors would be larger than this target when employing PKA74 CFER parameters, which are not shown in the following. Only one major outlier was seen for the drug set, which overestimated the pK_a by about 6 units. For the respective molecule, cimetidine, a proton transfer is observed in the conjugate acid upon geometry optimization at the r^{2}SCAN-3c level, hence explaining the mismatch. This value was excluded from the statistics. Note, that if only single-point energies at the r^{2}SCAN-3c level are calculated on GFN2–xTB geometries and the proton transfer is avoided, the computed pK_a is shifted by 3 units in direction of the experiment. Also at other levels of theory, rare outliers appear, *e.g.*, at the B97-3c/COSMO-RS level no pK_a could be calculated for cefadroxil due to SCF convergence problems for the zwitterionic conjugate base. GFN2–xTB provides excellent results also close to the target 1 pK_a error range for the drug set while for SAMPL6 only results of mixed quality are obtained. Out of the 31 reference values in the set, 15 showed a deviation much smaller than 1 pK_a units at GFN2–xTB+E_{mod}^{TB} level, but 7 systems (all with an anionic conjugate base) produced errors larger than 3 pK_a units. To get qualitatively correct results for these systems DFT calculations seem to be required, but due to the considerably lower computational costs compared to DFT, SQM results are nonetheless useful.

In comparison with methods from the literature the CFER results for all three methods were

slightly worse. As already observed by Jensen *et al.*, chemoinformatic tools such as Epik[244] or ACDlabs/pka[578] provide exceptionally good results at a fraction of the computational cost of FER based approaches for systems for which these tools were developed. This can be seen again in Tab. 7.4, where ACDlabs/pka and Epik outperform all QM and SQM results. However, for the drug set it is very likely that all experimental pK_a values were included within the training parameters of the chemoinformatic tools, hence explaining the good predictions.[560] For molecules presumably *not* included in the training data, *e.g.*, the SAMPL6 set, larger errors are seen and DFT results from the literature[221] are slightly better than the still surprisingly well-performing Epik and ACDlabs/pka predictions. The slightly worse results of the FER approaches are mainly attributed to the CFER parameters fitted for TR224, which are not the optimum for the drug and SAMPL6 set. In fact, if CFER parameters are re-adjusted for these sets (see Tab. A6.20 in Appendix A6), RMSDs for the drug molecules decrease to 0.74 and 0.69 pK_a units for the r^{2}SCAN-3c and B97-3c methods, respectively. For the SAMPL6 set, RMSDs decrease to 0.67 and 0.55 pK_a units, which even outperforms the "winning" submission of the SAMPL6 blind challenge, calculated with the DSD-BLYP-D3(BJ) double hybrid functional.[148,221] The respective CFER parameters are found in Appendix A6 and could easily be used in a problem specific manner (*i.e.*, for other drug-like molecules). For general purpose pK_a predictions the TR224 and PKA74 parameters yield sufficient accuracy. Note again that the presented CFER protocol only involves free energy computations for the acid and conjugate base, and does not require manual selection of tailored reference systems as, *e.g.*, suggested in Ref. 560. Hence, its simplicity and efficiency makes the protocol a good choice in screening applications.

It is unlikely that errors much below 0.5 pK_a units can be achieved with either QSAR/ML or FER methods since this is already the realistic regime of errors for the reference data from titration and NMR experiments.[588] For any of the FER based methods the quality of free energies is limited by the accuracy of the δG^T_{solv} term, which is partially compensated by the CFER fits. As mentioned earlier, one possible improvement for this is the combination with microsolvation models or the inclusion of single solvent molecules.[577] In other words, while QSAR models are already limited by the available training data, systematic improvements are possible for all FER methods.

7.4.5. Conformational Effects

Any theoretical approach that computes the pK_a value explicitly from the free energy of molecular structures may show a dependence on the conformation for non-rigid cases. Input geometries, *e.g.*, generated from SMILES strings, have a more or less random conformation that is typically several kcal mol^{-1} higher in energy than the respective global minimum, and this can occur randomly for the acid/base pair in pK_a calculations. The inclusion of conformational ensembles significantly increases the computational cost but is often necessary to achieve sufficient accuracy and represents the physically correct approach. To quantify this influence, the spread of individual microstate pK_a values, *i.e.*, values calculated for single conformers in the acid and base ensembles, were evaluated for the drug and SAMPL6 set. Here, the differences between

the experimental pK_a and the Boltzmann averaged, minimum and maximum possible micro pK_a were obtained from the respective averaged, minimum and maximum $\Delta G'_{diss}$ as shown in Fig. 7.1. For both benchmark sets the respective ΔpK_a at the r^{2}SCAN-3c/COSMO-RS level (using CFER parameters fitted on TR224) are shown as boxplots in Fig. 7.6. The significance of

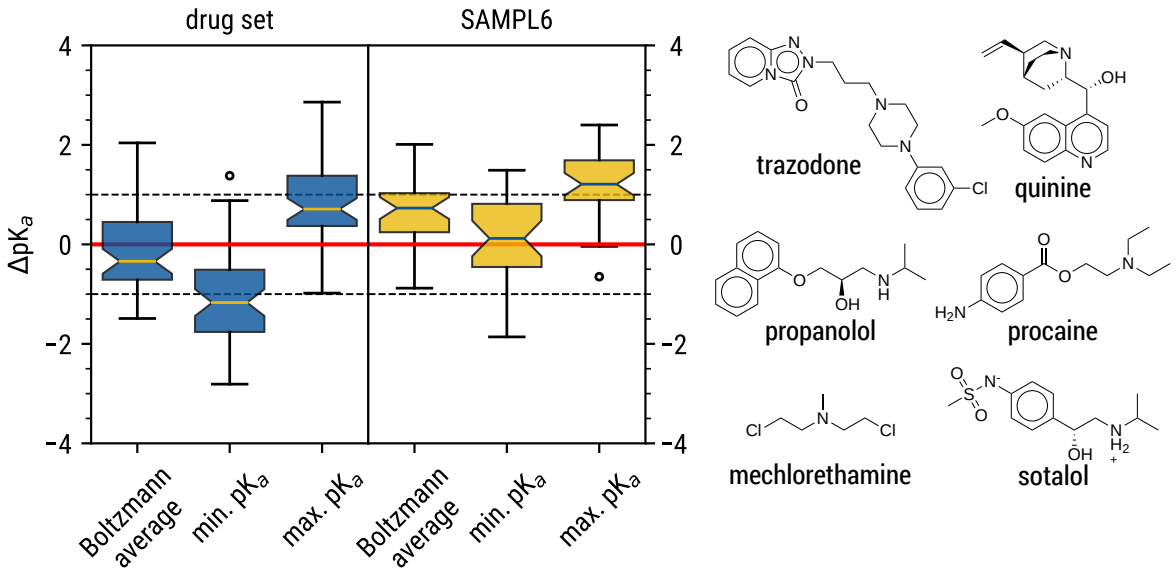

Figure 7.6.: Deviations (ΔpK_a) for the drug and SAMPL6 set at the r^{2}SCAN-3c/COSMO-RS level, visualized as box plots. Data is shown for the conformational (Boltzmann) average, as well as the minimum and maximum possible pK_a from the ensembles. The dashed lines denote a target error range of ± 1 pK_a units. On the right side the six molecules with the largest min./max. pK_a spread are shown in their neutral protonation state. The circles in the boxplots are outliers not included within 95% confidence interval around the median.

the conformational treatment is obvious from the plot. For the flexible drug molecules, the pK_a is either strongly over- or underestimated depending on the choice of conformers, and only the Boltzmann averaged value gives results close to the experiment. Herein, flexible molecules, such as the ones shown on the right side in Fig. 7.6, show differences of as much as three pK_a units between the minimum and maximum possible microstate pK_a. For the SAMPL6 sets, pK_a values seem to be overestimated in general, and taking the minimum pK_a gives the smallest mean deviation to the experiment. However, the SAMPL6 set contains several rigid molecules and furthermore consists of many cases with anionic conjugate bases. Hence, in accordance with previous the sections, an overestimation of pK_a values is expected here. In fact, at the r^{2}SCAN-3c/COSMO-RS level MADs for the drug set could increase from 0.71 pK_a units to as much as 1.23 pK_a units by neglect or wrong choice of conformations. In case of the SAMPL6 set, MADs could increase from 0.77 to 1.27 pK_a units. The corresponding MADs are visualized in Fig. 7.7 in comparison with GFN2–xTB+E^{TB}_{mod}/ALPB and B97-3c/COSMO-RS, where MADs resulting from a wrong conformer selection (either min/max $\Delta G'_{diss}$) are shown as light-colored bars. All

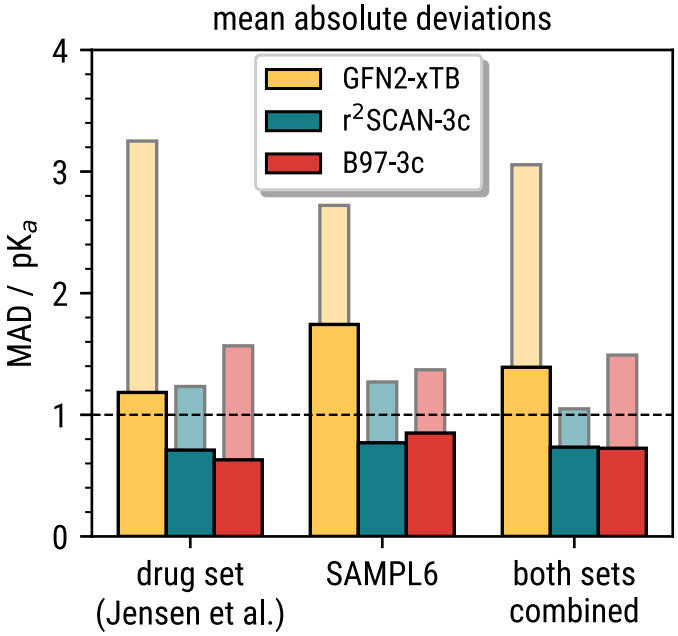

Figure 7.7.: pK_a MADs for GFN2–xTB+E_{mod}^{TB}/ALPB, r^{2}SCAN-3c/COSMO-RS and B97-3c/COSMO-RS. The third bin combines the drug and SAMPL6 set. The light-colored thinner bars show the upper limit MADs resulting from non-conformationally averaged dissociation free energies. The dashed line denotes a target error range of 1 pK_a unit.

three methods clearly owe their good performance to the conformational treatment and, in case of the two DFT levels, would exceed a MAD of 1 pK_a unit otherwise. Notice that the minimum and maximum $\Delta G'_{diss}$ values refer to conformational ensembles from the `censo` program's default sorting procedure, *i.e.*, only structures up to 2.5 kcal mol^{-1} relative to the lowest structure are included. The spread between minimum and maximum pK_a could be even larger in reality because relative conformational energies from a (semi-)random conformer selection can easily exceed 2.5 kcal mol^{-1} by up to an order of magnitude for large, flexible molecules. At the GFN2–xTB level, conformational ensembles up to a relative energy of 6 kcal mol^{-1} were included, which explains the comparably large MAD differences in Fig. 7.7. As an alternative to the Boltzmann averages one could also simply use the free energy of the global minima for the acid and base molecules. Compared to the averaged value $\overline{G} = \sum p_i G_i$, the global minimum free energy G_{min} will be slightly lower *for both species* and hence yield a pK_a close to the population-averaged value. But since free energies are required for the correct ranking and identification of the global minima anyways, calculating the Boltzmann average involves no additional computational cost.

7.5. Conclusion

The prediction of pK_a values is an important field of research with many different approaches available from chemoinformatics and computational chemistry. In the presented study, we have

applied the latter in combination with molecular free energy computations and a new CFER equation to compute acid dissociation constants. Four benchmark sets have been used for the evaluation of the CFER approach with a total of 390 experimental reference pK_a values. Herein, several theoretical methods have been tested, with a focus on the recently published r^{2}SCAN-3c/COSMO-RS(H_2O) DFT composite functional and the semiempirical GFN2–xTB/ALPB(H_2O) method.

In some initial tests it was shown that dissociation free energies at the GFN2–xTB level are underestimated approximately by a factor of two compared to DFT. Upon application of a new element-dependent correction term, heterolytic dissociation free energies were improved and could be used for the calculation of pK_a values. Using different orders of FER for both GFN2–xTB and r^{2}SCAN-3c, accurate pK_a values were obtained for a test set of 82 molecules spanning a large pK_a range of 74 units (PKA74 set). On the basis of the respective results, a new cubic FER (CFER) was chosen as final working equation which requires just two ensembles (*i.e.*, the acid and base) for the calculation of the dissociation free energy. More detailed insights into method performance were provided by fits and cross-evaluations on the PKA74 and TR224 sets with several levels of DFT. This revealed a strong functional group dependence of the error for the pK_a but only comparatively small performance differences between different levels of theory, with RMSDs between 1.03 and 1.22 pK_a units for DFT and 1.81 for GFN2–xTB. Surprisingly, it was observed that even cheap composite functionals such as B97-3c can provide higher accuracy than high level ωB97X-V/def2-TZVPP calculations, mainly due to error compensations. pK_a calculations for larger, non-rigid molecules were investigated at r^{2}SCAN-3c, GFN2–xTB, and B97-3c levels using two benchmark sets from the literature and CFER parameters determined for the TR224 set. Excellent correspondence to experimental pK_a was achieved for the DFT methods with RMSDs below 1.0 pK_a units, despite showing strong conformational dependencies. For the latter it was shown that the neglect of conformational effects can strongly influence the errors, increasing MADs by more than 0.5 p$K_a$units.

Some concise conclusions can be drawn from the presented data. In general, FER based approaches at the DFT level can perform similarly to highly specialized chemoinformatic pK_a predictors, although strongly exceeding them with regards to computational costs. Calculations at the GFN2–xTB/ALPB level are less expensive and accuracy-wise seem to be suitable at a semi-quantitative level and for trend recognition. Furthermore, FER approaches are attractive because they do not rely on proprietary software and allow more detailed insights, *e.g.*, into conformational dependencies or even uncommon chemical systems with highly positive or negative pK_a values. However, for optimum performance, method-specific FER parameters should be determined.

Future work on this topic will focus on the automatic inclusion of explicit solvent molecule (microsolvation) and investigations of conformational entropy effects. The latter could be important for molecules with very different ensembles for the conjugate acid and base.

Acknowledgements

This work was supported by the DFG in the framework of the "Gottfried Wilhelm Leibniz-Prize" awarded to S.G.

Supporting Information

Some additional supporting information can be found in Appendix A6. Detailed dissociation free energies and FER parameters are provided. Input structures can be obtained from the electronic supporting information under DOI 10.1021/acs.jpca.1c03463.

Part V.

Final Summary and Conclusion

In this thesis the capabilities of fast quantum chemical methods for the systematic exploration of the low-energy chemical space were investigated. The main focus was the application of the so-called GFN*n*–xTB schemes, semiempirical quantum mechanical (SQM) methods belonging to the (extended) tight-binding type. These methods are characterized by an element-wise parametrization of almost the entire periodic table ($Z \leq 86$) and a special-purpose construction for the good description of molecular geometries, frequencies, and non-covalent interactions. In combination with low computation times and relatively high robustness, large scale screening procedures, *i.e.*, the exploration of the low-energy chemical space, are the archetypal field of usage for GFN*n*–xTB SQM. With regards to the chemical space, the screening of molecular conformations is of special importance and requires a fine balance between non-covalent and strongly directional short-range covalent interactions. This led to the development of a broadly applicable and efficient program called CREST based on the GFN*n*–xTB methods as the central part of this thesis.

Knowledge of the low-energy chemical (conformational) space is imperative for the calculation of molecular properties such as spectroscopic features or free energy based equilibrium rate constants. The link between QM calculations for single structures and macroscopic observables is provided by statistical mechanics *via* Boltzmann population averages. Different members of the conformational space are discretized by their energy as a function of the nuclear spatial coordinates, which defines the potential energy surface (PES). It is the task of programs such as CREST to systematically explore the PES and find relevant stationary points (minima) defining the low-energy chemical space. The approach for finding molecular conformers pursued in CREST is based on a combination of specialized metadynamics (MTD) simulations and efficient geometry optimizations at a highly robust SQM level. A key component of the respective algorithm is an efficient sorting procedure based on energetic differences and purely structural descriptors such as the atomic RMSDs and differences between the rotational constants B_e. Other representatives of the low-energy chemical space, such as protomers, tautomers, and non-covalent aggregates can also be sampled with CREST due to the computational robustness of GFN*n*–xTB. Here, the quantum mechanical origin of GFN*n*–xTB provides the necessary methodological flexibility for these applications, *e.g.*, by the ability to freely form and break bonds. In reviewing all these procedures, Part II was dedicated to CREST and its prototypical applications. The performance of conformational searches was demonstrated for small organic compounds, a large polypeptide with 220 atoms, several macrocyclic and two organometallic molecules. While the comparison of conformers to experimental data is often difficult, in all of these examples diverse ensembles were obtained providing an excellent starting point for higher level QM post-processing.[22,589] Furthermore, it was shown that the general workflow in CREST can be employed to special problems such as conformational sampling with structural constraints applied to parts of the system as, *e.g.*, successfully demonstrated for tyrosine on a fixed graphene cut-out and conformers of a S_N2 methyl-group transfer transition state. With the latter, it was shown that due to the so-called Curtin–Hammett principle[219] knowledge of the conformational

space is especially important for mechanistic studies. Non-covalent aggregates were sampled with a conformational workflow employing similar constraints and the system was encapsulated within a spherical potential in order to avoid dissociation. Finally, protonation, cationization, and tautomerization schemes were discussed for some organic and inorganic compounds, extending the capabilities of CREST for applications to low-energy chemical space sampling. In conclusion, CREST is a powerful tool for simulations of molecules up to a few hundred atoms and will potentially find wide spread application in computational chemistry. Since first being published, the program has already been applied for a diverse number of projects, *e.g.*, the large scale conformer generation of organometallic compounds,[183,590] protein side chain conformational sampling,[591] gas docking in metal-organic frameworks (MOF),[444] input generation of machine-learning approaches,[414] and various other mechanistic,[592–594] conceptual,[595–598] and spectroscopic studies.[22,476,599–601]

In Part III, a recent extension to the CREST code was presented and discussed linking the conformational low-energy chemical space to statistical thermodynamics. The central quantity here is the absolute molecular entropy, or more specifically, the conformational contribution to it. A numerically robust and accurate algorithm was introduced for calculation of the latter. This approach is based on the separation of the molecular partition function (Eq. 2.62) into molecular electronic, translational, rotational, vibrational parts and a conformational contribution. Since this partitioning, especially with regards to rotational, vibrational and conformational components, is problematic due to a breakdown of the truncated Taylor expansion in the harmonic approximation, all treatments of the conformational entropy include some degree of uncertainty. Various schemes for the calculation of conformational entropies have been proposed in the past,[43,376] but no generally applicable workflow existed up to this point. The newly implemented algorithm in CREST generates conformational ensembles on an iterative basis with a converged conformational entropy and ensemble size as criteria for termination. As another introduced novelty, the intermediate ensembles were used to extrapolate the calculated entropy, which serves as a convergence enhancement. Herein, the conformational terms are basically treated as additional electronic energy levels for which standard thermodynamic expressions were employed. To treat the problematic separation of rotational, vibrational and conformational degrees of freedom (DOF), frequency calculations at DFT level were conducted in a modified and scaled msRRHO approximation.[42] While being computational less expensive than full anharmonic DFT treatments,[380] the new approach proved to provide exceptional accuracy in comparison with absolute molecular entropies from experimental measurements, in most cases with better than chemical accuracy (about $3\,\mathrm{cal\,mol^{-1}\,K^{-1}}$). Even for complicated non-rigid molecules up to one hundred atoms and hardship cases such as long *n*-alkanes up to $C_{16}H_{44}$ numerically stable and accurate entropies were obtained. Results of similar quality were obtained for absolute molecular heat capacities that are calculated from related thermodynamic equations, including the conformational contribution. The significance of the conformational terms was demonstrated in this part by some prototypical calculations of reaction free energies.

Due to a large change in DOF, for example by ring-closure during the reaction, large conformational free energy contributions up to several kcal mol^{-1} energy difference were observed. The computational efficiency and accuracy of the presented workflow allows its standard application for thermodynamical investigations of non-rigid molecules. As was shown, both GFNn–xTB and GFN–FF can be routinely applied for such calculations, where the latter may also enable applications of the workflow for molecular systems with up to 200 atoms.

Part IV included three computational studies where the fast and reasonably accurate GFNn–xTB methods and composite DFT-3c methods[158,180,387] have been applied for the calculation of gas-phase IR spectra and acid dissociation constants. CREST (and its conceptional predecessor) was used in all three chapters of Part IV to sample molecular conformations. For the calculation of pK_a values, it was furthermore employed to determine (de-)protonation sites. Chapter 5 presented the calculation of gas-phase IR spectra from harmonic vibrational frequencies and derivatives of the molecular dipole moment obtained at the GFN1–xTB, GFN2–xTB, GFN–FF and B3LYP-3c levels. Based on a sufficiently large sample size of more than seven thousand experimental gas-phase reference spectra, it was shown that all these methods can be applied for the IR spectra simulation with a varying degree of accuracy. At DFT level (B3LYP-3c), problems arise mainly from systematically overestimated frequencies, which could be sufficiently repaired by application of linear frequency scaling factors. For SQM calculations at GFNn–xTB level, errors in harmonic frequencies and molecular dipole moments are less systematic due to the more empirical nature of the methods. However, spectra at the SQM level were found to be only slightly worse than DFT ones, even for complicated organometallic molecules. Only at FF level problems arise due to wrong harmonic frequencies and errors of the dipole moments derived from classical EEQ charges. Hence, improvements for the SQM and FF levels are required. For IR frequencies this was achieved by an atomic mass scaling approach which was found to perform better than the often employed linear frequency scaling at all employed levels of theory. Overall, the low-level SQM and FF methods were found to be suitable for larger scale pre-screening of IR spectra in context of unknown compound identification processes,[44] but can not be recommended for final predictions where DFT should be used instead. Moreover, IR spectra simulation for non-rigid molecules requires conformational sampling as it was shown exemplary for a subset of the experimental reference data. Chapters 6 and 7 discussed the calculation of pK_a values from free energy relationships. Several different levels of theory were employed but in general free energies were obtained from Eq. 2.64, where the total energy was computed at DFT or SQM level, thermal contributions were generated at the GFNn–xTB level, and solvation free energies were calculated from implicit solvation models. Computations in this part are straight forward and require only a free energy for an acid and its conjugate base to calculate the dissociation constant. The general finding here was that excellent accuracy can be achieved in comparison with experimental data, often with lower than one pK_a unit error for drug like molecules. This performance was found to be rather independent of the employed level of DFT for the total energies. Also GFN2–xTB in combination with ALPB implicit solvation

and an empirical correction for heterolytic dissociation energies was able to provide reasonable pK_a predictions. However, for flexible molecules, a strong conformational dependence of the pK_a was found which requires thorough sampling. In summary, Part IV presented the application of robust and efficient computational protocols based on GFNn–xTB, CREST, and low-cost DFT for the prediction of molecular properties. Multilevel or "bottom-up" screening procedures (*cf.* Fig 1.3) enable efficient computational studies and are of essential importance for thermochemical investigations in modern computational chemistry.

As mentioned in the introduction, the *conquest of the combinatorial conundrum*, *i.e.*, the need for practical workflows that provide a full understanding of conformational ensembles of (bio-)molecules and associated thermodynamic properties such as the entropy, is one of the "holy grails" in computational chemistry.[3,32] The naturally arising question after this thesis is "Does CREST in combination with GFNn–xTB provide a suitable solution to the combinatorial conundrum?" No definite answer can be given to this. With good confidence one can state that CREST *will* provide an approximate solution to the conformational problem and accurate entropy calculations for reasonably sized molecules with up to about 100 atoms. In aspect of generalizability, due to the broad parametrization and robustness of GFNn–xTB, conformational sampling in CREST provides unprecedented capabilities and computational performance in comparison to similar tools. Similar conclusions can be made for the quality of generated conformers and their relative energies. The concept of conformers is intrinsically related to the PES and conformational energies at SQM (GFNn–xTB) level will for a wide range of systems be more accurate than at FF level, which is traditionally applied for large-scale sampling procedures. Molecular geometries are a target feature of GFNn–xTB and for most systems should yield good quality structures. However, (conformational) energies and geometries obtained at SQM level are worse than at higher (*ab initio*) levels of theory.[589] As extensively discussed in Part II, CREST at GFNn–xTB level is a compromise between computational cost and accuracy and in practice a re-ranking of the conformers at DFT level is often required.[22,589] The general construction of CREST workflows and sorting procedures would in principle allow running all parts at a DFT PES, which would yield the required ensembles directly and diminish the need of a multilevel procedure. However, this is unfeasible due to the enormous computational cost. Already at SQM level capabilities for conformational searches of non-rigid molecules are nearing the limit with approximately 200 atoms. Sampling at DFT level for such systems is practically impossible and computational or theoretical advances are obviously desired. For technological developments it is often reasonable "to expect the un-expected" and, despite current deviations from Moor's law, more powerful computers will most likely exist in the future further extending the capabilities in computational chemistry. Hence, running CREST at a cheap DFT level may be possible within the next decade or two. This will require some changes to the program in its current form, *e.g.*, implementation of molecular dynamics and geometry optimization routines in CREST but first steps in this direction have already been taken. From a theoretical point of view, mainly applications of CREST's conformational entropy procedures are of scientific

interest in the near future. The combination with fast implicit solvation models (ALPB, GBSA) enables for the first time the study of conformational entropy effects in different phases and has yet to be investigated as part of ongoing research.

In summary, the programs and workflows presented in this thesis provide powerful, efficient, and widely applicable methodologies for computational simulations of molecular systems. The developments will pave the way for a range of scientific projects across many fields of theoretical and computational chemistry.

Bibliography

[1] Löwdin, P.-O. *Int. J. Quantum Chem.* **1967**, *1*, 7–12.

[2] Thiel, W. *Angew. Chem. Int. Ed.* **2011**, *50*, 9216–9217.

[3] Grimme, S.; Schreiner, P. R. *Angew. Chem. Int. Ed.* **2017**, *57*, 4170–4176.

[4] Jensen, F. *Introduction to Computational Chemistry*; Wiley: Chichester, UK, 2017.

[5] Krylov, A.; Windus, T. L.; Barnes, T.; Marin-Rimoldi, E.; Nash, J. A.; Pritchard, B.; Smith, D. G. A.; 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.

[6] Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. *Pharmacol. Rev.* **2014**, *66*, 334–395.

[7] Vaissier Welborn, V.; Head-Gordon, T. *Chem. Rev.* **2019**, *119*, 6613–6630.

[8] Marrone, T. J.; Briggs, J. M., and; McCammon, J. A. *Annu. Rev. Pharmacol. Toxicol.* **1997**, *37*, 71–90.

[9] Chen, B. W. J.; Xu, L.; Mavrikakis, M. *Chem. Rev.* **2021**, *121*, 1007–1048.

[10] Ahn, S.; Hong, M.; Sundararajan, M.; Ess, D. H.; Baik, M.-H. *Chem. Rev.* **2019**, *119*, 6509–6560.

[11] Poree, C.; Schoenebeck, F. *Acc. Chem. Res.* **2017**, *50*, 605–608.

[12] Hafner, J.; Wolverton, C.; Ceder, G. *MRS Bulletin* **2006**, *31*, 659–668.

[13] Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. *Nature Mater.* **2013**, *12*, 191–201.

[14] Pedone, A.; Biczysko, M.; Barone, V. *ChemPhysChem* **2010**, *11*, 1812–1832.

[15] Bühl, M.; van Mourik, T. *WIREs Comput. Mol. Sci.* **2011**, *1*, 634–647.

[16] Puzzarini, C.; Bloino, J.; Tasinato, N.; Barone, V. *Chem. Rev.* **2019**, *119*, 8131–8191.

[17] Helgaker, T.; Jørgensen, P.; Olsen, J. *Molecular Electronic-Structure Theory*; Wiley: Chichester, UK, 2002.

[18] Becke, A. D. *J. Chem. Phys.* **2014**, *140*, 18A301.

[19] Burke, K. *J. Chem. Phys.* **2012**, *136*, 150901.

[20] Christensen, A. S.; Kubař, T.; Cui, Q.; Elstner, M. *Chem. Rev.* **2016**, *116*, 5301–5337.

[21] Thiel, W. *WIREs Comput. Mol. Sci.* **2014**, *4*, 145–157.

[22] Grimme, S.; Bohle, F.; Hansen, A.; Pracht, P.; Spicher, S.; Stahn, M. *J. Phys. Chem. A* **2021**, *125*, 4039–4054.

[23] Ufimtsev, I. S.; Martinez, T. J. *J. Chem. Theory Comput.* **2009**, *5*, 2619–2628.

[24] Yasuda, K. *J. Chem. Theory Comput.* **2008**, *4*, 1230–1236.

[25] Wu, X.; Koslowski, A.; Thiel, W. *J. Chem. Theory Comput.* **2012**, *8*, 2272–2281.

[26] Moore, G. *Electronics* **1965**, *38*, 114.

[27] Waldrop, M. M. *Nature* **2016**, *530*, 144–147.

[28] Shalf, J. *Philos. Trans. R. Soc. A* **2020**, *378*, 20190061.

[29] Leiserson, C. E.; Thompson, N. C.; Emer, J. S.; Kuszmaul, B. C.; Lampson, B. W.; Sanchez, D.; Schardl, T. B. *Science* **2020**, *368*.

[30] Lanyon, B. P.; Whitfield, J. D.; Gillett, G. G.; Goggin, M. E.; Almeida, M. P.; Kassal, I.; Biamonte, J. D.; Mohseni, M.; Powell, B. J.; Barbieri, M.; Aspuru-Guzik, A.; White, A. G. *Nature Chem.* **2010**, *2*, 106–111.

[31] McArdle, S.; Endo, S.; Aspuru-Guzik, A.; Benjamin, S. C.; Yuan, X. *Rev. Mod. Phys.* **2020**, *92*, 015003.

[32] Houk, K. N.; Liu, F. *Acc. Chem. Res.* **2017**, *50*, 539–543.

[33] Pracht, P.; Bohle, F.; Grimme, S. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

[34] Kirkpatrick, P.; Ellis, C. *Nature* **2004**, *432*, 823–823.

[35] Reymond, J.-L.; Awale, M. *ACS Chem. Neurosci.* **2012**, *3*, 649–657.

[36] Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. *WIREs Comput. Mol. Sci.* **2020**, e01493; https://doi.org/10.1002/wcms.1493.

[37] Pracht, P.; Caldeweyher, E.; Ehlert, S.; Grimme, S. *ChemRxiv preprint* **2019**, https://doi.org/10.26434/chemrxiv.8326202.v1.

[38] Grimme, S.; Bannwarth, C.; Shushkov, P. *J. Chem. Theory Comput.* **2017**, *13*, 1989–2009.

[39] Bannwarth, C.; Ehlert, S.; Grimme, S. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

[40] Grimme, S.; Bannwarth, C.; Dohm, S.; Hansen, A.; Pisarek, J.; Pracht, P.; Seibert, J.; Neese, F. *Angew. Chem. Int. Ed.* **2017**, *56*, 14763–14769.

[41] Grimme, S. *J. Chem. Theory Comput.* **2019**, *15*, 2847–2862.

[42] Grimme, S. *Chem. Eur. J.* **2012**, *18*, 9955–9964.

[43] Suárez, D.; Díaz, N. *WIREs Comput. Mol. Sci.* **2015**, *5*, 1–26.

[44] Henschel, H.; van der Spoel, D. *J. Phys. Chem. Lett.* **2020**, *11*, 5471–5475.

[45] Işık, M.; Levorse, D.; Rustenburg, A. S.; Ndukwe, I. E.; Wang, H.; Wang, X.; Reibarkh, M.; Martin, G. E.; Makarov, A. A.; Mobley, D. L.; Rhodes, T.; Chodera, J. D. *J. Comput.-Aided Mol. Des.* **2018**, *32*, 1117–1138.

[46] Işık, M.; Rustenburg, A. S.; Rizzi, A.; Gunner, M. R.; Mobley, D. L.; Chodera, J. D. *J. Comput.-Aided Mol. Des.* **2021**, *35*, 131–166.

[47] Engel, T.; Gasteiger, J. *Chemoinformatics: Basic Concepts and Methods*; Wiley-VCH: Weinheim, Germany, 2018.

[48] Born, M.; Oppenheimer, R. *Ann. Phys.* **1927**, *389*, 457–484.

[49] Stone, A. *The Theory of Intermolecular Forces*; Oxford University Press: Oxford, UK, 2016.

[50] Hartree, D. R. *Math. Proc. Cambridge* **1928**, *24*, 89–110.

[51] Fock, V. *Z. Physik* **1930**, *61*, 126–148.

[52] Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133–A1138.

[53] Ziegler, T. *Chem. Rev.* **1991**, *91*, 651–667.

[54] Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.

[55] Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

[56] Kohn, W. *Rev. Mod. Phys.* **1999**, *71*, 1253–1266.

[57] Bredow, T.; Jug, K. *Theor. Chem. Acc.* **2005**, *113*, 1–14.

[58] Pople, J. A.; Santry, D. P.; Segal, G. A. *J. Chem. Phys.* **1965**, *43*, S129–S135.

[59] Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899–4907.

[60] Bredow, T.; Jug, K. In *Encyclopedia of Computational Chemistry (online edition)*; von Ragué Schleyer, P., Ed.; Wiley: New York, 2004.

[61] Weber, W.; Thiel, W. *Theor. Chem. Acc.* **2000**, *103*, 495–506.

[62] Dral, P. O.; Wu, X.; Thiel, W. *J. Chem. Theory Comput.* **2019**, *15*, 1743–1760.

[63] Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.

[64] Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 221–264.

[65] Stewart, J. J. P. *J. Mol. Model.* **2004**, *10*, 155–164.

[66] Stewart, J. J. P. *J. Mol. Model.* **2007**, *13*, 1173.

[67] Stewart, J. J. P. *J. Mol. Model.* **2013**, *19*, 1–32.

[68] Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260–7268.

[69] Gaus, M.; Cui, Q.; Elstner, M. *J. Chem. Theory Comput.* **2011**, *7*, 931–948.

[70] Brandenburg, J. G.; Grimme, S. *Top. Curr. Chem.* **2014**, *345*, 1–23.

[71] Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. *WIREs Comput. Mol. Sci.* **2013**, *3*, 198–210.

[72] Zhu, X.; Lopes, P. E. M.; MacKerell, A. D. *WIREs Comput. Mol. Sci.* **2012**, *2*, 167–185.

[73] Brooks, B. R. et al. *J. Comput. Chem.* **2009**, *30*, 1545–1614.

[74] Banks, J. L. et al. *J. Comput. Chem.* **2005**, *26*, 1752–1780.

[75] Harrison, J. A.; Schall, J. D.; Maskey, S.; Mikulski, P. T.; Knippenberg, M. T.; Morrow, B. H. *App. Phys. Rev.* **2018**, *5*, 031104.

[76] Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

[77] Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490–519.

[78] Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.

[79] Shi, S.; Yan, L.; Yang, Y.; Fisher-Shaulsky, J.; Thacher, T. *J. Comput. Chem.* **2003**, *24*, 1059–1076.

[80] Spicher, S.; Grimme, S. *Angew. Chem. Int. Ed. 132*, 15795–15803.

[81] Riniker, S.; Allison, J. R.; van Gunsteren, W. F. *Phys. Chem. Chem. Phys.* **2012**, *14*, 12423–12430.

[82] Brini, E.; Algaer, E. A.; Ganguly, P.; Li, C.; Rodríguez-Ropero, F.; van der Vegt, N. F. A. *Soft Matter* **2013**, *9*, 2108–2119.

[83] Jacob, C. R.; Beyhan, S. M.; Bulo, R. E.; Gomes, A. S. P.; Götz, A. W.; Kiewisch, K.; Sikkema, J.; Visscher, L. *J. Comput. Chem.* **2011**, *32*, 2328–2338.

[84] Lin, H.; Truhlar, D. G. *Theor. Chem. Acc.* **2006**, *117*, 185.

[85] Brunk, E.; Rothlisberger, U. *Chem. Rev.* **2015**, *115*, 6217–6263.

[86] Schlegel, H. B. *WIREs Comput. Mol. Sci.* **2011**, *1*, 790–809.

[87] Larsson, P.; Hess, B.; Lindahl, E. *WIREs Comput. Mol. Sci.* **2011**, *1*, 93–108.

[88] Tuckerman, M. E. *J. Phys.: Condens. Matter* **2002**, *14*, R1297–R1355.

[89] Hutter, J. *WIREs Comput. Mol. Sci.* **2012**, *2*, 604–612.

[90] Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry*; Dover Publications: New York, 1996.

[91] Dirac, P. A. M. *Math. Proc. Cambridge* **1939**, *35*, 416—-418.

[92] Pauli, W. *Phys. Rev.* **1940**, *58*, 716–722.

[93] Roothaan, C. C. J. *Rev. Mod. Phys.* **1951**, *23*, 69–89.

[94] Hall, G. G. **1951**, *205*, 541–552.

[95] Hill, J. G. *Int. J. Quantum Chem.* **2013**, *113*, 21–34.

[96] Van Lenthe, E.; Baerends, E. J. *J. Comput. Chem.* **2003**, *24*, 1142–1156.

[97] Kresse, G.; Furthmüller, J. *Phys. Rev. B* **1996**, *54*, 11169–11186.

[98] Sure, R.; Brandenburg, J. G.; Grimme, S. *ChemistryOpen* **2016**, *5*, 94–109.

[99] Davidson, E. R.; Feller, D. *Chem. Rev.* **1986**, *86*, 681–696.

[100] Grimme, S.; Hansen, A.; Brandenburg, J. G.; Bannwarth, C. *Chem. Rev.* **2016**, *116*, 5105–5154.

[101] Bannwarth, C. *Development and Application of Efficient Methods for the Computation of Electronic Spectra of Large Systems*; Dissertation: University of Bonn, 2017.

[102] Sherrill, D. C.; Schaefer, H. F. In *The Configuration Interaction Method: Advances in Highly Correlated Approaches*; Löwdin, P.-O., Ed.; Advances in Quantum Chemistry; 1999; Vol. 34; pp 143–269.

[103] Szalay, P. G.; Müller, T.; Gidofalvi, G.; Lischka, H.; Shepard, R. *Chem. Rev.* **2012**, *112*, 108–181.

[104] Bartlett, R. J. *Annu. Rev. Phys. Chem.* **1981**, *32*, 359–401.

[105] Bartlett, R. J.; Musiał, M. *Rev. Mod. Phys.* **2007**, *79*, 291–352.

[106] Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618–622.

[107] Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F. *J. Chem. Phys.* **2013**, *139*, 134101.

[108] Riplinger, C.; Pinski, P.; Becker, U.; Valeev, E. F.; Neese, F. *J. Chem. Phys.* **2016**, *144*, 024109.

[109] Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. *Phys. Chem. Chem. Phys.* **2017**, *19*, 32184–32215.

[110] Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864–B871.

[111] Parr, R. G.; Yang, W. *Annu. Rev. Phys. Chem.* **1995**, *46*, 701–728.

[112] Chai, J.-D.; Weeks, J. D. *Phys. Rev. B* **2007**, *75*, 205122.

[113] Thomas, L. H. *Math. Proc. Cambridge Philos. Soc.* **1927**, *23*, 542–548.

[114] Fermi, E. *Z. Physik* **1928**, *48*, 73–79.

[115] Dirac, P. A. M. *Math. Proc. Cambridge* **1930**, *26*, 376–385.

[116] García-Aldea, D.; Alvarellos, J. E. *J. Chem. Phys.* **2007**, *127*, 144109.

[117] Laricchia, S.; Constantin, L. A.; Fabiano, E.; Della Sala, F. *J. Chem. Theory Comput.* **2014**, *10*, 164–179.

[118] Perdew, J. P.; Schmidt, K. *AIP Conf. Proc.* **2001**, *577*, 1–20.

[119] Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. *J. Phys. Chem. A* **2007**, *111*, 10439–10452.

[120] Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–1377.

[121] Medvedev, M. G.; Bushmarinov, I. S.; Sun, J.; Perdew, J. P.; Lyssenko, K. A. *Science* **2017**, *355*, 49–52.

[122] Sun, J.; Ruzsinszky, A.; Perdew, J. P. *Phys. Rev. Lett.* **2015**, *115*, 036402.

[123] Goerigk, L.; Grimme, S. *Phys. Chem. Chem. Phys.* **2011**, *13*, 6670–6688.

[124] Slater, J. C. *Phys. Rev. 81*, 385–390.

[125] Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.

[126] Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244–13249.

[127] Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.

[128] Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868; erratum *Phys. Rev. Lett.* **78**, 1396 (1997).

[129] Furness, J. W.; Kaplan, A. D.; Ning, J.; Perdew, J. P.; Sun, J. *J. Phys. Chem. Lett.* **2020**, *11*, 8208–8215.

[130] Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.

[131] Vahtras, O.; Almlöf, J.; Feyereisen, M. W. *Chem. Phys. Lett.* **1993**, *213*, 514–518.

[132] Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *242*, 652–660.

[133] Weigend, F. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057–1065.

[134] Csonka, G. I.; Perdew, J. P.; Ruzsinszky, A. *J. Chem. Theory Comput.* **2010**, *6*, 3688–3703.

[135] Harris, J. *Phys. Rev. A* **1984**, *29*, 1648–1659.

[136] Adamson, R. D.; Dombroski, J. P.; Gill, P. M. W. *J. Comput. Chem.* **1999**, *20*, 921–927.

[137] Gill, P. M. W.; Adamson, R. D.; Pople, J. A. *Mol. Phys.* **1996**, *88*, 1005–1010.

[138] Leininger, T.; Stoll, H.; Werner, H.-J.; Savin, A. *Chem. Phys. Lett.* **1997**, *275*, 151–160.

[139] Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540–3544.

[140] Chai, J.-D.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.

[141] Chai, J.-D.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 084106.

[142] Yu, H. S.; Li, S. L.; Truhlar, D. G. *J. Chem. Phys.* **2016**, *145*, 130901.

[143] Goerigk, L.; Grimme, S. *WIREs Comput. Mol. Sci.* **2014**, *4*, 576–600.

[144] Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108.

[145] Brémond, E.; Ciofini, I.; Sancho-García, J. C.; Adamo, C. *Acc. Chem. Res.* **2016**, *49*, 1503–1513.

[146] Eshuis, H.; Bates, J. E.; Furche, F. *Theor. Chem. Acc.* **2012**, *131*, 1084.

[147] Kalai, C.; Mussard, B.; Toulouse, J. *J. Chem. Phys.* **2019**, *151*, 074102.

[148] Kozuch, S.; Gruzman, D.; Martin, J. M. L. *J. Phys. Chem. C* **2010**, 20801–20808.

[149] Kozuch, S.; Martin, J. M. L. *J. Comput. Chem.* **2013**, *34*, 2327–2344.

[150] Santra, G.; Sylvetsky, N.; Martin, J. M. L. *J. Phys. Chem. A* **2019**, *123*, 5129–5143.

[151] Chan, B.; Goerigk, L.; Radom, L. *J. Comput. Chem.* **2016**, *37*, 183–193.

[152] Kruse, H.; Grimme, S. *J. Chem. Phys.* **2012**, *136*, 154101.

[153] Brandenburg, J. G.; Alessio, M.; Civalleri, B.; Peintinger, M. F.; Bredow, T.; Grimme, S. *J. Phys. Chem. A* **2013**, *117*, 9282–9292.

[154] Witte, J.; Neaton, J. B.; Head-Gordon, M. *J. Chem. Phys.* **2017**, *146*, 234105.

[155] Otero-de-la Roza, A.; DiLabio, G. A. *J. Chem. Theory Comput.* **2017**, *13*, 3505–3524.

[156] Kulik, H. J.; Seelam, N.; Mar, B. D.; Martinez, T. J. *J. Phys. Chem. A* **2016**, *120*, 5939–5949.

[157] Sure, R.; Grimme, S. *J. Comput. Chem.* **2013**, *34*, 1672–1685.

[158] Grimme, S.; Brandenburg, J. G.; Bannwarth, C.; Hansen, A. *J. Chem. Phys.* **2015**, *143*, 054107.

[159] Eisenschitz, R.; London, F. *Z. Phys.* **1930**, *60*, 491–527.

[160] London, F. *Z. Phys.* **1930**, *63*, 245–279.

[161] Klimeş, J.; Michaelides, A. *J. Chem. Phys.* **2012**, *137*, 120901.

[162] Kristyán, S.; Pulay, P. *Chem. Phys. Lett.* **1994**, *229*, 175–180.

[163] Wu, Q.; Yang, W. *J. Chem. Phys.* **2002**, *116*, 515–524.

[164] Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2005**, *123*, 154101.

[165] Johnson, E. R.; Becke, A. D. *J. Chem. Phys.* **2005**, *123*, 024101.

[166] Tkatchenko, A.; Scheffler, M. *Phys. Rev. Lett.* **2009**, *102*, 073005.

[167] Tkatchenko, A.; DiStasio, R. A.; Car, R.; Scheffler, M. *Phys. Rev. Lett.* **2012**, *108*, 236402.

[168] Lee, K.; Murray, E. D.; Kong, L.; Lundqvist, B. I.; Langreth, D. C. *Phys. Rev. B* **2010**, *82*, 081101.

[169] Vydrov, O. A.; Van Voorhis, T. *J. Chem. Phys.* **2010**, *133*, 244103.

[170] Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463–1473.

[171] Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787–1799.

[172] Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104.

[173] Grimme, S.; Ehrlich, S.; Goerigk, L. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

[174] Caldeweyher, E.; Ehlert, S.; Hansen, A.; Neugebauer, H.; Spicher, S.; Bannwarth, C.; Grimme, S. *J. Chem. Phys.* **2019**, *150*, 154122.

[175] Witte, J.; Mardirossian, N.; Neaton, J. B.; Head-Gordon, M. *J. Chem. Theory Comput.* **2017**, *13*, 2043–2052.

[176] Axilrod, B. M.; Teller, E. *J. Chem. Phys.* **1943**, *11*, 299–300.

[177] Muto, Y. *Proc. Phys. Math. Soc. Jpn.* **1943**, *17*, 629.

[178] Caldeweyher, E.; Bannwarth, C.; Grimme, S. *J. Chem. Phys.* **2017**, *147*, 034112.

[179] Bursch, M.; Caldeweyher, E.; Hansen, A.; Neugebauer, H.; Ehlert, S.; Grimme, S. *Acc. Chem. Res.* **2019**, *52*, 258–266.

[180] Brandenburg, J. G.; Bannwarth, C.; Hansen, A.; Grimme, S. *J. Chem. Phys.* **2018**, *148*, 064104.

[181] Brandenburg, J. G.; Caldeweyher, E.; Grimme, S. *Phys. Chem. Chem. Phys.* **2016**, *18*, 15519–15523.

[182] Grimme, S.; Hansen, A.; Ehlert, S.; Mewes, J.-M. *J. Chem. Phys.* **2021**, *154*, 064103.

[183] Bursch, M.; Hansen, A.; Pracht, P.; Kohn, J. T.; Grimme, S. *Phys. Chem. Chem. Phys.* **2021**, *23*, 287–299.

[184] Dohm, S.; Hansen, A.; Steinmetz, M.; Grimme, S.; Checinski, M. P. *J. Chem. Theory Comput.* **2018**, *14*, 2596–2608.

[185] Neugebauer, H.; Bohle, F.; Bursch, M.; Hansen, A.; Grimme, S. *J. Phys. Chem. A* **2020**, *124*, 7166–7176.

[186] Pracht, P.; Grimme, S. *Chem. Sci.* **2021**, *12*, 6551–6568.

[187] Elstner, M. *Theor. Chem. Acc.* **2006**, *116*, 316–325.

[188] Hoffmann, R. *J. Chem. Phys.* **1963**, *39*, 1397–1412.

[189] Imamura, A. *J. Chem. Phys.* **1970**, *52*, 3168–3175.

[190] Seifert, G.; Porezag, D.; Frauenheim, T. *Int. J. Quantum Chem.* **1996**, *58*, 185–192.

[191] Porezag, D.; Frauenheim, T.; Köhler, T.; Seifert, G.; Kaschner, R. *Phys. Rev. B* **1995**, *51*, 12947–12957.

[192] Nishimoto, K.; Mataga, N. *Z. Phys. Chem.* **1957**, *12*, 335–338.

[193] Ohno, K. *Theor. Chim. Act.* **1964**, *2*, 219.

[194] Klopman, G. *J. Am. Chem. Soc.* **1964**, *86*, 4450.

[195] Grimme, S. *J. Chem. Phys.* **2013**, *138*, 244104.

[196] Gaus, M.; Goez, A.; Elstner, M. *J. Chem. Theory Comput.* **2013**, *9*, 338–354.

[197] Gaus, M.; Lu, X.; Elstner, M.; Cui, Q. *J. Chem. Theory Comput.* **2014**, *10*, 1518–1537.

[198] Kubillus, M.; Kubař, T.; Gaus, M.; Řezáč, J.; Elstner, M. *J. Chem. Theory Comput.* **2015**, *11*, 332–342.

[199] Zheng, G.; Witek, H. A.; Bobadova-Parvanova, P.; Irle, S.; Musaev, D. G.; Prabhakar, R.; Morokuma, K.; Lundberg, M.; Elstner, M.; Köhler, C.; Frauenheim, T. *J. Chem. Theory Comput.* **2007**, *3*, 1349–1367.

[200] Gaus, M.; Jin, H.; Demapan, D.; Christensen, A. S.; Goyal, P.; Elstner, M.; Cui, Q. *J. Chem. Theory Comput.* **2015**, *11*, 4205–4219.

[201] Grimme, S.; Bannwarth, C. *J. Chem. Phys.* **2016**, *145*, 054103.

[202] Sokalski, W.; Poirier, R. *Chem. Phys. Lett.* **1983**, *98*, 86–92.

[203] Sokalski, W.; Sawaryn, A. *J. Mol. Struct. THEOCHEM* **1992**, *256*, 91–112.

[204] Spicher, S.; Grimme, S. *J. Phys. Chem. Lett.* **2020**, *11*, 6606–6611.

[205] Hehre, W. J.; Stewart, R. F.; Pople, J. A. *J. Chem. Phys.* **1969**, *51*, 2657–2664.

[206] Mermin, N. D. *Phys. Rev. A* **1965**, *137*, 1441–1443.

[207] Helgaker, T.; Coriani, S.; Jørgensen, P.; Kristensen, K.; Olsen, J.; Ruud, K. *Chem. Rev.* **2012**, *112*, 543–631.

[208] Güttinger, P. *Z. Physik* **1932**, *73*, 169–184.

[209] Hellmann, H. *Einführung in die Quantenchemie*; Franz Deuticke: Leipzig, Germany, 1937.

[210] Feynman, R. P. *Phys. Rev.* **1939**, *56*, 340–343.

[211] Suárez, A.; Silbey, R. *J. Chem. Phys.* **1991**, *94*, 4809–4816.

[212] Schäfer, M.; Peckelsen, K.; Paul, M.; Martens, J.; Oomens, J.; Berden, G.; Berkessel, A.; Meijer, A. J. H. M. *J. Am. Chem. Soc.* **2017**, *139*, 5779–5786.

[213] Hill, T. L. *An Introduction to Statistical thermodynamics*; Dover Publications Inc.: New York, USA, 2020.

[214] Tolman, R. C. *The Principles of Statistical Mechanics*; Dover Publications Inc.: New York, USA, 2019.

[215] Irikura, K.; Frurip, D. J. *Computational thermochemistry: prediction and estimation of molecular thermodynamics*; American Chemical Society, 1998.

[216] Jensen, J. H. *Phys. Chem. Chem. Phys.* **2015**, *17*, 12441–12451.

[217] Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161–2200.

[218] Klamt, A. *WIREs Comput. Mol. Sci.* **2018**, *8*, e1338.

[219] Seeman, J. I. *Chem. Rev.* **1983**, *83*, 83–134.

[220] Pracht, P.; Bauer, C. A.; Grimme, S. *J. Comput. Chem.* **2017**, *38*, 2618–2631.

[221] Pracht, P.; Wilcken, R.; Udvarhelyi, A.; Rodde, S.; Grimme, S. *J. Comput.-Aided Mol. Des.* **2018**, *32*, 1139–1149.

[222] Salsbury, F. R. *Curr. Opin. Pharmacol.* **2010**, *10*, 738–744.

[223] Xu, J.; Hagler, A. *Molecules* **2002**, *7*, 566–600.

[224] Ghose, A. K.; Herbertz, T.; Salvino, J. M.; Mallamo, J. P. *Drug Discovery Today* **2006**, *11*, 1107–1114.

[225] Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. *Drug Discovery Today* **2018**, *23*, 1538–1546.

[226] Lavecchia, A. *Drug Discovery Today* **2015**, *20*, 318–331.

[227] Kuhn, B. et al. *J. Med. Chem.* **2016**, *59*, 4087–4102.

[228] Hawkins, P. C. D. *J. Chem. Inf. Model.* **2017**, *57*, 1747–1756.

[229] Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. *J. Chem. Inf. Model.* **2010**, *50*, 572–584.

[230] Vainio, M. J.; Johnson, M. S. *J. Chem. Inf. Model.* **2007**, *47*, 2462–2474.

[231] Riniker, S.; Landrum, G. A. *J. Chem. Inf. Model.* **2015**, *55*, 2562–2574.

[232] Miteva, M. A.; Guyon, F.; Tufféry, P. *Nucleic Acids Res.* **2010**, *38*, W622–W627.

[233] Tai, K. *Biophys. Chem.* **2004**, *107*, 213–220.

[234] Dorfman, R. J.; Smith, K. M.; Masek, B. B.; Clark, R. D. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 681–691.

[235] Sauton, N.; Lagorce, D.; Villoutreix, B. O.; Miteva, M. A. *BMC Bioinformatics* **2008**, *9*, 184.

[236] Watts, K. S.; Dalal, P.; Murphy, R. B.; Sherman, W.; Friesner, R. A.; Shelley, J. C. *J. Chem. Inf. Model.* **2010**, *50*, 534–546.

[237] Kolossváry, I.; Guida, W. C. *J. Am. Chem. Soc.* **1996**, *118*, 5011–5019.

[238] Friedrich, N.-O.; Flachsenberg, F.; Meyder, A.; Sommer, K.; Kirchmair, J.; Rarey, M. *J. Chem. Inf. Model.* **2019**, *59*, 731–742.

[239] Bonnet, P.; Agrafiotis, D. K.; Zhu, F.; Martin, E. *J. Chem. Inf. Model.* **2009**, *49*, 2242–2259.

[240] Watts, K. S.; Dalal, P.; Tebben, A. J.; Cheney, D. L.; Shelley, J. C. *J. Chem. Inf. Model.* **2014**, *54*, 2680–2696.

[241] Coutsias, E. A.; Lexa, K. W.; Wester, M. J.; Pollock, S. N.; Jacobson, M. P. *J. Chem. Theory Comput.* **2016**, *12*, 4674–4687.

[242] Cleves, A. E.; Jain, A. N. *J. Comput.-Aided Mol. Des.* **2017**, *31*, 419–439.

[243] Fuster, F.; Silvi, B. *Chemical Physics* **2000**, *252*, 279–287.

[244] Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 681–691.

[245] Liao, C.; Nicklaus, M. C. *J. Chem. Inf. Model.* **2009**, *49*, 2801–2812.

[246] Sanz, J. F.; Anguiano, J.; Vilarrasa, J. *J. Comput. Chem.* **1988**, *9*, 784–789.

[247] El Yazal, J.; Prendergast, F. G.; Shaw, D. E.; Pang, Y.-P. *J. Am. Chem. Soc.* **2000**, *122*, 11411–11415.

[248] Namazian, M.; Heidary, H. *J. Mol. Struc. THEOCHEM* **2003**, *620*, 257–263.

[249] Kromann, J. C.; Larsen, F.; Moustafa, H.; Jensen, J. H. *PeerJ* **2016**, *4*, e2335.

[250] Klicić, J. J.; Friesner, R. A.; Liu, S.-Y.; Guida, W. C. *J. Phys. Chem. A* **2002**, *106*, 1327–1335.

[251] Cruciani, G.; Milletti, F.; Storchi, L.; Sforna, G.; Goracci, L. *Chem. Biodivers.* **2009**, *6*, 1812–1821.

[252] Klamt, A.; Eckert, F.; Diedenhofen, M.; Beck, M. E. *J. Phys. Chem. A* **2003**, *107*, 9380–9386.

[253] Martin, Y. C. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 693.

[254] Sitzmann, M.; Ihlenfeldt, W.-D.; Nicklaus, M. C. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 521–551.

[255] Asgeirsson, V.; Bauer, C. A.; Grimme, S. *Chem. Sci.* **2017**, *8*, 4879–4895.

[256] Grimme, S. *Angew. Chem. Int. Ed.* **2013**, *52*, 6306–6312.

[257] Bauer, C. A.; Grimme, S. *J. Phys. Chem. A* **2016**, *120*, 3755–3766.

[258] Maeda, S.; Morokuma, K. *J. Chem. Phys.* **2010**, *132*, 241102.

[259] Maeda, S.; Harabuchi, Y.; Takagi, M.; Taketsugu, T.; Morokuma, K. *Chem. Rec.* **2016**, *16*, 2232–2248.

[260] Haag, M. P.; Vaucher, A. C.; Bosson, M.; Redon, S.; Reiher, M. *ChemPhysChem* **2014**, *15*, 3301–3319.

[261] Stein, C. J.; Reiher, M. *J. Comput. Chem.* **2019**,

[262] Zeist, W.-J. V.; Guerra, C. F.; Bickelhaupt, F. M. *J. Comput. Chem.* **2008**, *29*, 312–315.

[263] Wang, L.-P.; McGibbon, R. T.; Pande, V. S.; Martinez, T. J. *J. Chem. Theory Comput.* **2016**, *12*, 638–649.

[264] Rappoport, D.; Galvin, C. J.; Zubarev, D. Y.; Aspuru-Guzik, A. *J. Chem. Theory Comput.* **2014**, *10*, 897–907.

[265] Zimmerman, P. M. *J. Comput. Chem.* **2013**, *34*, 1385–1392.

[266] Guan, Y.; Wheeler, S. E. *Angew. Chem. Int. Ed.* **2017**, *56*, 9101–9105.

[267] Zapata, F.; Ridder, L.; Hidding, J.; Jacob, C. R.; Infante, I.; Visscher, L. *J. Chem. Inf. Model.* **2019**, *59*, 3191–3197.

[268] Ohno, K.; Maeda, S. *J. Phys. Chem. A* **2006**, *110*, 8933–8941.

[269] Maeda, S.; Morokuma, K. *J. Chem. Theory Comput.* **2011**, *7*, 2335–2345.

[270] Maeda, S.; Ohno, K.; Morokuma, K. *Phys. Chem. Chem. Phys.* **2013**, *15*, 3683–3701.

[271] Maeda, S.; Taketsugu, T.; Morokuma, K. *J. Comput. Chem.* **2014**, *35*, 166–173.

[272] Maeda, S.; Harabuchi, Y.; Ono, Y.; Taketsug, T.; Morokuma, K. *Int. J. Quantum Chem.* **2015**, *115*, 258–269.

[273] Nagahata, Y.; Maeda, S.; Teramoto, H.; Horiyama, T.; Taketsugu, T.; Komatsuzaki, T. *J. Phys. Chem. B* **2016**, *120*, 1961–1971.

[274] Bergeler, M.; Simm, G. N.; Proppe, J.; Reiher, M. *J. Chem. Theory Comput.* **2015**, *11*, 5712–5722.

[275] Proppe, J.; Husch, T.; Simm, G. N.; Reiher, M. *Faraday Discuss.* **2016**, *195*, 497–520.

[276] Simm, G. N.; Vaucher, A. C.; Reiher, M. *J. Phys. Chem. A* **2018**, *123*, 385–399.

[277] Simm, G. N.; Reiher, M. *J. Chem. Theory Comput.* **2018**, *14*, 5238–5248.

[278] Simm, G. N.; Vaucher, A. C.; Reiher, M. *J. Phys. Chem. A* **2019**, *123*, 385–399.

[279] Proppe, J.; Reiher, M. *J. Chem. Theory Comput.* **2019**, *15*, 357–370.

[280] Schlegel, H. B. *J. Comput. Chem.* **2003**, *24*, 1514–1527.

[281] Dewyer, A. L.; Zimmerman, P. M. *Org. Biomol. Chem.* **2017**, *15*, 501–504.

[282] Larsen, A. H. et al. *J. Phys.: Condens. Matter* **2017**, *29*, 273002.

[283] te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Fonseca Guerra, C.; van Gisbergen, S. J. A.; Snijders, J. G.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 931–967.

[284] Baerends, E. J. e. ADF2017, SCM, Theoretical Chemistry, Vrije Universiteit, Amsterdam, The Netherlands, https://www.scm.com.

[285] Seibert, J.; Bannwarth, C.; Grimme, S. *J. Am. Chem. Soc.* **2017**, *139*, 11682–11685.

[286] Ben-Efraim, D.; Green, B. *Tetrahedron* **1974**, *30*, 2357–2364.

[287] Saito, H. *Magn. Reson. Chem.* **1986**, *24*, 835–852.

[288] Barone, G.; Duca, D.; Silvestri, A.; Gomez-Paloma, L.; Riccio, R.; Bifulco, G. *Chem. Eur. J.* **2002**, *8*, 3240–3245.

[289] Glättli, A.; Daura, X.; Seebach, D.; van Gunsteren, W. F. *J. Am. Chem. Soc.* **2002**, *124*, 12972–12978.

[290] Brkljača, Z.; Mališ, M.; Smith, D. M.; Smith, A.-S. *J. Chem. Theory Comput.* **2014**, *10*, 3270–3279.

[291] Mori, T.; Grimme, S.; Inoue, Y. *J. Org. Chem.* **2007**, *72*, 6998–7010.

[292] van Vlijmen, H. W.; Schaefer, M.; Karplus, M. *Proteins* **1998**, *33*, 145–158.

[293] Cavasin, A. T.; Hillisch, A.; Uellendahl, F.; Schneckener, S.; Göller, A. H. *J. Chem. Inf. Model.* **2018**, *58*, 1005–1020.

[294] Seddon, M. P.; Cosgrove, D. A.; Packer, M. J.; Gillet, V. J. *J. Chem. Inf. Model.* **2019**, *59*, 98–116.

[295] Friedrich, N.-O.; de Bruyn Kops, C.; Flachsenberg, F.; Sommer, K.; Rarey, M.; Kirchmair, J. *J. Chem. Inf. Model.* **2017**, *57*, 2719–2728.

[296] Mitzel, N. W.; Rankin, D. W. H. *Dalton Trans.* **2003**, 3650.

[297] Blomeyer, S.; Linnemannstöns, M.; Nissen, J. H.; Paulus, J.; Neumann, B.; Stammler, H.-G.; Mitzel, N. W. *Angew. Chem. Int. Ed.* **2017**, *56*, 13259–13263.

[298] Feng, Z.; Chen, L.; Maddula, H.; Akcan, O.; Oughtred, R.; Berman, H. M.; Westbrook, J. *Bioinformatics* **2004**, *20*, 2153–2155.

[299] Friedrich, N.-O.; Meyder, A.; de Bruyn Kops, C.; Sommer, K.; Flachsenberg, F.; Rarey, M.; Kirchmair, J. *J. Chem. Inf. Model.* **2017**, *57*, 529–539.

[300] Pavlov, T.; Todorov, M.; Stoyanova, G.; Schmieder, P.; Aladjov, H.; Serafimova, R.; Mekenyan, O. *J. Chem. Inf. Model.* **2007**, *47*, 851–863.

[301] Gutten, O.; Bím, D.; Řezáč, J.; Rulíšek, L. *J. Chem. Inf. Model.* **2018**, *58*, 48–60.

[302] Coutsias, E. A.; Seok, C.; Dill, K. A. *J. Comput. Chem.* **2004**, *25*, 1849–1857.

[303] Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci.* **2002**, *99*, 12562–12566.

[304] Barducci, A.; Bonomi, M.; Parrinello, M. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 826–843.

[305] Spiwok, V.; Lipovová, P.; Králová, B. *J. Phys. Chem. B* **2007**, *111*, 3073–3076.

[306] Wen, E. Z.; Hsieh, M.-J.; Kollman, P. A.; Luo, R. *J. Mol. Graph. Model* **2004**, *22*, 415–424.

[307] Leone, V.; Marinelli, F.; Carloni, P.; Parrinello, M. *Curr. Opin. Struct. Biol.* **2010**, *20*, 148–154.

[308] Deriu, M. A.; Grasso, G.; Tuszynski, J. A.; Gallo, D.; Morbiducci, U.; Danani, A. *PLOS Comput. Biol.* **2016**, *12*, 1–14.

[309] Vymětal, J.; Vondrášek, J. *J. Phys. Chem. B* **2010**, *114*, 5632–5642.

[310] Hamelberg, D.; Mongan, J.; McCammon, J. A. *J. Chem. Phys.* **2004**, *120*, 11919–11929.

[311] Torrie, G.; Valleau, J. *J. Comput. Phys.* **1977**, *23*, 187–199.

[312] Wu, X.; Wang, S. *J. Chem. Phys.* **1999**, *110*, 9401–9410.

[313] Rahman, J. A.; Tully, J. C. *J. Chem. Phys.* **2002**, *116*, 8750–8760.

[314] Berne, B. J.; Straub, J. E. *Curr. Opin. Struct. Biol.* **1997**, *7*, 181–189.

[315] Steiner, M. M.; Genilloud, P.-A.; Wilkins, J. W. *Phys. Rev. B* **1998**, *57*, 10236–10239.

[316] Pal, S.; Fichthorn, K. *Chem. Eng. J.* **1999**, *74*, 77–83.

[317] Gong, X. G.; Wilkins, J. W. *Phys. Rev. B* **1999**, *59*, 54–57.

[318] de Oliveira, C. A. F.; Hamelberg, D.; McCammon, J. A. *J. Chem. Theory Comput.* **2008**, *4*, 1516–1525.

[319] Kamenik, A. S.; Lessel, U.; Fuchs, J. E.; Fox, T.; Liedl, K. R. *J. Chem. Inf. Model* **2018**, *58*, 982–992.

[320] Miao, Y.; Sinko, W.; Pierce, L.; Bucher, D.; Walker, R. C.; McCammon, J. A. *J. Chem. Theory Comput.* **2014**, *10*, 2677–2689.

[321] Hibbert, D. B. *Chemometrics Intell. Lab. Sys.* **1993**, *19*, 277–293.

[322] Leardi, R. *J. Chemometrics* **2001**, *15*, 559–569.

[323] Valdes, H.; Pluhackova, K.; Pitonák, M.; Řezáč, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2747–2757.

[324] Bursch, M.; Neugebauer, H.; Grimme, S. *Angew. Chem. Int. Ed.* **2019**, *58*, 11078–11087.

[325] Marianski, M.; Supady, A.; Ingram, T.; Schneider, M.; Baldauf, C. *J. Chem. Theory Comput.* **2016**, *12*, 6157–6168.

[326] 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.

[327] Řezáč, J.; Hobza, P. *J. Chem. Theory Comput.* **2012**, *8*, 141–151.

[328] Řezáč, J.; Bím, D.; Gutten, O.; Rulíšek, L. *J. Chem. Theory Comput.* **2018**, *14*, 1254–1266.

[329] Sharapa, D. I.; Genaev, A.; Cavallo, L.; Minenkov, Y. *ChemPhysChem* **2019**, *20*, 92–102.

[330] Wiberg, K. B. *Tetrahedron* **1968**, *24*, 1083–1096.

[331] Mayer, I. *J. Comput. Chem.* **2007**, *28*, 204–221.

[332] Kier, L. B. *QSAR* **1989**, *8*, 221–224.

[333] Fisanick, W.; Cross, K. P.; Rusinko, A. *Tetrahedron Comput. Methodol.* **1990**, *3*, 635–652.

[334] Weigend, F. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057.

[335] Domingos, S. R.; Pérez, C.; Medcraft, C.; Pinacho, P.; Schnell, M. *Phys. Chem. Chem. Phys.* **2016**, *18*, 16682–16689.

[336] National Institute of Advanced Industrial Science and Technology, SDBSWeb. 2019; https://sdbs.db.aist.go.jp.

[337] Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

[338] Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

[339] Klamt, A.; Schüürmann, G. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799–805.

[340] Klamt, A.; Diedenhofen, M. *J. Phys. Chem. A* **2015**, *119*, 5439–5445.

[341] Chen, I.-J.; Foloppe, N. *Bioorg. Med. Chem.* **2013**, *21*, 7898–7920.

[342] Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. *Acta Crystallogr. B* **2016**, *72*, 171–179.

[343] Hudgins, R. R.; Jarrold, M. F. *J. Am. Chem. Soc.* **1999**, *121*, 3494–3501.

[344] Jarrold, M. F. *Phys. Chem. Chem. Phys.* **2007**, *9*, 1659–1671.

[345] Schubert, F.; Rossi, M.; Baldauf, C.; Pagel, K.; Warnke, S.; von Helden, G.; Filsinger, F.; Kupser, P.; Meijer, G.; Salwiczek, M.; Koksch, B.; Scheffler, M.; Blum, V. *Phys. Chem. Chem. Phys.* **2015**, *17*, 7373–7385.

[346] Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604–1608.

[347] Hansmann, U. H. *Chem. Phys. Lett.* **1997**, *281*, 140–150.

[348] Marinari, E.; Parisi, G. *Europhysics Letters (EPL)* **1992**, *19*, 451–458.

[349] Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

[350] Marković, M.; Judaš, N.; Sabolović, J. *Inorg. Chem.* **2011**, *50*, 3632–3644.

[351] Bette, M.; Kluge, T.; Schmidt, J.; Steinborn, D. *Organometallics* **2013**, *32*, 2216–2227.

[352] Pandit, S.; De, M. *J. Phys. Chem. C* **2017**, *121*, 600–608.

[353] Goenka, S.; Sant, V.; Sant, S. *J. Control. Release* **2014**, *173*, 75–88.

[354] McNaught, A. D.; Wilkinson, A. *IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book").*; Blackwell Scientific Publications: Oxford, 1997.

[355] Medvedev, M. G.; Panova, M. V.; Chilov, G. G.; Bushmarinov, I. S.; Novikov, F. N.; Stroganov, O. V.; Zeifman, A. A.; Svitanko, I. V. *Mendeleev Communications* **2017**, *27*, 224–227.

[356] Jindal, G.; Warshel, A. *J. Phys. Chem. B* **2016**, *120*, 9913–9921.

[357] Bryantsev, V. S.; Diallo, M. S.; van Duin, A. C. T.; Goddard, W. A. *J. Chem. Theory Comput.* **2009**, *5*, 1016–1026.

[358] Seifert, N. A.; Hazrah, A. S.; Jäger, W. *J. Phys. Chem. Lett.* **2019**, *10*, 2836–2841.

[359] Risthaus, T.; Steinmetz, M.; Grimme, S. *J. Comput. Chem.* **2014**, *35*, 1509–1516.

[360] Lalli, P. M.; Iglesias, B. A.; Toma, H. E.; de Sa, G. F.; Daroda, R. J.; Silva Filho, J. C.; Szulejko, J. E.; Araki, K.; Eberlin, M. N. *J. Mass Spectrom.* **2012**, *47*, 712–719.

[361] Warnke, S.; Seo, J.; Boschmans, J.; Sobott, F.; Scrivens, J. H.; Bleiholder, C.; Bowers, M. T.; Gewinner, S.; Schöllkopf, W.; Pagel, K.; von Helden, G. *J. Am. Chem. Soc.* **2015**, *137*, 4236–4242.

[362] Seo, J.; Warnke, S.; Gewinner, S.; Schöllkopf, W.; Bowers, M. T.; Pagel, K.; von Helden, G. *Phys. Chem. Chem. Phys.* **2016**, *18*, 25474–25482.

[363] Xia, H.; Attygalle, A. B. *J. Am. Soc. Mass Spectrom.* **2017**, *28*, 2580–2587.

[364] Zhu, Y.; Hamlow, L. A.; He, C. C.; Strobehn, S. F.; Lee, J. K.; Gao, J.; Berden, G.; Oomens, J.; Rodgers, M. T. *J. Phys. Chem. B* **2016**, *120*, 8892–8904.

[365] Antonov, L.; Deneva, V.; Simeonov, S.; Kurteva, V.; Nedeltcheva, D.; Wirz, J. *Angew. Chem. Int. Ed.* **2009**, *48*, 7875–7878.

[366] Pospisil, P.; Ballmer, P.; Scapozza, L.; Folkers, G. *J. Recept. Signal Transduct.* **2003**, *23*, 361–371.

[367] Guasch, L.; Yapamudiyansel, W.; Peach, M. L.; Kelley, J. A.; Barchi, J. J.; Nicklaus, M. C. *J. Chem. Inf. Model.* **2016**, *56*, 2149–2161.

[368] Mezey, P. G.; Ladik, J. J. *Theor. Chim. Acta* **1979**, *52*, 129–145.

[369] Nir, E.; Janzen, C.; Imhof, P.; Kleinermanns, K.; de Vries, M. S. *J. Chem. Phys.* **2001**, *115*, 4604–4611.

[370] Langer, H.; Doltsinis, N. L. *J. Chem. Phys.* **2003**, *118*, 5400–5407.

[371] Chen, H.; Li, S. *J. Phys. Chem. A* **2006**, *110*, 12360–12362.

[372] Marian, C. M. *J. Phys. Chem. A* **2007**, *111*, 1545–1553.

[373] Quintana, L. M. A.; Johnson, S. I.; Corona, S. L.; Villatoro, W.; Goddard, W. A.; Takase, M. K.; VanderVelde, D. G.; Winkler, J. R.; Gray, H. B.; Blakemore, J. D. *Proc. Natl. Acad. Sci.* **2016**, *113*, 6409–6414.

[374] Clausius, R. *Ann. Phys.* **1865**, *201*, 353–400.

[375] Shannon, C. E.; Weaver, W. *The Mathematical Theory of Communication*; The University of Illinois Press: Urbana, IL, 1964.

[376] Gilson, M.; Given, J.; Bush, B.; McCammon, J. *Biophys. J.* **1997**, *72*, 1047–1069.

[377] East, A. L. L.; Radom, L. *J. Chem. Phys.* **1997**, *106*, 6655–6674.

[378] Njegic, B.; Gordon, M. S. *J. Chem. Phys.* **2006**, *125*, 224102.

[379] DeTar, D. F. *J. Phys. Chem. A* **2007**, *111*, 4464–4477.

[380] Li, Y.-P.; Bell, A. T.; Head-Gordon, M. *J. Chem. Theory Comput.* **2016**, *12*, 2861–2870.

[381] Scott, A. P.; Radom, L. *J. Phys. Chem.* **1996**, *100*, 16502–16513.

[382] Merrick, J. P.; Moran, D.; Radom, L. *J. Phys.Chem. A* **2007**, *111*, 11683–11700.

[383] Kesharwani, M. K.; Brauer, B.; Martin, J. M. L. *J. Phys. Chem. A* **2015**, *119*, 1701–1714.

[384] Johnson, R. D.; Irikura, K. K.; Kacker, R. N.; Kessel, R. *J. Chem. Theory Comput.* **2010**, *6*, 2822–2828.

[385] Baker, J.; Jarzecki, A. A.; Pulay, P. *J. Phys. Chem. A* **1998**, *102*, 1412–1424.

[386] Laury, M. L.; Boesch, S. E.; Haken, I.; Sinha, P.; Wheeler, R. A.; Wilson, A. K. *J. Comput. Chem.* **2011**, *32*, 2339–2347.

[387] Pracht, P.; Grant, D. F.; Grimme, S. *J. Chem. Theory Comput.* **2020**, *16*, 7044–7060.

[388] Piccini, G.; Sauer, J. *J. Chem. Theory Comput.* **2013**, *9*, 5038–5045.

[389] Piccini, G.; Sauer, J. *J. Chem. Theory Comput.* **2014**, *10*, 2479–2487.

[390] Piccini, G.; Alessio, M.; Sauer, J.; Zhi, Y.; Liu, Y.; Kolvenbach, R.; Jentys, A.; Lercher, J. A. *J. Phys. Chem. C* **2015**, *119*, 6128–6137.

[391] Van Speybroeck, V.; Van Neck, D.; Waroquier, M. *J. Phys. Chem. A* **2002**, *106*, 8945–8950.

[392] Vansteenkiste, P.; Van Neck, D.; Van Speybroeck, V.; Waroquier, M. *J. Chem. Phys.* **2006**, *124*, 044314.

[393] Simón-Carballido, L.; Bao, J. L.; Alves, T. V.; Meana-Pañeda, R.; Truhlar, D. G.; Fernńdez-Ramos, A. *J. Chem. Theory Comput.* **2017**, *13*, 3478–3492.

[394] Zheng, J.; Yu, T.; Papajak, E.; Alecu, I. M.; Mielke, S. L.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2011**, *13*, 10885–10907.

[395] Yu, T.; Zheng, J.; Truhlar, D. G. *Chem. Sci.* **2011**, *2*, 2199–2213.

[396] Zheng, J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2013**, *9*, 1356–1367.

[397] Wales, D. J. *Phys. Rev. E* **2017**, *95*, 030105.

[398] Wales, D. J. *Annu. Rev. Phys. Chem.* **2018**, *69*, 401–425.

[399] Gao, X.; Gallicchio, E.; Roitberg, A. E. *J. Chem. Phys.* **2019**, *151*, 034113.

[400] Chen, W.; Chang, C.-E.; Gilson, M. K. *Biophys. Jo.* **2004**, *87*, 3035–3049.

[401] DeTar, D. F. *J. Phys. Chem. A* **1998**, *102*, 5128–5141.

[402] Guthrie, J. P. *J. Phys. Chem. A* **2001**, *105*, 8495–8499.

[403] Chang, C.-E.; Potter, M. J.; Gilson, M. K. *J. Phys. Chem. B* **2003**, *107*, 1048–1055.

[404] Chang, C.-e. A.; Chen, W.; Gilson, M. K. *Proc. Nat. Acad. Sci. USA* **2007**, *104*, 1534–1539.

[405] Pereira, G. P.; Cecchini, M. *J. Chem. Theory Comput.* **2021**, *17*, 1133–1142.

[406] Killian, B. J.; Yundenfreund Kravitz, J.; Gilson, M. K. *J. Chem. Phys.* **2007**, *127*, 024107.

[407] Hnizdo, V.; Tan, J.; Killian, B. J.; Gilson, M. K. *J. Comput. Chem.* **2008**, *29*, 1605–1614.

[408] King, B. M.; Tidor, B. *Bioinformatics* **2009**, *25*, 1165–1172.

[409] King, B. M.; Silver, N. W.; Tidor, B. *J. Phys. Chem. B* **2012**, *116*, 2891–2904.

[410] Suárez, E.; Díaz, N.; Suárez, D. *J. Chem. Theory Comput.* **2011**, *7*, 2638–2653.

[411] Suárez, E.; Díaz, N.; Méndez, J.; Suárez, D. *J. Comput. Chem.* **2013**, *34*, 2041–2054.

[412] Suárez, D.; Díaz, N. *J. Chem. Theory Comput.* **2014**, *10*, 4718–4729.

[413] Jain, A.; Yang, G.; Yalkowsky, S. H. *Ind. Eng. Chem. Res.* **2004**, *43*, 4376–4379.

[414] Chan, L.; Morris, G.; Hutchison, G. *J. Chem. Theory Comput.* **2021**, *17*, 2099–2106.

[415] Karplus, M.; Kushick, J. N. *Macromolecules* **1981**, *14*, 325–332.

[416] Karplus, M.; Ichiye, T.; Pettitt, B. *Biophys. J.* **1987**, *52*, 1083–1085.

[417] Doig, A. J.; Sternberg, M. J. E. *Protein Sci.* **1995**, *4*, 2247–2251.

[418] Ribeiro, R. F.; Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2011**, *115*, 14556–14562.

[419] Ghahremanpour, M. M.; van Maaren, P. J.; Ditz, J. C.; Lindh, R.; van der Spoel, D. *J. Chem. Phys.* **2016**, *145*, 114305.

[420] Spicher, S.; Grimme, S. *J. Chem. Theory Comput.* **2021**, *17*, 1701–1714.

[421] Levenberg, K. *Q. Appl. Math.* **1944**, *2*, 164–168.

[422] Marquardt, D. *J. Soc. Ind. Appl. Math.* **1963**, *11*, 431–441.

[423] Conformer-Rotamer Ensemble Sampling Tool based on the xtb Semiempirical Extended Tight-Binding Program Package `crest`. https://github.com/grimme-lab/crest, Accessed: 2021-4-14.

[424] Pearson, K. *Philos. Mag.* **1901**, *2*, 559–572.

[425] Hotelling, H. *J. Educ. Psychol.* **1933**, *24*, 417–441.

[426] Lloyd, S. *IEEE Trans. Inf. Theorie* **1982**, *28*, 129–137.

[427] Shao, J.; Tanner, S. W.; Thompson, N.; Cheatham, T. E. *J. Chem. Theory Comput.* **2007**, *3*, 2312–2334.

[428] Kästner, J. *WIREs Comput. Mol. Sci.* **2011**, *1*, 932–942.

[429] Wales, D. J.; Doye, J. P. K. *J. Phys. Chem. A* **1997**, *101*, 5111–5116.

[430] Wales, D. J.; Scheraga, H. A. *Science* **1999**, *285*, 1368–1372.

[431] Semiempirical Extended Tight-Binding Program Package `xtb`. https://github.com/grimme-lab/xtb, Accessed: 2021-4-14.

[432] Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165–169.

[433] Furche, F.; Ahlrichs, R.; Hättig, C.; Klopper, W.; Sierka, M.; Weigend, F. *WIREs Comput. Mol. Sci.* **2014**, *4*, 91–100.

[434] Linstrom, E. P.; Mallard, W. NIST Chemistry WebBook, NIST Standard Reference Database Number 69. https://webbook.nist.gov/chemistry/, accessed December 18, 2020.

[435] Frenkel, M. Thermodynamics of Organic Compounds in the Gas State. TRC Data Series; Thermodynamics Research Center, 1994; p. 395, p. 460.

[436] Bootsma, S., Andrea N.; Wheeler Popular Integration Grids Can Result in Large Errors in DFT-Computed Free Energies. 2019; Preprint. https://doi.org/10.26434/chemrxiv.8864204.v5.

[437] Lüttschwager, N. O. B.; Wassermann, T. N.; Mata, R. A.; Suhm, M. A. *Angew. Chem. Int. Ed.* **2013**, *52*, 463–466.

[438] Byrd, J. N.; Bartlett, R. J.; Montgomery, J. A. *J. Phys. Chem. A* **2014**, *118*, 1706–1712.

[439] Vansteenkiste, P.; Van Speybroeck, V.; Marin, G. B.; Waroquier, M. *J. Phys. Chem. A* **2003**, *107*, 3139–3145.

[440] Scott, D. W. *J. Chem. Phys.* **1974**, *60*, 3144–3165.

[441] Rossini, F. D.; American Petroleum Institute Research Project 44, *Selected Values of Properties of Hydrocarbons and Related Compounds*; Thermodynamics Research Center, Texas Engineering Experiment Station, Texas A & M University, 1980.

[442] Mizuno, N.; Misono, M. *Chem. Rev.* **1998**, *98*, 199–218.

[443] De Moor, B. A.; Reyniers, M.-F. c.; Gobin, O. C.; Lercher, J. A.; Marin, G. B. *J. Phys. Chem. C* **2011**, *115*, 1204–1219.

[444] Spicher, S.; Bursch, M.; Grimme, S. *J. Phys. Chem. C* **2020**, *124*, 27529–27541.

[445] Fráter, G.; Schröder, F. *J. Org. Chem.* **2007**, *72*, 1112–1120.

[446] Brill, Z. G.; Condakes, M. L.; Ting, C. P.; Maimone, T. J. *Chem. Rev.* **2017**, *117*, 11753–11795.

[447] Lipp, A.; Selt, M.; Ferenc, D.; Schollmeyer, D.; Waldvogel, S. R.; Opatz, T. *Org. Lett.* **2019**, *21*, 1828–1831.

[448] Grubbs, R.; Tumas, W. *Science* **1989**, *243*, 907–915.

[449] Astruc, D. *New J. Chem.* **2005**, *29*, 42–56.

[450] Sure, R.; Grimme, S. *J. Chem. Theory Comput.* **2015**, *11*, 3785–3801.

[451] Mock, W. L.; Shih, N. Y. *J. Am. Chem. Soc.* **1989**, *111*, 2697–2699.

[452] Zhang, S.; Grimm, L.; Miskolczy, Z.; Biczók, L.; Biedermann, F.; Nau, W. M. *Chem. Commun.* **2019**, *55*, 14131–14134.

[453] Kolář, M.; Fanfrlík, J.; Lepšík, M.; Forti, F.; Luque, F. J.; Hobza, P. *J. Phys. Chem. B* **2013**, *117*, 5950–5962.

[454] Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J. C. W. *J. Phys. Chem. A* **1998**, *102*, 5074–5085.

[455] Schrader, B. *Infrared and Raman Spectroscopy: Methods and Applications*; VCH: New York, 2018.

[456] McCarthy, M.; Lee, K. L. K. *J. Phys. Chem. A* **2020**, *124*, 3002–3017.

[457] Pople, J. A.; Schlegel, H. B.; Krishnan, R.; Defrees, D. J.; Binkley, J. S.; Frisch, M. J.; Whiteside, R. A.; Hout, R. F.; Hehre, W. J. *Int. J. Quantum Chem.* **1981**, *20*, 269–278.

[458] Fogarasi, G.; Pulay, P. *Annu. Rev. Phys. Chem.* **1984**, *35*, 191–213.

[459] Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 553–586.

[460] Nevins, N.; Allinger, N. L. *J. Comput. Chem.* **1996**, *17*, 730–746.

[461] Stratmann, R. E.; Burant, J. C.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1997**, *106*, 10175–10183.

[462] Witek, H. A.; Morokuma, K. *J. Comput. Chem.* **2004**, *25*, 1858–1864.

[463] Fekete, Z. A.; Hoffmann, E. A.; Körtvélyesi, T.; Penke, B. *Mol. Phys.* **2007**, *105*, 2597–2605.

[464] Henschel, H.; Andersson, A. T.; Jespers, W.; Mehdi Ghahremanpour, M.; van der Spoel, D. *J. Chem. Theory Comput.* **2020**, *16*, 3307–3315.

[465] Blom, C.; Otto, L.; Altona, C. *Mol. Phys.* **1976**, *32*, 1137–1149.

[466] Blom, C.; Altona, C. *Mol. Phys.* **1976**, *31*, 1377–1391.

[467] Blom, C.; Altona, C. *Mol. Phys.* **1977**, *34*, 177–192.

[468] Pulay, P.; Fogarasi, G.; Pongor, G.; Boggs, J. E.; Vargha, A. *J. Am. Chem. Soc.* **1983**, *105*, 7037–7047.

[469] Fogarasi, G.; Zhou, X.; Taylor, P. W.; Pulay, P. *J. Am. Chem. Soc.* **1992**, *114*, 8191–8201.

[470] Laury, M. L.; Carlson, M. J.; Wilson, A. K. *J. Comput. Chem.* **2012**, *33*, 2380–2387.

[471] Gastegger, M.; Behler, J.; Marquetand, P. *Chem. Sci.* **2017**, *8*, 6924–6935.

[472] Lam, J.; Abdul-Al, S.; Allouche, A.-R. *J. Chem. Theory Comput.* **2020**, *16*, 1681–1689.

[473] Irikura, K. K. *Chem. Phys. Lett.* **2005**, *403*, 275–279.

[474] Rauhut, G.; Pulay, P. *J. Phys. Chem.* **1995**, *99*, 3093–3100.

[475] Katsyuba, S. A.; Zvereva, E. E.; Grimme, S. *J. Phys. Chem. A* **2019**, *123*, 3802–3808.

[476] Katsyuba, S. A.; Spicher, S.; Gerasimova, T. P.; Grimme, S. *J. Phsy. Chem. B* **2020**, *124*, 6664–6670.

[477] Penchev, P. N.; Sohou, A. N.; Andreev, G. N. *Spectroscopy Letters* **1996**, *29*, 1513–1522.

[478] Gussoni, M.; Castiglioni, C.; Ramos, M.; Rui, M.; Zerbi, G. *J. Mol. Struc.* **1990**, *224*, 445–470.

[479] Barone, V.; Biczysko, M.; Bloino, J. *Phys. Chem. Chem. Phys.* **2014**, *16*, 1759–1787.

[480] Porezag, D.; Pederson, M. R. *Phys. Rev. B* **1996**, *54*, 7830–7836.

[481] Galabov, B. S.; Dudev, T. *Vibrational Intensities*; Elsevier: Amsterdam, 1996.

[482] Fan, L.; Ziegler, T. *J. Chem. Phys.* **1992**, *96*, 9005–9012.

[483] Schwenke, D. W.; Partridge, H. *J. Chem. Phys.* **2000**, *113*, 6592–6597.

[484] Neugebauer, J.; Reiher, M.; Kind, C.; Hess, B. A. *J. Comput. Chem.* **2002**, *23*, 895–910.

[485] Luber, S.; Neugebauer, J.; Reiher, M. *J. Chem. Phys.* **2009**, *130*, 064105.

[486] Kiewisch, K.; Neugebauer, J.; Reiher, M. *J. Chem. Phys.* **2008**, *129*, 204103.

[487] Jacob, C. R.; Reiher, M. *J. Chem. Phys.* **2009**, *130*, 084106.

[488] Chu, P. M.; Guenther, F. R.; Rhoderick, G. C.; Lafferty, W. J. *J. Res. Natl. Inst. Stand. Technol.* **1999**, *104*, 59–81.

[489] Baumann, K.; Clerc, J. *Anal. Chim. Acta* **1997**, *348*, 327–343.

[490] Tan, X.; Chen, X.; Song, S. *J. Raman Spectrosc.* **2017**, *48*, 113–118.

[491] Vrancic, C.; Petrich, W. *J. Phys. Chem. A* **2011**, *115*, 12373–12379.

[492] Zapata, F.; García-Ruiz, C. *Spectrochim. Acta A* **2018**, *189*, 535–542.

[493] Balasubramani, S. G. et al. *J. Chem. Phys.* **2020**, *152*, 184107.

[494] Kruse, H.; Goerigk, L.; Grimme, S. *J. Org. Chem.* **2012**, *77*, 10824–10834.

[495] Neese, F. *WIREs Comput. Mol. Sci.* **2018**, *8*, e1327.

[496] Neese, F.; Wennmohs, F.; Hansen, A.; Becker, U. *Chem. Phys.* **2009**, *356*, 98–109.

[497] Alecu, I. M.; Zheng, J.; Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2010**, *6*, 2872–2887.

[498] Yamaguchi, Y.; Frisch, M.; Gaw, J.; Schaefer, H. F.; Binkley, J. S. *J. Chem. Phys.* **1986**, *84*, 2262–2278.

[499] Thomas, J. R.; DeLeeuw, B. J.; Vacek, G.; Crawford, T. D.; Yamaguchi, Y.; Schaefer, H. F. *J. Chem. Phys.* **1993**, *99*, 403–416.

[500] Jensen, F. *J. Chem. Phys.* **2003**, *118*, 2459–2463.

[501] Hait, D.; Head-Gordon, M. *J. Chem. Theory Comput.* **2018**, *14*, 1969–1981.

[502] Sinha, P.; Boesch, S. E.; Gu, C.; Wheeler, R. A.; Wilson, A. K. *J. Phys. Chem. A* **2004**, *108*, 9213–9217.

[503] Andersson, M. P.; Uvdal, P. *J. Phys. Chem. A* **2005**, *109*, 2937–2941.

[504] Chan, B. *J. Chem. Theory Comput.* **2017**, *13*, 6052–6060.

[505] Rozanska, X.; Stewart, J. J. P.; Ungerer, P.; Leblanc, B.; Freeman, C.; Saxe, P.; Wimmer, E. *J. Chem. Eng. Data* **2014**, *59*, 3136–3143.

[506] Savitzky, A.; Golay, M. J. E. *Anal. Chem.* **1964**, *36*, 1627–1639.

[507] Limbach, H.; Hennig, J.; Stulz, J. *J. Chem. Phys.* **1983**, *78*, 5432–5436.

[508] Kozlowski, P. M.; Jarzecki, A. A.; Pulay, P. *J. Phys. Chem.* **1996**, *100*, 7007–7013.

[509] Picard, R. R.; Cook, R. D. *J. Am. Stat. Assoc.* **1984**, *79*, 575–583.

[510] Klaeboe, P. *Vibrational Spectroscopy* **1995**, *9*, 3–17.

[511] Morel, P.; Schaffer, P.; Britten, J. F.; Valliant, J. F. *Acta Crystal. C* **2002**, *58*, m601–m604.

[512] Kingsbury, J. S.; Harrity, J. P. A.; Bonitatebus, P. J.; Hoveyda, A. H. *J. Am. Chem. Soc.* **1999**, *121*, 791–799.

[513] Manallack, D. T. *Perspect. Medicin. Chem.* **2007**, *1*, 1177391X0700100003.

[514] D3R, SAMPL6 homepage. https://drugdesigndata.org/about/sampl6, 2018.

[515] Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Bayly, C. I.; Cooper, M. D.; Pande, V. S. *J. Med. Chem.* **2008**, *51*, 769–779.

[516] Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Dill, K. A. *J. Phys. Chem. B* **2009**, *113*, 4533–4537.

[517] Klimovich, P. V.; Mobley, D. L. *J. Comput.–Aided Mol. Des.* **2010**, *724*, 307–316.

[518] Geballe, M. T.; Skillman, A. G.; Nicholls, A.; Guthrie, J. P.; Taylor, P. J. *J. Comput.–Aided Mol. Des.* **2010**, *24*, 259–279.

[519] Geballe, M. T.; Guthrie, J. P. *J. Comput.–Aided Mol. Des.* **2012**, *26*, 489–496.

[520] Mobley, D. L.; Wymer, K. L.; Lim, N. M.; Guthrie, J. P. *J. Comput.–Aided Mol. Des.* **2014**, *28*, 135–150.

[521] Bannan, C. C.; Burley, K. H.; Chiu, M.; Shirts, M. R.; Gilson, M. K.; Mobley, D. L. *J. Comput.–Aided Mol. Des.* **2016**, *30*, 927–944.

[522] Darvey, I. G. *Biochem. Educ.* **1995**, *23*, 80–82.

[523] Bodner, G. M. *J. Chem. Educ.* **1986**, *63*, 246.

[524] Murray, R. *Anal. Chem.* **1995**, *67*, 462a–462a.

[525] Pliego, J. R. *Chem. Phys. Lett.* **2003**, *367*, 145–149.

[526] da Silva, C. O.; da Silva, E. C.; Nascimento, M. A. C. *J. Phys. Chem. A* **1999**, *103*, 11194–11199.

[527] Dissanayake, D. P.; Senthilnithy, R. *J. Mol. Struc. THEOCHEM* **2009**, *910*, 93–98.

[528] Ripin, D.; Evans, D. Evans pKa Table. http://evans.rc.fas.harvard.edu/pdf/evans_pKa_table.pdf.

[529] Eckert, F.; Klamt, A. *AIChE Journal* **2002**, *48*, 369–385.

[530] Rappoport, D.; Furche, F. *J. Chem. Phys.* **2010**, *133*, 134105.

[531] Neese, F. *WIREs Comput. Mol. Sci.* **2012**, *2*, 73–78.

[532] Udvarhelyi, A.; Rodde, S.; Wilcken, R. *J. Comput.–Aided Mol. Des.* **2021**, *35*, 399–415.

[533] Molecular Networks GmbH, Nuremberg, Germany, CORINA v3.6. https://www.mn-an.com/, 2016.

[534] Schrödinger, LLC, New York, NY, Schrödinger Release 2017-3: MacroModel. 2017.

[535] Chang, G.; Guida, W. C.; Still, W. C. *J. Am. Chem. Soc.* **1989**, *111*, 4379–4386.

[536] Saunders, M.; Houk, K. N.; Wu, Y. D.; Still, W. C.; Lipton, M.; Chang, G.; Guida, W. C. *J. Am. Chem. Soc.* **1990**, *112*, 1419–1427.

[537] Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.

[538] Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571–2577.

[539] Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829–5835.

[540] Klamt, A.; Schüürmann, G. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799–805.

[541] Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 5656–5667.

[542] Bochevarov, A. D.; Watson, M. A.; Greenwood, J. R.; Philipp, D. M. *J. Chem. Theory Comput.* **2016**, *12*, 6001–6019.

[543] Eckert, F.; Klamt, A. *J. Comput. Chem.* **2006**, *27*, 11–19.

[544] Andersson, M. P.; Jensen, J. H.; Stipp, S. L. S. *PeerJ* **2013**, *1*, e198.

[545] Muckerman, J. T.; Skone, J. H.; Ning, M.; Wasada-Tsutsui, Y. *Biochim. Biophys. Acta, Bioenerg.* **2013**, *1827*, 882–891.

[546] Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009**, *113*, 6378–6396.

[547] Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009**, *113*, 4538–4543.

[548] Ribeiro, R. F.; Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Comput.–Aided Mol. Des.* **2010**, *24*, 317–333.

[549] Molecular Discovery Ltd., Borehamwood, United Kingdom, MoKa v2.5.4. https://www.mn-an.com/.

[550] Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029–5036.

[551] Milletti, F.; Storchi, L.; Sforna, G.; Cruciani, G. *J. Chem. Inf. Model.* **2007**, *47*, 2172–2181.

[552] Gedeck, P.; Lu, Y.; Skolnik, S.; Rodde, S.; Dollinger, G.; Jia, W.; Berellini, G.; Vianello, R.; Faller, B.; Lombardo, F. *J. Chem. Inf. Model.* **2015**, *55*, 1449–1459.

[553] Ho, J. *Aust. J. Chem.* **2014**, *67*, 1441.

[554] Seybold, P. G.; Shields, G. C. *WIREs Comput. Mol. Sci.* **2015**, *5*, 290–297.

[555] Mansouri, K.; Cariello, N. F.; Korotcov, A.; Tkachenko, V.; Grulke, C. M.; Catherine S. Sprankle, D. A.; Casey, W. M.; Kleinstreuer, N. C.; Williams, A. J. *J. Cheminform.* **2019**, *11*, 60.

[556] Yang, Q.; Li, Y.; Yang, J.-D.; Liu, Y.; Zhang, L.; Luo, S.; Cheng, J.-P. *Angew. Chem. Int. Ed.* **2020**, *59*, 19282–19291.

[557] Hunt, P.; Hosseini-Gerami, L.; Chrien, T.; Plante, J.; Ponting, D. J.; Segall, M. *J. Chem. Inf. Model.* **2020**, *60*, 2989–2997.

[558] Manchester, J.; Walkup, G.; Rivin, O.; You, Z. *J. Chem. Inf. Model.* **2010**, *50*, 565–571.

[559] Balogh, G. T.; Tarcsay, A.; Keserü, G. M. *J. Pharm. Biomed. Anal.* **2012**, *67-68*, 63–70.

[560] Jensen, J. H.; Swain, C. J.; Olsen, L. *J. Phys. Chem. A* **2017**, *121*, 699–707.

[561] Haworth, N. L.; Wang, Q.; Coote, M. L. *J. Phys. Chem. A* **2017**, *121*, 5217–5225.

[562] Grüber, C.; Buß, V. *Chemosphere* **1989**, *19*, 1595–1609.

[563] Soriano, E.; Cerdán, S.; Ballesteros, P. *J. Mol. Struct. THEOCHEM* **2004**, *684*, 121–128.

[564] Gross, K. C.; Seybold, P. G.; Hadad, C. M. *Int. J. Quantum Chem.* **2002**, *90*, 445–458.

[565] Svobodová Vařeková, R.; Geidl, S.; Ionescu, C.-M.; Skřehota, O.; Kudera, M.; Sehnal, D.; Bouchal, T.; Abagyan, R.; Huber, H. J.; Koča, J. *J. Chem. Inf. Model.* **2011**, *51*, 1795–1806.

[566] Ugur, I.; Marion, A.; Parant, S.; Jensen, J. H.; Monard, G. *J. Chem. Inf. Model.* **2014**, *54*, 2200–2213.

[567] Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999–3094.

[568] Schmidt am Busch, M.; Knapp, E.-W. *ChemPhysChem* **2004**, *5*, 1513–1522.

[569] Liptak, M. D.; Shields, G. C. *J. Am. Chem. Soc.* **2001**, *123*, 7314–7319.

[570] Zhang, S.; Baker, J.; Pulay, P. *J. Phys. Chem. A* **2010**, *114*, 425–431.

[571] Mirzaei, S.; Ivanov, M. V.; Timerghazin, Q. K. *J. Phys. Chem. A* **2019**, *123*, 9498–9504.

[572] Sakti, A. W.; Nishimura, Y.; Nakai, H. *J. Chem. Theory Comput.* **2018**, *14*, 351–356.

[573] Sure, R.; el Mahdali, M.; Plajer, A.; Deglmann, P. *J. Comput.-Aided Mol. Des.* **2021**,

[574] Thapa, B.; Schlegel, H. B. *J. Phys. Chem. A* **2015**, *119*, 5134–5144.

[575] Thapa, B.; Schlegel, H. B. *J. Phys. Chem. A* **2016**, *120*, 8916–8922.

[576] Thapa, B.; Schlegel, H. B. *J. Phys. Chem. A* **2016**, *120*, 5726–5735.

[577] Thapa, B.; Raghavachari, K. *J. Chem. Theory Comput.* **2019**, *15*, 6025–6035.

[578] ACD/LABS - Acid Dissociation (pKa) Calculation with ACD/pKa. https://www.acdlabs.com/products/percepta/predictors/pka/, Accessed: 2021-2-11.

[579] ChemAxon pKa plugin. hhttps://docs.chemaxon.com/display/docs/pka-plugin.md, Accessed: 2021-2-11.

[580] Fraczkiewicz, R.; Lobell, M.; Göller, A. H.; Krenz, U.; Schoenneis, R.; Clark, R. D.; Hillisch, A. *J. Chem. Inf. Model.* **2015**, *55*, 389–397.

[581] Greenwood, J. R.; Calkins, D.; Sullivan, A. P.; Shelley, J. C. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 591–604.

[582] Zhou, H.; Szabo, A. *J. Chem. Phys.* **1995**, *103*, 3481–3494.

[583] Mardirossian, N.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2014**, *16*, 9904–9924.

[584] COSMOtherm, Release 19. COSMOlogic GmbH & Co. KG, http://www.cosmologic.de, Accessed: 2021-4-6.

[585] Commandline ENergetic SOrting of Conformer Rotamer Ensembles `censo`. https://github.com/grimme-lab/censo, Accessed: 2021-4-14.

[586] O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *J. Cheminform.* **2011**, *3*, 33.

[587] PubChem entry 3-mercaptopropionic acid, `https://pubchem.ncbi.nlm.nih.gov/compound/6514`, Accessed: 2021-4-5.

[588] Webb, H.; Tynan-Connolly, B. M.; Lee, G. M.; Farrell, D.; O'Meara, F.; Sondergaard, C. R.; Teilum, K.; Hewage, C.; McIntosh, L. P.; Nielsen, J. E. *Proteins* **2011**, *79*, 685–702.

[589] Vuckovic, S.; Burke, K. *J. Phys. Chem. Lett.* **2020**, *11*, 9957–9964.

[590] Dohm, S.; Bursch, M.; Hansen, A.; Grimme, S. *J. Chem. Theory Comput.* **2020**, *16*, 2002–2012.

[591] Spicher, S.; Abdullin, D.; Grimme, S.; Schiemann, O. *Phys. Chem. Chem. Phys.* **2020**, *22*, 24282–24290.

[592] Young, T. A.; Silcock, J. J.; Sterling, A. J.; Duarte, F. *Angew. Chem. Int. Ed.* **2021**, *60*, 4266–4274.

[593] Funk, P.; Richrath, R. B.; Bohle, F.; Grimme, S.; Gansäuer, A. *Angew. Chem. Int. Ed.* **2021**, *60*, 5482–5488.

[594] Ruhl, J.; Ahles, S.; Strauss, M. A.; Leonhardt, C. M.; Wegner, H. A. *Org. Lett.* **2021**, *23*, 2089–2093.

[595] Bursch, M.; Kunze, L.; Vibhute, A. M.; Hansen, A.; Sureshan, K. M.; Jones, P. G.; Grimme, S.; Werz, D. B. *Chem. Eur. J.* **2021**, *27*, 4627–4639.

[596] Chan, L.; Hutchison, G. R.; Morris, G. M. *J. Chem. Inf. Model.* **2021**, *61*, 743–755.

[597] Strauss, M. A.; Wegner, H. A. *Angew. Chem. Int. Ed.* **2021**, *60*, 779–786.

[598] Zhang, Y.; Ye, J.; Liu, Z.; Liu, Q.; Guo, X.; Dang, Y.; Zhang, J.; Wei, Z.; Wang, Z.; Wang, Z.; Dong, H.; Hu, W. *J. Mater. Chem. C* **2020**, *8*, 10868–10879.

[599] Kirschbaum, C.; Greis, K.; Mucha, E.; Kain, L.; Deng, S.; Zappe, A.; Gewinner, S.; Schöllkopf, W.; von Helden, G.; Meijer, G.; Savage, P. B.; Marianski, M.; Teyton, L.; Pagel, K. *Nat. Comm.* **2021**, *12*, 1201.

[600] de Wergifosse, M.; Seibert, J.; Grimme, S. *J. Chem. Phys.* **2020**, *153*, 084116.

[601] Seibert, J.; Champagne, B.; Grimme, S.; de Wergifosse, M. *J. Phys. Chem. B* **2020**, *124*, 2568–2578.

Part VI.

Appendix

A1. Supporting Information to Chapter 2

Appendix A1 contains:

- Outline of energy terms in the GFNn–xTB schemes
- Equations for thermostatistical contributions

Energy Terms of GFNn–xTB

The following equations were taken from Ref. 36. The general GFNn–xTB energy terms are:

$$E_{GFN1-xTB} = E_{rep} + E_{disp}^{D3} + E_{XB} + E_{EHT} + E_{\gamma} + E_{\Gamma} + G_{Fermi} \tag{A1.1}$$

$$E_{GFN2-xTB} = E_{rep} + E_{EHT} + E_{disp}^{D4'} + E_{\gamma} + E_{AES} + E_{AXC} + E_{\Gamma} + G_{Fermi} \tag{A1.2}$$

$$E_{GFN0-xTB} = E_{rep} + E_{disp}^{D4} + E_{SRB} + E_{EEQ} + E_{EHT} + G_{Fermi} \tag{A1.3}$$

Empirical (fitted) parameters will be pointed out, but no detailed information about the terms origin is given, which instead can be found in the original literature.[36–39] Also, all auxiliary functions such as damping functions or formulations for coordination numbers (CNs) will be omitted here.

Common GFNn–xTB Ingredients

The extended Hückel term is calculated in all GFNn–xTB schemes as

$$E_{EHT} = \sum_{\mu\nu} P_{\mu\nu} H_{\nu\mu}^{EHT} \ , \tag{A1.4}$$

where the Hamiltonian elements are constructed *via*

$$H_{\mu\nu}^{EHT} = \frac{1}{2} \, K_{AB}^{ll'} \, S_{\mu\nu} \, (H_{\mu\mu} + H_{\nu\nu}) \ X(EN_A, EN_B) \, \Pi(R_{AB}, l, l') \, Y(\zeta_l^A, \zeta_{l'}^B) \tag{A1.5}$$

Herein the indices μ/ν indicate AOs, A/B are atoms and l/l' indicate shells. Furthermore, $\mu \in l(A)$ and $\nu \in l'(B)$. $K_{AB}^{ll'}$ is a shell-specific scaling constant, $X(EN_A, EN_B)$ is electronegativity dependent function, and $Y(\zeta_l^A, \zeta_{l'}^B)$ is a shell-exponent dependent term that is only present for GFN2–xTB. The distance-dependent polynomial scaling function $\Pi(R_{AB}, l, l')$ is

$$\Pi(R_{AB}, l, l') = \left(1 + k_{A,l}^{poly} \left(\frac{R_{AB}}{R_{cov,AB}}\right)^{\frac{1}{2}}\right) \left(1 + k_{B,l'}^{poly} \left(\frac{R_{AB}}{R_{cov,AB}}\right)^{\frac{1}{2}}\right) \tag{A1.6}$$

with the summed covalent radii $R_{cov,AB}$ and the fitted parameters $k_{A,l}^{poly}$.

The other term included in all GFNn–xTB methods is

$$E_{rep} = \frac{1}{2}\sum_{A,B}\frac{Z_A^{eff}Z_B^{eff}}{R_{AB}}e^{-\sqrt{\alpha_A\alpha_B}(R_{AB})^{k_f}} \ , \tag{A1.7}$$

where Z^{eff} and α are element-specific parameters and k_f is a global parameter. This term is related to the zeroth order potential used in other DFTB schemes,[20] but employs no pair-wise parameters.

Energies in GFN1–xTB

The EHT Hamiltonian components $H_{\mu\mu}$ and $X(EN_A, EN_B)$ in GFN1–xTB are

$$H_{\mu\mu}^{GFN1} = h_A^l\,(1 + k_{CN,l}CN_A) \tag{A1.8}$$

with the element-specific parameter h_A^l and the global angular momentum-specific parameter $k_{CN,l}$, and

$$X(EN_A, EN_B) = \left(1 + k_{EN}\Delta EN_{AB}^2\right) \ , \tag{A1.9}$$

with the global parameter k_{EN}. CN_A is the coordination number of atom A, calculated by the well-known D3 coordination number formula.[172,173]

The second order isotropic ES/XC tight-binding energy in GFN1–xTB is similar to the one employed in DFTB[20], but shell-dependent

$$E_\gamma^{GFN1} = \frac{1}{2}\sum_{A,B}\sum_{l\in A}\sum_{l'\in B} q_l\, q_{l'}\, \gamma_{AB,ll'} \ , \tag{A1.10}$$

where $\gamma_{AB,ll'}$ is a Mataga–Nishimoto–Ohno–Klopman damping function for the Coulomb interaction, employing several global and element-specific parameters.

At third order the GFN1–xTB energy is given as

$$E_\Gamma^{GFN1} = \frac{1}{3}\sum_A (q_A)^3\Gamma_A \ , \tag{A1.11}$$

i.e., it is (as the second order term) similar to DFTB, but modified to include only the on-site interaction. Γ_A is an element-specific parameter.

The remaining two terms in GFN1–xTB are the D3 dispersion energy (*cf.* Eq. 2.34) and a

correction for halogen bonds (XB)

$$E_{XB}^{GFN1} = \sum_{AB}^{N_{XB}} f_{damp}^{AXB} k_X \left[\left(\frac{k_{XR} R_{cov,AX}}{R_{AX}} \right)^{12} - k_{X2} \left(\frac{k_{XR} R_{cov,AX}}{R_{AX}} \right)^{6} \right] \left[\left(\frac{k_{XR} R_{cov,AX}}{R_{AX}} \right)^{12} - 1 \right]^{-1} \tag{A1.12}$$

with the global parameters k_{XR} and k_{X2}, the halogen-specific parameter k_X and the angular three-body damping function f_{damp}^{AXB}.

Energies in GFN2–xTB

While the the electronegativitiy dependent scaling function $X(EN_A, EN_B)$ in GFN2–xTB has the same form as Eq. A1.9, the Hamiltonian elements are reformulated as

$$H_{\mu\mu}^{GFN2} = h_A^l - \delta h_{CN'_A}^l CN'_A \tag{A1.13}$$

where h_A^l and $\delta h_{CN'_A}^l$ are shell- and element-specific parameters and CN'_A is a modified[39] version of the D3 coordination number. Additionally, the AO exponent ζ dependent term

$$Y(\zeta_A, \zeta_B) = \left(\frac{2\sqrt{\zeta_l^A \zeta_{l'}^B}}{\zeta_l^A + \zeta_{l'}^B} \right)^{\frac{1}{2}} \tag{A1.14}$$

is present as an additional scaling function of the Hamiltonian elements.

At second order, the GFN2–xTB isotropic electrostatic and exchange-correlation (IES, IXC) energy is the same as Eq. A1.10, but slightly differs with regards to the employed damping function.[39] As a novelty, GFN2–xTB is the first TB scheme that also employs anisotropic electrostatic (AES) and exchange-correlation (AXC) terms, which are given by

$$\begin{aligned} E_{AES} =& E_{q\mu} + E_{q\Theta} + E_{\mu\mu} && \text{(A1.15)} \\ =& \frac{1}{2} \sum_{A,B} \{ f_3(R_{AB}) \left[q_A(\boldsymbol{\mu}_B^T \mathbf{R}_{BA}) + q_B(\boldsymbol{\mu}_A^T \mathbf{R}_{AB}) \right] \\ & + f_5(R_{AB}) [q_A \mathbf{R}_{AB}^T \boldsymbol{\Theta}_B \mathbf{R}_{AB} + q_B \mathbf{R}_{AB}^T \boldsymbol{\Theta}_A \mathbf{R}_{AB} \\ & - 3(\boldsymbol{\mu}_A^T \mathbf{R}_{AB})(\boldsymbol{\mu}_B^T \mathbf{R}_{AB}) + (\boldsymbol{\mu}_A^T \boldsymbol{\mu}_B) R_{AB}^2] \} && \text{(A1.16)} \\ E_{AXC} =& \sum_A \left(f_{XC}^{\mu_A} |\boldsymbol{\mu}_A|^2 + f_{XC}^{\Theta_A} ||\boldsymbol{\Theta}_A||^2 \right) . && \text{(A1.17)} \end{aligned}$$

These so-called cumulative atomic multipole moments (CAMM) describe the anisotropic interaction between the atomic charge q_A, the dipole moment μ_A and quadrupole moment Θ_A. $f_n(R_{AB})$ are distance dependent damping functions employing global parameters and $f_{XC}^{\mu_A}$ and $f_{XC}^{\Theta_A}$ are element-specific parameters.

Additionally at second order a self-consistent charge version of the D4 dispersion model[174,178]

is employed according to

$$E_{disp}^{D4'} = -\sum_{A>B}\sum_{n=6,8} s_n \frac{C_n^{AB}(q_A, CN_A, q_B, CN_B)}{R_{AB}^n} f_{damp,BJ}^{(n)}(R_{AB}) - s_9 \sum_{A>B>C} \frac{(3\cos(\theta_{ABC})\cos(\theta_{BCA})\cos(\theta_{CAB})+1)\, C_9^{ABC}(CN_{cov}^A, CN_{cov}^B, CN_{cov}^C)}{(R_{AB}R_{AC}R_{BC})^3} \times f_{damp,zero}^{(9)}(R_{AB}, R_{AC}, R_{BC}) \ , \tag{A1.18}$$

which depends on the atomic charges, coordination numbers and employs both Becke–Johnson $f_{damp,BJ}^{(n)}$ and zero damping $f_{damp,zero}^{(9)}$ functions.

The third order IES and IXC term in GFN2–xTB is closely related to the one of GFN1–xTB, but is shell-dependent and additionally depends on a global shell-specific parameter K_l^Γ

$$E_\Gamma^{GFN2} = \frac{1}{3}\sum_A \sum_{l\in A} (q_l)^3 \, K_l^\Gamma \, \Gamma_A \ . \tag{A1.19}$$

Due to the sophisticated AES/AXC description in GFN2–xTB no additional empirical XB correction is needed.

Energies in GFN0–xTB

GFN0–xTB is the most empirical GFNn–xTB scheme and employs no second and third order charge dependent terms. Instead, a charge dependence is built-in into the EHT Hamiltonian elements

$$H_{\mu\mu}^{GFN1} = h_A^l - \delta h_{mCN_A}^l \, mCN_A - \delta h_{q_A}^l \, q_A - \Gamma_{q_a}^l q_A^2 \ , \tag{A1.20}$$

where h_A^l and $\delta h_{mCN_A}^l$ are element-specific parameters as in GFN2–xTB, and $\delta h_{q_A}^l$ and $\Gamma_{q_A}^l$ are element-specific parameters related to the chemical hardness. mCN_A is a modified coordination number based on the D4 CN.[37] As a further modification to the EHT Hamiltonian in GFN0–xTB, the electronegativity dependent scaling function $X(EN_A, EN_B)$ is modified and made shell dependent according to

$$X^{ll'}(EN_A, EN_B) = 1 + k_{EN}^{ll'} \Delta EN_{AB}^2 + k_{EN}^{ll'} b_{EN} \Delta EN_{AB}^4 \ , \tag{A1.21}$$

with the shell-specific parameter $k_{EN}^{ll'}$ and the global parameter b_{EN}. The electronegativity is also used for the "short-range basis" correction[157,180]

$$E_{SRB} = k_{srb} \sum_{A,B} \exp\left[-\eta_{srb}\left(1 + g_{scal}\Delta EN_{AB}^2\right)\left(R_{AB} - R_{AB}^{srb}\right)\right] \ , \tag{A1.22}$$

where k_{srb}, η_{srb}, and g_{scal} are global fit parameters and the covalent bond radii R_{srb}^{AB} are modified by the electronegativities.[37] The dispersion energy is included *via* the standard D4 model.[174,178] This (and the Hamiltonian) requires atomic charges, which are not obtained from Mulliken populations but an classical charge equilibrium model (EEQ). From this also the zeroth order electrostatic energy is obtained according to

$$E_{EEQ} = \sum_A \left[\chi_A \, q_A + \frac{1}{2} \left(J_{AA} + \frac{2}{\sqrt{\pi}} \gamma_{AA} \right) \right] + \frac{1}{2} \sum_{A,B} q_A \, q_B \frac{\mathrm{erf}(\gamma_{AB} R_{AB})}{R_{AB}} \; . \tag{A1.23}$$

Here, J_{AA}, $\chi_A(EN_A, \kappa_A, mCN_A)$ are fitted element-specific parameters and γ_{AB} is related to the inverse root mean square of the atomic radii.[37]

Additional Equations for Thermostatistical Contributions

The following symbols are used in this section:

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

The formulas were taken from Refs. 4,215. Conformational contributions are discussed in Part III and are derived from the electronic partition function.

Partition Functions

$$Q_{elec} = \sum_i g_i e^{-\epsilon_i / kT} \tag{A1.24}$$

$$Q_{trans} = (2\pi MkT)^{\frac{3}{2}} \, h^{-3} V \tag{A1.25}$$

$$Q_{rot} = \frac{8\pi^2}{\sigma h^3} \, (2\pi kT)^{\frac{3}{2}} \, \sqrt{I_A \, I_B \, I_C} \tag{A1.26}$$

$$Q_{vib} = \prod_i^{modes} \frac{e^{-h\nu_i / 2kT}}{1 - e^{-h\nu_i / kT}} \tag{A1.27}$$

The Enthalpy

Note that the enthalpies given below refer to $[H(T) - H(0)]$. The translational and rotational enthalpy only depend on the temperature and the ideal gas constant

$$H_{trans} = \frac{5}{2}RT \tag{A1.28}$$

$$H_{rot} = \frac{3}{2}RT \tag{A1.29}$$

$$H_{rot}^{linear} = RT\ . \tag{A1.30}$$

The vibrational and electronic enthalpies given by

$$H_{vib} = RT \sum_{i}^{modes} \left(\frac{h\nu_i}{kT}\right) \frac{e^{-h\nu_i/kT}}{1 - e^{-h\nu_i/kT}} \tag{A1.31}$$

$$H_{elec} = RT \frac{\sum_i g_i(\epsilon_i/kT)e^{-\epsilon_i/kT}}{\sum_i g_i e^{-\epsilon_i/kT}}\ . \tag{A1.32}$$

The Entropy

The translational and rotational entropies are given by

$$S_{trans} = \frac{5}{2}R + R\ln\left(\frac{V}{N_A}\left(\frac{2\pi MkT}{h^2}\right)^{\frac{3}{2}}\right) \tag{A1.33}$$

$$S_{rot} = R\left[\frac{3}{2} + \ln\left(\frac{\sqrt{\pi}}{\sigma}\left(\frac{8\pi^2 kT}{h^2}\right)^{\frac{3}{2}}\sqrt{I_A I_B I_C}\right)\right] \tag{A1.34}$$

$$S_{rot}^{linear} = R\left[1 + \ln\left(\frac{8\pi^2 IkT}{\sigma h^2}\right)\right]\ . \tag{A1.35}$$

The vibrational and electronic entropies are

$$S_{vib} = R\sum_{i}^{modes}\left[\left(\frac{h\nu_i}{kT}\frac{e^{-h\nu_i/kT}}{1 - e^{-h\nu_i/kT}}\right) - \ln\left(1 - e^{-h\nu_i/kT}\right)\right] \tag{A1.36}$$

$$S_{elec} = R\ln\left(\sum_i g_i e^{-\epsilon_i/kT}\right) + R\frac{\sum_i g_i(\epsilon_i/kT)e^{-\epsilon_i/kT}}{\sum_i g_i e^{-\epsilon_i/kT}}\ . \tag{A1.37}$$

A2. Supporting Information to Chapter 3

Appendix A2 contains:

- List of important changes to the CREST program since publication
- Information about RMSD and B_e calculations
- Front cover of the associated publication

Important Changes to the CREST Code since Publication

- Revision of the rotational constant comparison. Instead of a fixed 15.0 MHz threshold, comparison are now made using a dynamical relative threshold of 1.0–2.5 %, based on the anisotropy of rotational constants in the ensemble.
- For GFNn–xTB the energy threshold (ETHR) between conformers was lowered to 0.05 kcal mol^{-1}. For GFN–FF the threshold is still 0.1 kcal mol^{-1}.
- The CREGEN sorting procedure now includes a topology comparison and the technical performances (computation times) of the comparisons were improved.
- The flexibility measure was revised and now includes a non-covalent component. See Appendix A3.
- Additional procedures for the conformational entropy (see Chapter 4) and pK_a values (see Chapter 7) were implemented.

Note on the Calculation of Atomic RMSDs

The atomic RMSDs are referred to above as

$$RMSD = \sqrt{\frac{\sum_i^N |\boldsymbol{x}_i - \boldsymbol{y}_i|^2}{N}} \ , \tag{A2.1}$$

where N is the number of atoms and $\boldsymbol{x}_i/\boldsymbol{y}_i$ are the spatial coordinate vectors for atom i in the two structures X and Y. However, the actual task of finding the RMSD is a problem of linear algebra, which results in some computational overhead. The goal is to find an orthogonal transformation $\mathbf{U}$ that minimizes a residual E (=the RMSD)

$$E := \frac{1}{N} \sum_i^N |\mathbf{U}\boldsymbol{x}_i + \boldsymbol{r} - \boldsymbol{y}_i|^2 \ , \tag{A2.2}$$

where $\boldsymbol{r}$ is a spatial translation. In CREST a quaternion algorithm is employed for this minimization problem. The mathematical formulation is quite lengthy and can be found in Ref. 302.

Note on the Calculation of Rotational Constants

The rotational constants B_e are, as the RMSD, a purely structure based quantity that can be used for structure comparisons. In contrast to the RMSD however, there is no dependence on the atomic order, allowing a distinction between conformers and rotamers (*i.e.*, all rotamers of a conformer have the same B_e). B_e herein refers to the equilibrium (*i.e.*, optimized) geometry and is derived from B. The rotational B constant will have three spatial components (for linear molecules only two), depending on three moments of inertia I_A, I_B, I_C, around the respective orthogonal principle axes of inertia A, B, and C. Both, moments and axes of inertia can be obtained from diagonalization of a 3×3 inertia matrix

$$\mathbf{I} = \begin{pmatrix} \sum_i m_i(y_i^2 + z_i^2) & -\sum_i m_i x_i y_i & -\sum_i m_i x_i z_i \\ -\sum_i m_i x_i y_i & \sum_i m_i(x_i^2 + z_i^2) & -\sum_i m_i y_i z_i \\ -\sum_i m_i x_i z_i & -\sum_i m_i y_i z_i & \sum_i m_i(x_i^2 + y_i^2) \end{pmatrix} , \tag{A2.3}$$

with the mass m_i and spatial coordinates (relative to the center of mass) $\{x_i, y_i, z_i\}$ for atom i. Upon diagonalization the moments of inertia are obtained as eigenvalues and the principle axes as eigenvectors of $\mathbf{I}$. The rotational constants B is then calculated *via*

$$B_n = \frac{h}{8\pi^2 c I_n} \qquad \forall n \in \{A, B, C\} , \tag{A2.4}$$

where h is Planck's constant and c is the constant speed of light.

Associated Publication Front Cover

Figure A2.1.: Front cover associated with the publication *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

A3. Supporting Information to Chapter 4

Appendix A3 contains:

- Additional technical details of the calculations
- Supplementary figures
- Tables with detailed results

Implementation, Algorithmic and Calculation Details

RMSD based metadynamics

The RMSD metadynamics (MTD) were introduced in Ref. 41 and are based on a bias potential

$$V_{bias} = \sum_i k_i \exp\left(-\alpha_i \Delta_i^2\right) \ , \tag{A3.1}$$

where Δ_i is the atomic RMSD[302] between a reference structure i and the calculated molecule. k_i and α_i empirical or automatically determined parameters that shape the potential. During a metadynamics simulation points on the simulated PES, *i.e.* snapshots of the MD simulation are saved for the calculation of Δ_i, which is then used to generate a repulsive V_{bias} contribution at the respective geometry. By a continuous collection and update of reference structures (from new snapshots) over the whole length of the simulation, V_{bias} will dynamically increase and form a history-dependent potential. This way previously found regions of the PES are blocked for the exploration and new conformers (PES minima) are found more safely.

As a new alternative we introduce another type of metadynamics, called static metadynamics (sMTD). In contrast to the MTD discussed in Refs. 33,41, this simulation is initialized with a given set of reference geometries and the MD will hence exhibit one global (and unchanged) V_{bias} potential. This version of MTD is more similar in nature to the well-known umbrella sampling or global optimization procedures. With regards to the PES sampling, sMTD has a less explorative character than MTD for finding the global minimum, but will more continuously expand the conformational ensemble with new higher-energetic structures.

Molecular flexibility description

Many settings for the here discussed workflow are generated automatically and based on the individual structure of the investigated molecule. An important parameter is the molecular

flexibility, because it is directly related to the molecules accessible low-energy space. In Ref. 33 we proposed a molecular flexibility measure $\xi_{\text{f,cov}}$, defined by

$$\xi_{\text{f,cov}} = \sqrt{\frac{1}{N_{\text{bonds}}}} \left(\sum_{i}^{N_{\text{bonds}}} \left(1 - \mathrm{e}^{-5(\mathbf{B}_{\text{AB}}-2)^{10}}\right)^2 \frac{4}{N_{\text{A}}^{\text{neigh}} N_{\text{B}}^{\text{neigh}}} \left(R_i^{(\text{f})}\right)^2 \right)^{\frac{1}{2}} . \tag{A3.2}$$

The summation over all *non*-terminal bonds i with the atoms A,B $\in i$ includes the Wiberg-Mayer bond order[330,331] (WBO) $\mathbf{B}_{\text{AB}}$ between the two atoms. It is always obtained from a GFN0-xTB calculation because no WBO is accessible from FF data. $N_{\text{A,B}}^{\text{neigh}}$ are the numbers of neighboring atoms of A and B, respectively. $R^{(\text{f})}$ is a predefined factor of value 1 if the bond i is not part of a ring and < 1 (depending on the ring size) otherwise. The measure $\xi_{\text{f,cov}}$ works well for assigning a quantitative covalent (as indicated by the subscript addendum) flexibility, where values close to unity indicate an highly flexible system and values $\ll 1$ indicate rigid systems. It fails, however, for systems that are stabilized by non-covalent interactions like hydrogen-bonds or dispersion. Reasonably sized organic molecules, such as polypeptides, are often much more rigid as described by $\xi_{\text{f,cov}}$, due to the formation of intramolecular hydrogen-bonding networks. Likewise, dispersion interactions are always present and might stabilize certain conformations, but do not contribute to the flexibility in Eq. A3.2. Therefore, a modified molecular flexibility is proposed that includes non-covalent contributions $\xi_{\text{f,NCI}}$ from hydrogen bonds and dispersion to the total molecular flexibility

$$\xi_{\text{f,tot}} = \frac{1}{2}\xi_{\text{f,cov}} + \frac{1}{2}\left(\xi_{\text{f,NCI}}\xi_{\text{f,cov}}^{\frac{1}{2}}\right) . \tag{A3.3}$$

Non-covalent interactions are quantified from the total hydrogen-bond energy E_{HB} and D4 dispersion energy E_{disp}, *relative* to the respective energies of a known reference system. In order to be comparable to the reference, the energies must be normalized to the number of atoms N in the system. The final formulation for $\xi_{\text{f,NCI}}$ then is given by

$$\xi_{\text{f,NCI}} = 1.0 - \frac{1}{2}\left(\frac{E_{\text{HB}}}{E_{\text{HB,ref}}} + \frac{E_{\text{disp}}}{E_{\text{disp,ref}}}\right)\left(\frac{N}{N_{\text{ref}}}\right)^{-1} . \tag{A3.4}$$

The energy contributions E_{HB} and E_{disp} are readily available from a simple GFN-FF singlepoint energy calculation. Respective reference contributions $E_{\text{HB,ref}}$ and $E_{\text{disp,ref}}$ are then assumed to be a calibration standard for all further computations of the molecular flexibility. As a reasonably flexible reference in which NCI interactions are important, the crambin protein was chosen for the calculation of $E_{\text{disp,ref}}$ and $E_{\text{HB,ref}}$.

Rotamer Numbers

One of the key assumptions in the proposed scheme is that every contributing conformer i can be effectively represented by a number of energetically degenerate rotamer structures with its

degeneracy number g_i. This number is composed of three parts

$$g_i = \frac{g_{rot}\, g_{core}}{g_{sym}} \ , \tag{A3.5}$$

where g_{rot} is a factor arising from single-bond rotations, g_{core} denotes a factor resulting from complex inversion and g_{sym} includes the molecular symmetry into g_i. Here, the factor g_{rot} is a constant that is the same for all *unique* conformers and (pseudo-)enantiomers. All conformers of a molecule have the same number of rotatable groups, each resulting in a fixed prefactor equal to the number of equivalent nuclei exchanged by the rotation (*i.e.*, 3 for methyl, 2 for phenyl, 5 for η^5-$C_5H_5^-$, and so forth). We assume this factor to be constant for a given molecules since all combinations of the rotations would be observed at some point in time ($t \to \infty$). The factor g_{core} results from more complicated inversion-type processes that are responsible for the generation of other degenerate structures such as (pseudo-)enantiomers of a conformer. g_{core} is unique for every conformer since it is linked to the molecular symmetry of the respective structure.

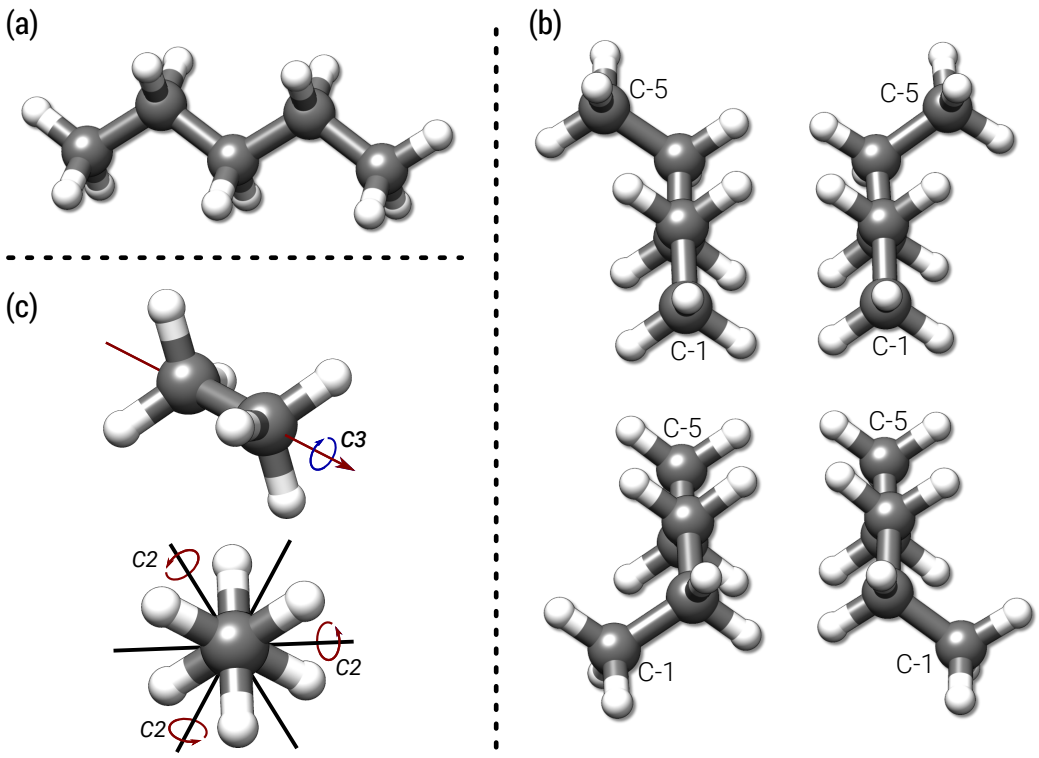

Figure A3.1.: Structures of ethane and *n*-pentane. a) Lowest conformer of *n*-pentane in the gas-phase. b) Second lowest conformer of *n*-pentane. The conformer has four different pseudo-enantiomers, for better distinguishability the first and fifth carbon atom are labeled as such. c) Main symmetry elements within the ecliptic (D_{3d}) ethane molecule.

As an example the two lowest conformers of the n-pentane molecule are shown in Fig. A3.1a

and A3.1b. With two terminal methyl groups n-pentane has a rotamer degeneration of $g_{rot} = 3^2$. The lowest conformer of n-pentane in the gas-phase has C_{2v} symmetry and no other enantiomeric structures exist ($g_{core} = 1$). In the second conformer, however, one of the terminal methyl groups is slightly twisted resulting in a total of four different (pseudo-)enantiomers ($g_{core} = 4$). Hence, 9 rotamers are to be expected for the lowest conformer of n-pentane, but there are 36 degenerate rotamers for the second conformer.

The rotamer number g_i also depends on the molecular symmetry. If symmetry operations exist that coincide with some of the rotations included in g_{rot} and can impose a nucleus on itself, g_i has to be reduced by a factor g_{sym}. A simple example is the ethane molecule as shown in Fig. A3.1c. With two terminal methyl groups one could expect 3^2 rotamers, but since ethane has D_{3d} symmetry (D_{3h} for the eclipsed form) there are only three different rotamers for the molecule. Here, g_{rot} equals the symmetry number of the primary rotation axis. Other examples are neopentane (T_d), isobutane (C_{3v}), or ferrocene (D_{5d}/D_{5h}). For most molecules g_{rot} simply is unity and is only important for high symmetry cases.

The rotamer number g_i is generated automatically from the CRE obtained by a conformational search as implemented in the CREST program and information of chemically equivalent nuclei. Nuclear equivalencies are obtained as a by-product of the conformational search directly from the structure comparison as described in Ref. 40. For the identification of rotational groups, the topology of the molecule is set up for the lowest energy conformer and analyzed. Herein, the topology can be either based on quantum chemical data (covalent bond orders) or just set up directly from the coordination numbers (CNs). Molecular rings are identified in a graph representation of the topology, using an custom depth-first all-pair-shortest-path algorithm. Rotational groups are obtained form groups of equivalent nuclei and must obey some simple heuristic rules:

- The equivalent nuclei must be connected to a common neighboring atom.
- The neighboring atom may have a maximum of one neighbor other than the equivalent nuclei to be considered "freely rotatable". (An alternative definition via the WBO is possible).
- Rotations from different groups of equivalent atoms in the *same* ring must only be counted once to avoid double counting.
- The rotation number of the group is equivalent to the number of its members (*i.e.*, the equivalent nuclei).

These rules work recursively (*e.g.*, a *tert*-butyl group results in 3^4 rotamers), but special consideration has to be paid to freely coordinated rings (*e.g.*, Cp in ferrocene), which will not be discussed here any further.

The factor g_{core} is generated from a Cartesian RMSD comparison[302] of all structures in a CRE that belong to the same conformer. In this comparison all atoms that are rotationally equivalent must be neglected in the RMSD, leading to the identification of all the different

"core" structures for each conformer, *i.e.*, its (pseudo-)enantiomers. For an example again see Fig. A3.1b. The key assumption for this is, that the conformational search was able to generate all relevant enantiomeric structures of a single conformer at least once.

Single Point Hessian Procedure

Within the described workflow and the calculation of the $\overline{S}_{msRRHO}$ population average, the entropy for a reference structure $S_{msRRHO,ref}$ has to be calculated. For consistency this reference term has to be calculated at the same level of theory as the population average in $\overline{S}_{msRRHO}$, that is GFN2-xTB or GFN-FF. A geometry optimization at this level might lead to an alteration of the frequencies and hence calculated entropy. On the other hand, if calculated directly for the DFT reference geometry, there is a high probability to observe imaginary modes because the DFT geometry will not necessarily be a minimum on the GFN2 or GFN-FF PES. To account for this problem in a "best of two worlds" approach we employ a new procedure called single point hessians (SPH). Details of the SPH approach will be published elsewhere,[420] but basically it works by applying an additive potential[33,41] similar to Eq. 3.3 above,

$$V_{SPH} = k \exp\left(-\alpha \Delta^2\right) \ , \tag{A3.6}$$

where Δ is the atomic RMSD[302] between two molecular structures, and k and α define the potential shape. Within the SPH procedure k and α are calculated automatically in an iterative process, by repeatedly calculating the RMSD between the DFT input structure and a GFNn-xTB or GFN-FF re-optimized structure and updating V_{bias}, until no change in the geometry is observed. This essentially reshapes the PES at GFN2-xTB (or GFN-FF) level and removes any imaginary modes for frequencies calculated directly for the DFT geometry. Entropies calculated with frequencies from SPH resemble those at the DFT level, but retain a slight level of theory dependent shift, which makes them compatible with $\overline{S}_{msRRHO}$. With regards to computational cost the SPH approach is much cheaper than calculating frequencies at the DFT level, but more expensive as standard GFN2-xTB or GFN-FF Hessian calculations.

Test Sets

The LBH Test Set

Table A3.1.: Absolute entropies for the LBH benchmark[380] set. Entropies are given for a combination of S_{msRRHO} entropy calculated at DFT (B97-3c or B3LYP-D3/def2-TZVP) and S_{conf} calculated at a lower (GFN2-xTB or GFN-FF) level. Mean deviation (MD), mean average deviation (MAD), root-mean-square deviation (RMSD) and standard deviation (SD) are given below. "plain" values correspond to msRRHO without S_{conf}. All values correspond to cal mol^{-1} K^{-1}.

			B97-3c			B3LYP-D3/TZ			UM-VT[a]
		experiment	plain	GFN-FF	GFN2-xTB	plain	GFN-FF	GFN2-xTB	
1	ethane	54.79	54.48	54.50	54.51	54.48	54.49	54.49	54.75
2	propane	64.61	64.32	64.36	64.52	64.28	64.28	64.44	64.57
3	n-butane	74.21	72.07	74.21	74.14	72.07	74.05	74.08	74.21
4	isobutane	70.63	70.22	70.25	70.35	70.20	70.21	70.25	70.22
5	n-pentane	83.55	79.71	83.76	83.53	79.74	83.72	83.83	83.56
6	isopentane	82.16	80.05	82.19	82.10	80.00	82.12	82.25	81.51
7	neopentane	73.14	72.86	72.87	73.02	73.01	72.97	73.03	73.18
8	n-hexane	92.94	87.39	93.12	93.18	87.52	93.30	93.39	93.49
9	2,2-dimethylbutane	85.66	85.91	85.95	86.07	85.70	85.81	85.64	84.97
10	2,3-dimethylbutane	87.46	85.70	88.08	87.19	85.32	87.66	87.04	85.17
11	2-methylpentane	91.06	87.68	91.48	91.57	87.72	91.27	91.01	90.56
12	3-methylpentane	91.54	87.86	91.00	90.84	87.88	90.71	90.25	90.54
13	n-heptane	102.32	94.97	102.61	102.99	95.19	102.79	102.60	102.88
14	2,2-dimethylpentane	93.86	93.06	94.37	94.73	93.13	94.45	94.39	92.71
15	2,3-dimethylpentane	99.11	94.44	98.63	97.68	94.20	98.11	97.10	94.72
16	2,4-dimethylpentane	94.89	91.86	95.39	95.46	92.13	95.48	95.54	92.8
17	3,3-dimethylpentane	95.20	90.55	102.11[b]	96.48	90.46	101.62[b]	96.21	93.66
18	3-ethylpentane	98.37	93.61	99.33	97.71	94.04	99.42	98.16	96.25
19	2-methylhexane	100.50	95.24	100.92	100.76	95.46	100.83	100.46	100.48
20	3-methylhexane	101.84	95.47	102.52	101.89	95.61	102.57	101.29	99.94
21	2,2,3-trimethylbutane	91.63	92.56	92.65	92.08	92.52	92.43	92.21	90.06
22	n-octane	111.70	102.52	112.00	112.18	102.87	112.07	112.04	113.51
23	1-butene	73.58	71.22	74.55	73.33	71.14	74.65	73.17	73.36
24	1,3-butadiene	66.63	66.34	67.57	66.88	66.22	67.49	66.78	66.05
25	ethyl methyl ether	73.91	72.79	75.28	75.26	72.78	74.49	75.14	73.07
26	ethanol	67.07	64.58	66.65	66.28	64.57	66.79	66.24	66.41
27	propionaldehyde	72.75	70.01	73.03	72.44	70.05	72.38	72.87	72.73
28	2-butanone	81.12	79.63	84.22	83.74	80.19	84.70	83.86	80.63
29	acetic acid	67.75	69.97	70.04	69.98	69.08	69.60	69.11	68.06
30	propylamine	77.78	72.80	76.85	76.33	72.91	76.90	76.40	76.63
31	1-nitropropane	83.80	81.68	84.03	85.50	81.79	84.13	86.00	84.68
32	1-fluoropropane	72.85	70.68	72.67	73.04	70.65	72.61	72.96	73.11
33	1-chloropropane	75.43	73.27	75.52	75.67	73.17	75.38	75.49	75.54
34	1-bromopropane	79.07	76.04	78.27	78.45	75.93	78.21	78.36	78.08
35	ethyl methyl sulfide	79.64	77.13	80.13	79.54	76.92	80.15	79.14	79.41
36	methyl disulfide	80.16	78.97	80.14	80.99	78.61	79.79	80.52	80.12
37	ethanethiol	70.79	68.21	70.49	70.52	68.22	70.54	70.16	71.01
38	ethylene glycol	72.61	69.89	73.60	74.54	69.84	73.55	74.51	74.56
39	acrylic acid	73.54	71.90	73.29	72.99	71.71	73.08	72.61	72.18
	MD	—	-2.62	0.32	0.23	-2.63	0.23	0.09	-0.52
	MAD	—	2.79	0.59	0.65	2.74	0.60	0.65	0.86
	RMSD	—	3.46	0.84	0.91	3.39	0.85	0.93	1.24
	SD	—	2.29	0.79	0.89	2.18	0.83	0.93	1.14

[a] Values taken from Ref. 380. [b] Outlier neglected from the statistics.

Table A3.2.: Absolute entropies for the AS23 benchmark set. Entropies are given for a combination of S_{msRRHO} entropy calculated at DFT (B97-3c or B3LYP-D3/def2-TZVP) and S_{conf} calculated at a lower (GFN2-xTB or GFN-FF) level. Mean deviation (MD), mean average deviation (MAD), root-mean-square deviation (RMSD) and standard deviation (SD) are given for the combined LBH+AS23 below. "plain" values correspond to msRRHO without S_{conf}. All values correspond to cal mol^{-1} K^{-1}.

			B97-3c			B3LYP-D3/TZ		
		experiment	plain	GFN-FF	GFN2-xTB	plain	GFN-FF	GFN2-xTB
40	cyclohexane	71.27	71.33	71.31	71.31	71.26	71.27	71.24
41	cycloheptane	81.82	80.21	80.48	80.24	80.18	81.15	80.15
42	cyclooctane	87.66	86.51	87.52	88.61	86.47	87.78	91.44
43	perfluorheptane	158.88	152.65	164.32[b]	159.90	152.51	165.84[b]	160.04
44	2,2,4,4-tetramethylpentane	103.13	101.30	102.10	102.76	101.06	102.07	102.41
45	2,2,3,4,4-pentamethylpentane	108.70	106.47	107.14	106.12	106.35	106.99	105.77
46	3,3-Diethyl-2-methylpentane	116.00	109.55	121.75[b]	115.81	109.32	121.87[b]	115.00
47	dipropylether	100.98	92.59	103.52	103.86	93.00	103.84	103.81
48	triethylamine	96.90	92.78	101.57	101.63	93.31	102.57	101.05
49	1-heptanol	114.83	102.72	113.61	113.83	102.99	114.33	113.74
50	Thiacycloheptane	86.50	83.62	87.95	87.43	83.54	87.66	87.65
51	nonane	121.06	109.97	121.30	121.06	110.40	121.46	121.44
52	decane	130.44	117.31	130.96	130.51	117.79	131.02	130.77
53	dodecane	148.78	131.60	149.42	149.21	132.09	149.81	149.11
54	butyl-propyl-sulfide	117.90	105.43	118.31	120.15	105.58	118.54	120.05
55	1-hexanol	105.50	94.92	104.11	104.29	95.13	104.61	104.22
56	1-pentanol	96.20	87.41	94.90	95.06	87.42	95.17	94.78
57	1-butanol	86.80	79.75	85.60	85.79	79.75	85.84	85.65
58	1-propanol	77.10	72.12	76.37	76.21	72.03	76.47	76.02
59	1-butanthiol	89.70	83.39	89.67	89.17	83.46	89.74	89.22
60	1-pentanthiol	99.30	90.98	99.00	98.45	91.17	99.18	98.53
61	1-hexanthiol	108.60	98.62	108.50	107.80	98.89	108.64	107.95
62	1-heptanthiol	117.90	106.18	118.00	117.09	106.61	118.21	116.96
	LBH+AS23 errors							
	MD	—	-4.36	0.21	0.15	-4.32	0.24	0.07
	MAD	—	4.48	0.73	0.83	4.40	0.73	0.92
	RMSD	—	5.90	1.09	1.19	5.77	1.16	1.29
	SD	—	4.00	1.08	1.19	3.85	1.15	1.30

[b] Outlier neglected from the statistics.

Linear Alkanes

Table A3.3.: Entropies calculated for linear alkanes up to octadecane. All values correspond to cal mol^{-1} K^{-1}.

alkane	carbon atoms	experiment	S_{msRRHO} B97-3c	S_{abs} B97-3c + GFN-FF	S_{abs} B97-3c + GFN2-xTB
ethane	2	54.79	54.48	54.50	54.51
propane	3	64.61	64.32	64.36	64.52
n-butane	4	74.21	72.07	74.21	74.14
n-pentane	5	83.55	79.71	83.76	83.53
n-hexane	6	92.94	87.39	93.12	93.18
n-heptane	7	102.32	94.97	102.61	102.99
n-octane	8	111.70	102.52	112.00	112.18
nonane	9	121.06	109.97	121.30	121.06
decane	10	130.44	117.31	130.96	130.51
dodecane	12	148.78	131.60	149.42	149.21
tetradecane(linear)	14	167.40	144.73	165.85	166.01
tetradecane(folded)	14	167.40	140.44	165.68	167.10
hexadecane(linear)	16	186.02	158.25	180.02	182.79
hexadecane(folded)	16	186.02	153.76	182.08	184.85
octadecane(linear)	18	204.50	171.72	193.19	—
octadecane(folded)	18	204.50	164.09	193.66	—

LBH Set Heat Capacities

Table A3.4.: Heat capacities for *n*-octane in the range of 300 up to 1500K. All values correspond to cal mol^{-1} K^{-1}.

T / K	experiment[a]	$C_{p,RRHO}$	$C_{p,msRRHO}$	$C_{p,msRRHO}$+$C_{p,conf}$
300	45.10	40.6	42.4	45.95
400	57.30	53.2	55.4	58.17
500	68.55	65.2	67.6	69.71
600	78.10	75.7	78.2	79.80
700	86.10	84.7	87.2	88.47
800	92.80	92.5	94.9	95.94
900	98.40	99.2	101.5	102.38
1000	103.10	105.0	107.3	107.94
1100	107.20	110.0	112.2	112.75
1200	110.70	114.3	116.4	116.91
1300	114.00	118.0	120.1	120.50
1400	117.00	121.3	123.2	123.61
1500	119.00	124.1	126.0	126.31

[a]Values taken from Refs. 434,439.

Table A3.5.: Heat capacities for a subset of the LBH benchmark set. C_p are given for a combination of $C_{p,msRRHO}$ calculated at DFT(B97-3c or B3LYP-D3/def2-TZVP) and $C_{p,conf}$ calculated at a lower (GFN2-xTB or GFN-FF) level. Mean deviation (MD), mean average deviation (MAD), root-mean-square deviation (RMSD) and standard deviation (SD) are given below. All values correspond to cal mol^{-1} K^{-1}.

			B97-3c		B3LYP-D3/TZ		UM-VT[a]
	T / K	experiment	GFN-FF	GFN2-xTB	GFN-FF	GFN2-xTB	
isopentane	317.2	29.95	29.47	29.61	29.17	29.65	29.97
	358.2	33.25	32.79	32.91	32.44	32.91	33.04
	402.3	36.72	36.33	36.45	35.95	36.41	36.29
	449.2	40.24	39.98	40.08	39.57	40.00	39.59
	487.1	42.93	42.79	42.88	42.36	42.76	42.11
n-hexane	333.9	37.35	37.16	37.75	36.80	37.20	36.86
	365.2	40.22	40.22	40.72	39.81	40.13	39.54
	398.9	43.30	43.47	43.89	43.03	43.28	42.42
	433.7	46.39	46.74	47.09	46.27	46.46	45.33
	468.9	49.46	49.92	50.21	49.42	49.56	48.17
2,2-dimethylbutane	341.6	38.10	37.97	37.97	37.58	37.89	39.24
	353.2	39.25	39.10	39.10	38.70	39.04	40.33
	376.1	41.50	41.32	41.32	40.90	41.30	42.45
	412.4	44.95	44.77	44.77	44.32	44.82	45.73
	449.4	48.33	48.16	48.16	47.68	48.27	48.92
2,3-dimethylbutane	341.6	37.78	37.57	37.57	37.18	37.18	38.91
	371.2	40.69	40.45	40.45	40.02	40.02	41.58
	402.3	43.63	43.43	43.43	42.98	42.98	44.32
	436	46.73	46.57	46.57	46.10	46.10	47.17
	471.2	49.77	49.72	49.72	49.23	49.23	50.00
2-methylpentane	325.1	36.77	37.31	37.69	36.71	36.94	36.51
	362.2	40.30	40.98	41.27	40.33	40.52	39.79
	402.3	44.08	44.83	45.04	44.16	44.32	43.26
	436.2	47.14	47.95	48.12	47.28	47.42	46.10
	471.2	50.16	51.03	51.15	50.35	50.49	48.90
3-methylpentane	332.1	36.88	36.98	36.77	36.29	36.38	37.54
	367.6	40.25	40.46	40.20	39.74	39.76	40.54
	402.4	43.43	43.82	43.52	43.08	43.05	43.46
	436.2	46.52	46.98	46.67	46.23	46.18	46.23
	471.2	49.55	50.12	49.81	49.37	49.30	48.99
n-heptane	357.1	45.77	46.02	46.54	45.61	45.80	46.18
	373.2	47.51	47.84	48.30	47.42	47.55	47.66
	400.4	50.37	50.89	51.25	50.44	50.47	50.18
	434.4	53.85	54.58	54.84	54.10	54.06	53.29
	466.1	57.00	57.89	58.08	57.38	57.29	56.12
2,2,3-trimethylbutane	328.8	42.74	41.83	41.83	42.04	42.68	44.11
	348.9	45.09	44.13	44.13	44.30	45.00	46.22
	369.2	47.39	46.43	46.43	46.57	47.32	48.33
	400.4	50.92	49.90	49.90	50.01	50.81	51.49
	434.3	54.54	53.55	53.55	53.63	54.47	54.79
	461.8	57.36	56.39	56.39	56.46	57.33	57.36
n-octane	405.7	58.00	58.34	58.89	57.67	58.19	57.51
	462.5	64.70	65.25	65.58	64.52	64.84	63.38
	522.7	70.60	71.99	72.17	71.22	71.40	69.24
	MD	—	0.05	0.17	-0.39	-0.11	-0.05
	MAD	—	0.47	0.57	0.47	0.25	0.68
	RMSD	—	0.58	0.69	0.54	0.32	0.78
	SD	—	0.58	0.68	0.38	0.31	0.79

[a] Values taken from Ref. 380.

The CD25 Set: Drug Molecules

Table A3.6.: Comparison of a qualitative empirical flexibility measure ξ_f (see above) and the conformational entropy per atom for all molecules of the CD25 set at GFN2-xTB and GFN-FF level. Entropy values are given in $\mathrm{cal\,mol^{-1}\,K^{-1}}$ and are normalized to the number of atoms N_{at}. Values for tetra- and octadecane are given as further reference.

molecule	ξ_f	S_{conf}/N_{at} GFN-FF	GFN2-xTB
Apixaban	0.188	0.20	0.18
Aripiprazole	0.317	0.26	0.25
Celecoxib	0.161	0.15	0.13
Chloroquine	0.395	0.25	0.40
Duloxetine	0.376	0.30	0.25
Enzalutamide	0.160	0.10	0.12
Esomeprazole	0.379	0.27	0.24
Ezetimibe	0.258	0.21	0.18
Guaiol	0.230	0.10	0.15
Ibrutinib	0.189	0.17	0.19
Ibuprofen	0.341	0.18	0.13
Imatinib	0.214	0.19	0.16
Lenalidomid	0.135	0.18	0.11
Lisdexamfetamin	0.452	0.22	0.50
Oxycodone	0.160	0.05	0.01
Palbociclib	0.235	0.16	0.18
Penicilin	0.301	0.20	0.19
Pregabalin	0.515	0.33	0.46
Ritonavir	0.348	0.16	0.16
Rivaroxaban	0.280	0.24	0.12
(z)-Rosuvastatin	0.318	0.16	0.25
Sitagliptin	0.265	0.27	0.25
Sofosbuvir	0.292	0.03	0.15
Tamiflu	0.471	0.30	0.32
Tenofovir	0.302	0.25	0.14
C14	0.852	0.48	0.48
C18	0.836	0.53	–

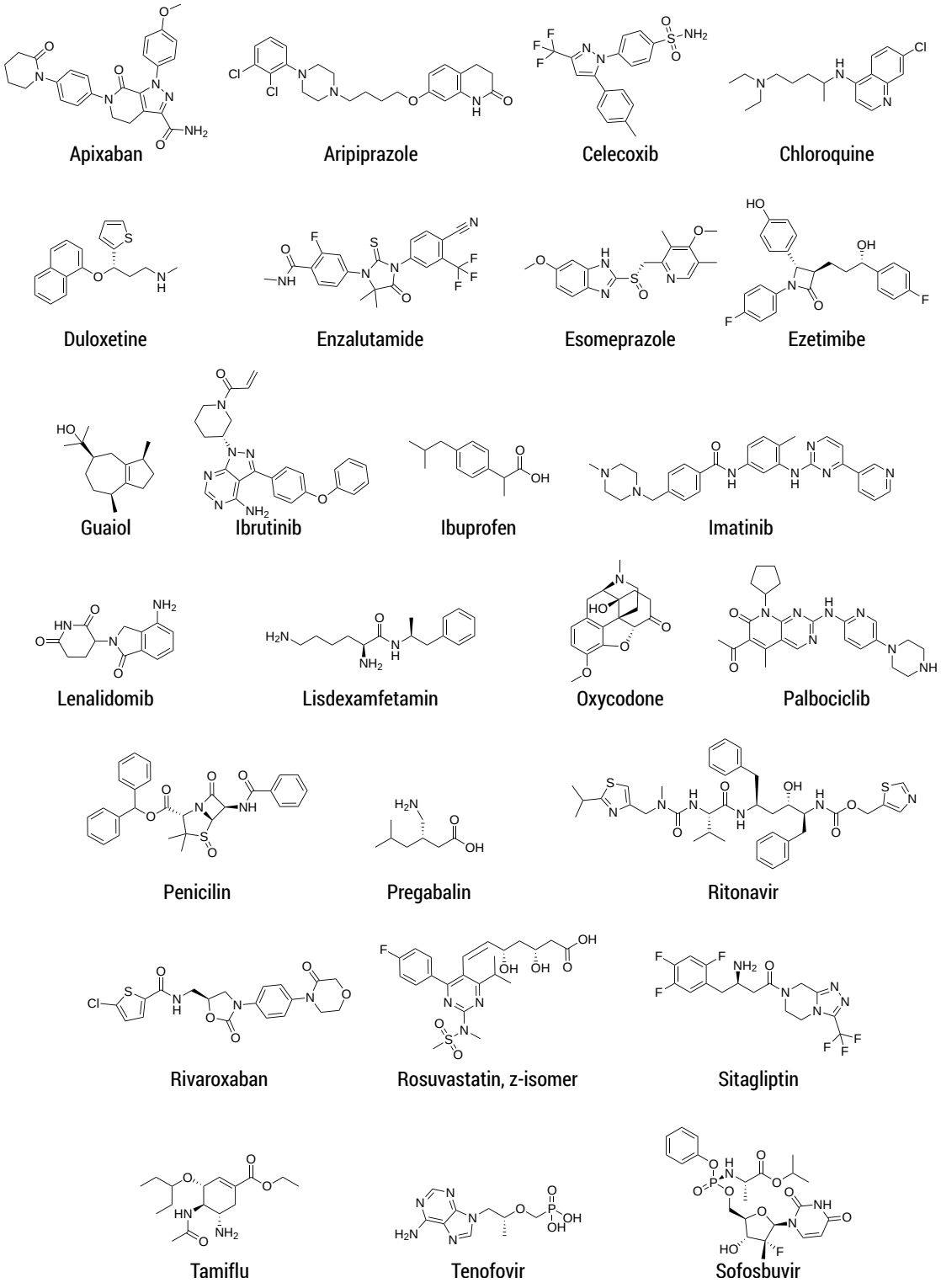

Figure A3.2.: Lewis structures for all 25 molecules included in the CD25 set. Molecule names are given below the respective structure. Input structures are available from https://github.com/grimme-lab/mol-entropy.

Empirical Entropy Estimates

As mentioned in the manuscript, the empirical formulation

$$S_{simple} = R \ln (N_{conf}) \tag{A3.7}$$

is used in some studies[402,419] to estimate the conformational entropy. However, while this formulation may be used for very simple molecules, it breaks down for challenging energy surfaces. One could easily imagine a case where only a few conformers of an otherwise large ensemble contribute to the entropy (*e.g.*, sofosbuvir at GFN-FF level), or an opposite case with many high-energetic conformers that individually contribute nothing, but in sum make a large part of the entropy (*e.g.*, continuous ensembles, large *n*-alkanes). Population differences can thus lead to significant differences even for ensembles of same size, and would therefore not be captured by the approximation via N_{conf}. The approximation is further unable to capture vibrational entropy averages as in $\overline{S}_{msRRHO}$. Differences $\Delta S_{conf/simple}$ between this estimated and the fully converged entropy often exceed several $\mathrm{cal\,mol^{-1}\,K^{-1}}$ in either direction, which is shown for the CD25 in Tabs. A3.7,A3.8 below.

Table A3.7.: Conformational entropies and standard deviations (SD) calculated for the CD25 set from repeated CREST entropy sampling runs at GFN2-xTB level. Also shown is the simple entropy $S_{simple} = R \ln (N_{conf})$, estimated only from the number of conformers N_{conf} for each structure. All entropy values correspond to cal mol^{-1} K^{-1}.

molecule	S_{conf}(GFN2-xTB)	SD	N_{conf}	S_{simple}	$\Delta S_{conf/simple}$
Apixaban	10.42	0.04	123	9.56	0.85
Aripiprazole	14.22	0.49	4273	16.61	-2.39
Celecoxib	5.08	0.38	10	4.51	0.57
Chloroquine	18.99	0.16	6499	17.45	1.54
Duloxetine	10.06	0.11	726	13.09	-3.02
Enzalutamide	5.96	0.11	23	6.20	-0.24
Esomeprazole	10.16	0.24	440	12.10	-1.93
Ezetimibe	9.33	0.15	664	12.91	-3.59
Guaiol	6.22	0.15	125	9.59	-3.38
Ibrutinib	10.76	0.54	470	12.23	-1.47
Ibuprofen	4.36	0.03	14	5.20	-0.84
Imatinib	11.10	0.22	851	13.41	-2.30
Lenalidomib	3.50	0.06	12	4.99	-1.49
Lisdexamfetamin	22.07	1.58	10044	18.31	3.76
Oxycodone	0.30	0.00	9	4.37	-4.07
Palbociclib	11.24	0.23	613	12.75	-1.52
Penicilin	7.62	0.14	193	10.46	-2.84
Pregabalin	12.86	0.31	870	13.45	-0.59
Ritonavir	15.34	0.17	11895	18.65	-3.31
Rivaroxaban	5.65	0.12	40	7.35	-1.69
(z)-Rosuvastatin	15.30	0.13	572	12.62	2.68
Sitagliptin	10.83	0.49	930	13.58	-2.75
Sofosbuvir	9.58	0.03	1756	14.85	-5.27
Tamiflu	16.04	0.13	8863	18.06	-2.03
Tenofovir	4.63	0.03	65	8.29	-3.66
average	—	0.25	—	—	-1.56

Table A3.8.: Conformational entropies and standard deviations (SD) calculated for the CD25 set from repeated CREST entropy sampling runs at GFN-FF level. Also shown is the simple entropy $S_{simple} = R \ln (N_{conf})$, estimated only from the number of conformers N_{conf} for each structure. All entropy values correspond to cal mol^{-1} K^{-1}.

molecule	S_{conf}(GFN-FF)	SD	N_{conf}	S_{simple}	$\Delta S_{conf/simple}$
Apixaban	11.69	0.34	262	11.07	0.62
Aripiprazole	14.89	0.16	3602	16.27	-1.39
Celecoxib	6.14	0.44	19	5.80	0.34
Chloroquine	11.95	0.14	4972	16.91	-4.97
Duloxetine	12.07	0.69	1157	14.02	-1.95
Enzalutamide	4.87	0.21	17	5.65	-0.78
Esomeprazole	11.62	0.61	490	12.31	-0.69
Ezetimibe	10.47	0.33	214	10.66	-0.20
Guaiol	4.21	0.45	67	8.35	-4.14
Ibrutinib	9.45	0.35	1110	13.94	-4.48
Ibuprofen	5.98	0.24	54	7.94	-1.96
Imatinib	13.03	0.19	1916	15.02	-1.99
Lenalidomib	5.86	0.22	9	4.37	1.50
Lisdexamfetamin	9.76	0.70	2429	15.49	-5.73
Oxycodone	2.23	0.02	28	6.61	-4.38
Palbociclib	9.90	0.33	571	12.61	-2.72
Penicilin	8.10	0.21	582	12.65	-4.55
Pregabalin	9.19	0.15	771	13.21	-4.02
Ritonavir	16.02	0.72	1467	14.49	1.53
Rivaroxaban	11.20	0.56	190	10.43	0.77
(z)-Rosuvastatin	9.65	0.41	628	12.80	-3.15
Sitagliptin	11.52	0.32	917	13.55	-2.03
Sofosbuvir	1.66	0.38	103	9.21	-7.55
Tamiflu	15.21	0.24	9559	18.21	-3.00
Tenofovir	8.36	0.36	236	10.85	-2.50
average	—	0.35	—	—	-2.30

A4. Supporting Information to Chapter 5

Appendix A2 contains:

- Revised results obtained with `xtb` 6.4.1

Revised Calculations of IR Spectra

In the article [*J. Chem. Theor. Comput.* **2020**, *16*, 7044–7060.] calculations were conducted with version 6.3.2 of the `xtb` code. Unfortunately, up to program version 6.4.0 the calculation of IR intensities was wrongly implemented. In all calculations at the GFNn levels in Chapter 5 IR intensities were hence obtained as

$$\text{IR intensity} \propto \frac{\partial \mu}{\partial q_i} \,, \tag{A4.1}$$

whereas it correctly (as stated in the article) should be

$$\text{IR intensity} \propto \left(\frac{\partial \mu}{\partial q_i} \right)^2 \,. \tag{A4.2}$$

DFT results at the B3LYP-3c level are not affected by this as they were obtained with TURBOMOLE. For completeness in scope of this thesis, results obtained for the 6556 structures in the HCNO set at the GFNn levels were revised with `xtb` 6.4.1 and are shown in Tab. A4.1 and Fig. A4.1.

Table A4.1.: Revised average metrics r_{match}, r_{euclid}, $r_{spearman}$, and $r_{pearson}$ for the 6556 unscaled and ideally (MSFS) scaled IR spectra calculated at the GFN1–xTB, GFN2–xTB and GFN-FF levels of theory.

	unscaled				molecule specific scaling			
method	r_{match}	r_{euclid}	$r_{spearman}$	$r_{pearson}$	r_{match}	r_{euclid}	$r_{spearman}$	$r_{pearson}$
GFN1–xTB	0.629	0.712	0.745	0.604	0.804	0.768	0.981	0.771
GFN2–xTB	0.670	0.739	0.749	0.665	0.814	0.785	0.990	0.758
GFN–FF	0.470	0.618	0.604	0.386	0.648	0.455	0.993	0.618

As can be seen from Tab. A4.1, average similarity measures for unscaled GFN1– and GFN2–xTB are similar but slightly worse than the original values (*cf.* Tab. 5.1). This is the same trend for GFN–FF, but here the differences are more significant. However, upon MSFS the correct IR intensities seem to improve predictions for all GFNn methods. Since atomic mass scaling

showed performances more close to MSFS than to the unscaled results, an improvement would be expected therein also, but is not tested here again.

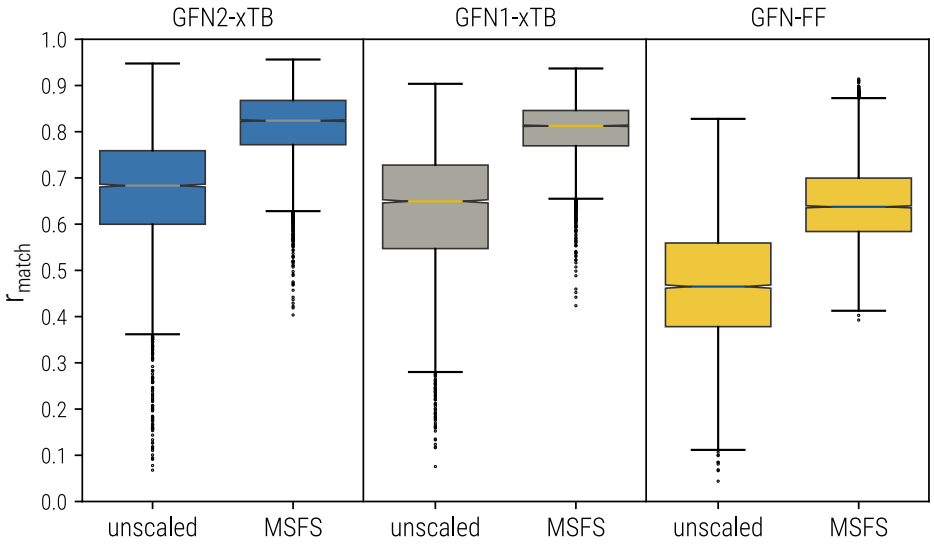

Figure A4.1.: Similarity measures (r_{match}) for the revised 6556 spectra comparisons (unscaled and with molecule specific frequency scaling) at the GFN1–xTB, GFN2–xTB and GFN–FF levels of theory, visualized as box plots.

General conclusions drawn in the article remain unchanged. An improved version of the mass scaling is already being investigated and will not be revised in its current form.

A5. Supporting Information to Chapter 6

Appendix A5 contains:

- Tables with detailed results for submission xvxzd
- LFER fit set data for submission xvxzd

Submission: xvxzd

Calculated Free Energies

Free energies for all the 24 molecules were calculated at the DSD-BLYP-D3/def2-TZVPD//PBEh-3c[DCOSMO-RS(water)] level including RRHO(GFN-xTB[GBSA(water)]) entropic and COSMO-RS(fine) free solvation energy contributions. For every molecule the chemical ensemble was generated as described in the manuscript. The Boltzmann averaged free energies for the 24 ensembles in the neutral, anionic and cationic form are shown in Table A5.1 below.

Table A5.1.: Calculated and Boltzmann averaged free energies for the 24 SAMPL6 molecules in their neutral ($|G^{neutral}|$), anionic ($|G^{anion}|$) and cationic ($|G^{cation}|$) form. All free energies are in E_h, calculated at the DSD-BLYP-D3/def2-TZVPD//PBEh-3c[DCOSMO-RS]+RRHO(GFN-xTB[GBSA])+COSMO-RS level.

molecule	$\|G^{neutral}\|$	$\|G^{anion}\|$	$\|G^{cation}\|$
SM01	-743.7625031037	-743.304788054	-744.1797905832
SM02	-1040.7353285759	-1040.2726514917	-1041.173044208
SM03	-1575.0340919676	-1574.5864421631	-1575.4597665047
SM04	-1202.4860068249	-1202.0146845191	-1202.9246444728
SM05	-1339.1374887118	-1338.6714899641	-1339.5730059192
SM06	-3390.0017408338	-3389.5397698749	-3390.4335912621
SM07	-743.0317437548	-742.5601021141	-743.4713811308
SM08	-974.6507758069	-974.210045212	-975.071439758
SM09	-818.2140852338	-817.7501438044	-818.6523439353
SM10	-1329.5898905202	-1329.1393702205	-1330.0145992036
SM11	-697.7974209236	-697.3544887555	-698.2313575709
SM12	-1163.2298569976	-1162.7664450465	-1163.6679815996
SM13	-971.9133840241	-971.4475719151	-972.3518768465
SM14	-665.6950182003	-665.1988446134	-666.1315341035
SM15	-685.5782512469	-685.1235888474	-686.0130711228
SM16	-1566.4043299299	-1565.9487240186	-1566.8422637859
SM17	-1176.9574534087	-1176.46109157	-1177.3841878893
SM18	-1776.1133054991	-1775.6618360661	-1776.5400521634
SM19	-2233.5806110999	-2233.128985007	-2234.0042211518
SM20	-2252.2700686168	-2251.8262655113	-2252.6698786386
SM21	-6081.5402750583	-6081.0749829135	-6081.9738638345
SM22	-1070.5772599332	-1070.1320102089	-1071.0078985211
SM23	-1409.7020267963	-1409.2351107628	-1410.1381492314
SM24	-1315.0474897735	-1314.5727724932	-1315.4762931794
$(H_2O)_4$	-306.1838613663	—	—
$H_3O^+(H_2O)_3$	—	—	-306.6040567912

$(H_2O)_4$ and $H_3O^+(H_2O)_3$ were used as reference molecules in the calculation of $|\Delta G_{diss}|$.

Fit Data

The LFER parameters c_0 and c_1 were obtained by fitting calculated free dissociation energies of 59 small organic and inorganic molecules to their corresponding experimental pK_a values. The free dissociation energies $|\Delta G_{diss}|$ (for the reaction according to Eq. 6.6 below) were calculated at the DSD-BLYP-D3(BJ)/def2-TZVPD//PBEh-3c[DCOSMO-RS(water)] level including RRHO contributions at the GFN-xTB[GBSA(water)] level and contributions to the free solvation energy calculated with COSMO-RS(fine). Most of the molecules are small and show only a single protonation or deprotonation site. Hence, the conformational search and averaging of the free energies was only performed for the more flexible molecules for which several conformations can be expected, such as pentanoic acid. The (averaged) $|\Delta G_{diss}|$, experimental and fitted pK_a values of the 59 molecules are shown in Table A5.2.

Calculated Macroscopic pKa

Macroscopic pK_a values were calculated according to the free dissociation energy of either one of the two reactions:

$$\mathrm{AH}_{solv} + (\mathrm{H_2O})_{4,solv} \rightarrow \mathrm{A}^{-1}_{solv} + \mathrm{H_3O^+}(\mathrm{H_2O})_{3,solv} \tag{A5.1}$$

$$\mathrm{AH}^{+}_{solv} + (\mathrm{H_2O})_{4,solv} \rightarrow \mathrm{A}_{solv} + \mathrm{H_3O^+}(\mathrm{H_2O})_{3,solv} \tag{A5.2}$$

Table A5.2.: Set of 59 small organic and inorganic molecules used to fit calculated free dissociation energies to experimental pK_a values via the LFER. Conformational searches were conducted where deemed necessary. $|\Delta G_{diss}|$ in kcal mol^{-1} calculated at the DSD-BLYP-D3/def2-TZVPD//PBEh-3c level.

molecule	$\lvert\Delta G_{diss}\rvert$	$\frac{\lvert\Delta G_{diss}\rvert}{\ln(10)RT}$	pK_a(exptl.)	pK_a(calc., fitted)
2,2-Dimethylsuccinimide	23.2986	17.0780	9.50	8.53
2,3,4,6-Tetrachlorphenol	16.5083	12.1007	5.62	5.71
2,3,4-Trichlorphenol	19.4950	14.2899	7.10	6.95
2,3-Dichlorphenol	21.9997	16.1259	7.76	7.99
2,4,5-Trichlorphenol	19.5371	14.3208	7.07	6.96
2,4,6-Trichlorphenol	18.2140	13.3509	6.21	6.42
2,4-Dichlorphenol	22.5525	16.5311	8.09	8.22
2,5-Dichlorphenol	21.6007	15.8335	7.51	7.82
2,6-Dihclorphenol	19.8357	14.5396	6.79	7.09
2-Acetylbutanedioic acid	8.0680	5.9139	2.86	2.20
2-Chloropropanoic acid	11.8961	8.7199	2.83	3.79
2-Chlorphenol	24.0541	17.6318	8.29	8.84
2-Hydroxypropanoic acid	11.3521	8.3211	3.86	3.57
3,4-Dichlorphenol	22.8374	16.7399	8.68	8.34
3,5-Dichlorophenol	21.9947	16.1222	8.27	7.99
3-Chloropropanoic acid	13.9477	10.2238	3.98	4.64
3-Chlorphenol	24.2560	17.7798	8.78	8.92
4-Chlorphenol	25.7218	18.8542	9.14	9.53
4-Nitrophenol	18.2826	13.4012	7.14	6.44
5-Formyluracil	15.7631	11.5544	6.84	5.40
5-Nitrouracil	13.9895	10.2544	5.30	4.66
Acetic acid	15.1742	11.1227	4.76	5.15
Acrylic acid	14.3777	10.5389	4.25	4.82
Benzoic acid	15.6034	11.4374	4.20	5.33
Boric acid	24.2896	17.8044	9.23	8.94
Bromoacetic acid	11.9388	8.7512	2.86	3.81
Carbonic acid	11.0107	8.0709	3.58	3.42
Chloroaceticacid	11.5827	8.4902	2.86	3.66
Cyanoacetic acid	10.2971	7.5478	2.43	3.13
Dichloroacetic acid	7.7051	5.6479	1.29	2.05
Dimethadione	26.5509	19.4619	6.10	9.88
Fluoro acetic	10.1897	7.4691	2.66	3.08
Fluorouracil	19.8093	14.5203	8.00	7.08
Formic acid	12.1143	8.8798	3.77	3.88
Fumaric acid	11.6516	8.5407	3.02	3.69
Hypobromous acid	24.0432	17.6238	8.55	8.84
Hypochlorous acid	21.7570	15.9480	7.40	7.89
Hypoiodous acid	25.7266	18.8578	10.50	9.53
Iodoacetic acid	13.0036	9.5317	3.12	4.25
Maleic acid	12.9390	9.4843	1.93	4.23
Methylperoxide	27.1417	19.8950	11.50	10.12
Methylthiouracil	20.7597	15.2170	8.20	7.47
Nitroacetic acid	8.1799	5.9959	1.68	2.25
Nitrous acid	6.3838	4.6794	3.29	1.50
Oxalic acid	5.6623	4.1505	1.23	1.20
Pentanoic acid	15.5898	11.4274	4.84	5.33
Peroxide	27.2617	19.9830	11.60	10.17
Phenol	26.4243	19.3691	9.82	9.82
Phenytoin	23.1427	16.9637	8.30	8.46
Phosphoric acid	4.0589	2.9752	2.16	0.54
Phthalic acid	12.6270	9.2556	2.98	4.10
Phthalimide	23.2423	17.0367	8.30	8.50
Pivalic acid	15.3076	11.2206	5.03	5.21
Succinimide	22.9325	16.8096	9.60	8.37
Sulfurous acid	1.6421	1.2036	1.90	-0.47
Thymine	23.8240	17.4631	9.75	8.74
Trichloracetic acid	4.1161	3.0171	0.65	0.56
Trifluoroethanol	32.0117	23.4648	12.50	12.14
Uracil	22.4289	16.4405	9.42	8.17

Table A5.3.: Macroscopic pK_a values calculated with the fitted LFER for the reaction shown in Eq. A5.1. $|\Delta G_{diss}|$ in kcal mol^{-1}. "*" denotes pK_a calculations that were omitted from the SAMPL6 submission.

molecule	$\lvert\Delta G_{diss}\rvert$	$\frac{\lvert\Delta G_{diss}\rvert}{\ln(10)RT}$	pK_a(calc.)
SM01	27.1940	19.9334	10.14
SM02	30.3077	22.2157	11.44
SM03	20.8780	15.3037	7.52
SM04*	35.7327	26.1923	13.69
SM05	32.3921	23.7436	12.30
SM06	29.8646	21.8909	11.25
SM07*	35.9331	26.3391	13.77
SM08	16.5361	12.1211	5.72
SM09	31.1011	22.7973	11.77
SM10	22.6792	16.6240	8.27
SM11	17.9176	13.1337	6.29
SM12	30.7689	22.5538	11.63
SM13	32.2750	23.6577	12.25
SM14*	51.3271	37.6230	20.16
SM15	25.2785	18.5292	9.35
SM16	25.8705	18.9632	9.59
SM17*	51.4452	37.7096	20.21
SM18	23.2748	17.0606	8.52
SM19	23.3731	17.1326	8.56
SM20	18.4641	13.5343	6.52
SM21	31.9487	23.4186	12.12
SM22	19.3719	14.1997	6.90
SM23*	32.9677	24.1655	12.54
SM24*	37.8631	27.7538	14.57

Table A5.4.: Macroscopic pK_a values calculated with the fitted LFER for the reaction shown in Eq. A5.2. $|\Delta G_{diss}|$ in kcal mol^{-1}. "*" denotes pK_a calculations that were omitted from the SAMPL6 submission.

molecule	$\lvert\Delta G_{diss}\rvert$	$\frac{\lvert\Delta G_{diss}\rvert}{\ln(10)RT}$	pK_a(calc.)
SM01*	1.8253	1.3380	-0.39
SM02	14.6442	10.7343	4.93
SM03	7.0883	5.1958	1.80
SM04	15.2228	11.1584	5.17
SM05	13.2647	9.7231	4.36
SM06	10.9637	8.0365	3.41
SM07	15.8501	11.6182	5.43
SM08*	3.9441	2.8910	0.49
SM09	14.9850	10.9841	5.07
SM10*	6.4822	4.7515	1.54
SM11	12.2728	8.9961	3.95
SM12	14.9008	10.9224	5.04
SM13	15.1319	11.0918	5.14
SM14	13.8914	10.1824	4.62
SM15	12.8271	9.4023	4.18
SM16	14.7811	10.8346	4.99
SM17	7.7534	5.6833	2.07
SM18	7.7611	5.6889	2.08
SM19*	5.7928	4.2462	1.26
SM20*	-9.1419	-6.7011	-4.94
SM21	12.0546	8.8361	3.86
SM22	10.2033	7.4791	3.09
SM23	13.6444	10.0015	4.52
SM24	9.0517	6.6349	2.61

A6. Supporting Information to Chapter 7

Appendix A6 contains:

- Additional computational details
- Tables with fit data
- CFER parameters c_{0-3}
- Tables with detailed results
- Figures with molecular structures

Additional Computational Details

In the manuscript the three steps for computation of dissociation free energies and consequently pK_a values were discussed. Some additional technical aspects are given here:

- Ionic starting geometries for the acid and its conjugate base can either be taken from the literature (as was the case, *i.e.* for the SAMPL6 benchmark sets), or generated from the neutral species at GFN2-xTB level with the command `crest inp.xyz --protonate --alpb h2o` (respectively `--deprotonate` for the anion).

- Alternatively, if the acid input geometry is known the pK_a at GFN2-xTB level can be calculated in a single-structure approach *via* `crest inp.xyz --pka <acidic H>`, where `<acidic H>` is the number of the acidic proton in the acid input file. This will automatically generate the base and calculate a pK_a value. Note that the charge of the acid should be specified via the `--chrg` command.

- Conformational ensembles for consecutive DFT treatment were generated separately for the acid and base with `crest`. To reduce computational cost, GFN2-xTB structures were clustered using the `--cluster` command.

- Clustered ensembles were passed to the `censo` program, which interfaces Turbomole and yields ensemble files containing the final free energies at DFT level.

- The respective `censo` ensembles can directly be read by `crest`, which calculates the Boltzmann averaged $\overline{G}^A/\overline{G}^B$ and from this the pK_a. The corresponding command is `crest --pka --pkaensemble acid-ensemble.xyz base-ensemble.xyz`. CFER parameters can be read from a plain text file via `--pkaparam <file>`.

Detailed Results

Adjustment of GFN2-xTB Dissociation Energies

Table A6.1.: Element specific parameters to calculate an energy correction term E_{mod}^{TB} for dissociation energies at the GFN2-xTB/ALPB(H_2O) level.

element X	$\varepsilon(X)$	k_1	k_2	k_3	k_4
C	0.243240	0.006819	-0.012581	0.154801	0.143015
N	0.257074	0.037121	-0.055775	0.106734	0.023837
O	0.194618	0.024046	0.028236	0.014825	0.149677
F	0.122274	-0.147062	0.328353	-0.327531	0.234198
Si	0.212852	0.019048	-0.061229	0.026066	0.008854
P	0.274942	-0.023403	0.025728	-0.013699	-0.026747
Si	0.301025	0.008263	-0.080627	0.093954	-0.051587
Cl	0.278736	-0.008886	-0.008372	0.014058	-0.057756

Table A6.2.: Energy differences $\Delta E = E_{base} - E_{acid}$ at the r^{2}SCAN-3c/COSMO-RS(H_2O), GFN2-xTB/ALPB(H_2O) and GFN2-xTB+E_{mod}^{TB}/ALPB(H_2O) levels. The element specific parameters of E_{mod}^{TB} in Tab. A6.1 were fitted to these energies. All energies are in Eh. The "tag" is a trivial label corresponding to the system as provided in the separate input structure zip file.

fit element	tag	ΔE_{DFT}	ΔE_{GFN2}	$\Delta E_{GFN2} + E_{mod}^{TB}$	$\Delta\Delta E_{DFT/GFN2} + E_{mod}^{TB}$
C	acetaldehyd	0.518258	0.297300	0.528032	-0.009774
C	aceton	0.479308	0.246949	0.472988	0.006320
C	acetylaceton	0.446127	0.205680	0.437789	0.008339
C	barbituric	0.434592	0.181474	0.420908	0.013684
C	benzol	0.529184	0.325102	0.527222	0.001962
C	butadien	0.527351	0.323373	0.525149	0.002202
C	bzh+	0.375641	0.126545	0.400290	-0.024648
C	c4oen	0.512839	0.280436	0.522994	-0.010154
C	cf3h	0.500405	0.217474	0.503984	-0.003579
C	ch2no22	0.421813	0.147368	0.404471	0.017342
C	ch3cn	0.485975	0.268872	0.493184	-0.007209
C	ch3nh2	0.547339	0.328233	0.551091	-0.003752
C	chbr3	0.484328	0.196193	0.485281	-0.000953
C	chcl3	0.489328	0.205443	0.473581	0.015747
C	chexanon	0.479771	0.245278	0.474416	0.005356
C	chlorethin	0.475893	0.265922	0.471883	0.004011
C	chme2+	0.489907	0.241325	0.488522	0.001385
C	cpdien	0.458051	0.236263	0.477236	-0.019185
C	cyclohexandion	0.435026	0.198086	0.432699	0.002327
C	cyclononyl	0.448908	0.230845	0.471853	-0.022945
C	dihydrofuran	0.536414	0.306699	0.534830	0.001583
C	diphenylmethan	0.503804	0.262988	0.495229	0.008575
C	ethan	0.547117	0.333451	0.548228	-0.001111
C	ethen	0.526964	0.321775	0.527700	-0.000736
C	ethin	0.470002	0.302895	0.488944	-0.018943
C	fluoren	0.483967	0.245019	0.482208	0.001759
C	lacton5	0.487312	0.241447	0.469354	0.017958
C	malonitril	0.442189	0.209823	0.449159	-0.006970
C	pyridin	0.525787	0.314269	0.531305	-0.005518
C	tbutyl+	0.403114	0.159821	0.408595	-0.005481
C	toluol	0.515513	0.285823	0.510551	0.004962
N	acetamid	0.473020	0.228146	0.468414	0.004606
N	anilin	0.488184	0.254263	0.490695	-0.002512
N	ch3cnh+	0.395281	0.156339	0.413997	-0.018715
N	diphenylamineh+	0.422952	0.159406	0.425642	-0.002690
N	guanidinium+	0.465492	0.214756	0.457551	0.007941
N	hn3	0.428340	0.203318	0.435842	-0.007501
N	hydrazinium+	0.444366	0.182031	0.442370	0.001997
N	imidazol	0.461993	0.221536	0.468007	-0.006014
N	imidazolh+	0.447593	0.192420	0.445405	0.002188
N	isocyanic	0.429788	0.199369	0.433278	-0.003489
N	morpholineh+	0.450425	0.180689	0.442596	0.007829
N	Naphthyridinium	0.438535	0.181843	0.439902	-0.001367
N	nh3	0.508929	0.291057	0.525937	-0.017008
N	nh4+	0.450836	0.200003	0.451506	-0.000669
N	nph3h+	0.411569	0.147858	0.416860	-0.005291
N	phnh3+	0.434460	0.174413	0.434555	-0.000095
N	phtalimid	0.449301	0.205287	0.452585	-0.003284
N	pyridinium+	0.444893	0.182091	0.440050	0.004843
N	pyrimidinium	0.433031	0.166103	0.424219	0.008811
N	pyrrol	0.470363	0.240632	0.486836	-0.016472
N	sacharin	0.428778	0.178392	0.421435	0.007343

Table A6.2.: Continue previous table...

fit element	tag	ΔE_{DFT}	ΔE_{GFN2}	$\Delta E_{GFN2} + E^{\mathrm{TB}}_{mod}$	$\Delta\Delta E_{DFT/GFN2} + E^{\mathrm{TB}}_{mod}$
O	acetamidh+	0.424986	0.171974	0.424438	0.000548
O	acetonh+	0.413485	0.160852	0.419367	-0.005881
O	ala	0.437842	0.193654	0.440525	-0.002683
O	ccl3cooh	0.427043	0.161688	0.417721	0.009321
O	cf3cooh	0.425982	0.159606	0.417870	0.008113
O	ch3cooh2+	0.412401	0.153267	0.410678	0.001723
O	ch3cooh	0.438220	0.198504	0.444044	-0.005824
O	dmsoh+	0.427113	0.172271	0.419320	0.007793
O	enol	0.454712	0.225257	0.463282	-0.008570
O	h2co3	0.431799	0.182002	0.430707	0.001092
O	h2o2	0.462396	0.220957	0.474283	-0.011887
O	h2o	0.476512	0.257044	0.486801	-0.010289
O	h2po4-	0.438474	0.204968	0.444697	-0.006224
O	h2so4	0.407784	0.168973	0.408284	-0.000501
O	h3o+	0.426699	0.165183	0.422124	0.004575
O	h3po4	0.428554	0.183783	0.428061	0.000493
O	hclo3	0.408558	0.163076	0.418918	-0.010361
O	hclo4	0.399627	0.134184	0.393783	0.005844
O	hco3-	0.452962	0.224171	0.456057	-0.003095
O	hcooh	0.436114	0.185716	0.436182	-0.000068
O	hno2	0.429749	0.179599	0.432729	-0.002980
O	hno3	0.415951	0.160589	0.413361	0.002590
O	hocl	0.451763	0.200210	0.451526	0.000237
O	hso4-	0.418432	0.193737	0.432452	-0.014020
O	meoh2+	0.417264	0.153797	0.418185	-0.000921
O	meoh	0.484921	0.239711	0.477903	0.007017
O	mesulfons	0.412470	0.176101	0.419065	-0.006595
O	odiphenol	0.453611	0.217875	0.455454	-0.001843
O	oxal1	0.426535	0.179940	0.433176	-0.006641
O	oxal2	0.430682	0.194329	0.443352	-0.012670
O	phboron	0.456836	0.227545	0.462502	-0.005666
O	phcooh	0.438015	0.197940	0.443596	-0.005581
O	phenol	0.457422	0.223836	0.463281	-0.005859
O	pikrin	0.426762	0.166828	0.419855	0.006907
O	quadratic1	0.425330	0.175504	0.426235	-0.000905
O	quadratic2	0.427702	0.197336	0.443395	-0.015693
O	salicylat	0.435005	0.192297	0.438312	-0.003307
O	tbutanol	0.487440	0.254921	0.489114	-0.001674
O	thfh+	0.413682	0.148052	0.414586	-0.000905
O	trifluorethanol	0.471973	0.216740	0.460124	0.011849
O	tropolon	0.446291	0.210060	0.452503	-0.006212
O	uracil	0.450681	0.208806	0.456783	-0.006102
O	uronium+	0.428774	0.178438	0.423027	0.005746
O	vitcmod	0.436900	0.187780	0.435886	0.001014
F	etfh+	0.357602	0.081647	0.347701	0.009901
F	f2h+	0.293090	0.029602	0.297628	-0.004539
F	fethenh+	0.354159	0.123378	0.406676	-0.052517
F	hfdimer	0.414417	0.148024	0.398024	0.016393
F	lifh+	0.423745	0.129430	0.425556	-0.001811
F	mefh+	0.351819	0.079171	0.349800	0.002020
F	phfh+	0.332818	0.072970	0.337974	-0.005156
Si	sif2h	0.475726	0.255362	0.471139	0.004587
Si	sih2cl2	0.470784	0.252445	0.477973	-0.007188
Si	sih3+	0.423495	0.214151	0.415387	0.008109
Si	sih3cn	0.469097	0.279804	0.485893	-0.016796
Si	sihbr3	0.448203	0.202407	0.423487	0.024716
Si	sihcl2+	0.384558	0.177256	0.403282	-0.018724
Si	sihcl3	0.457577	0.229032	0.456155	0.001422
Si	sihclch2	0.472265	0.253635	0.503635	-0.031370
Si	sihcls	0.439781	0.228606	0.455766	-0.015985
Si	sihf-ethene	0.470471	0.274598	0.487869	-0.017398
Si	sihfo	0.493141	0.256100	0.506100	-0.012959
Si	sihme2+	0.449635	0.229321	0.438321	0.011315

Table A6.2.: Continue previous table...

fit element	tag	ΔE_{DFT}	ΔE_{GFN2}	$\Delta E_{GFN2} + E_{mod}^{\mathrm{TB}}$	$\Delta\Delta E_{DFT/GFN2} + E_{mod}^{\mathrm{TB}}$
P	p2me2h+	0.386058	0.159947	0.404984	-0.018926
P	p2meh	0.452386	0.223084	0.484567	-0.032180
P	p4o6h+	0.388565	0.139646	0.389646	-0.001082
P	p-betaine	0.433219	0.201049	0.451049	-0.017830
P	pbr3h+	0.371490	0.046515	0.300179	0.071312
P	pcl3h+	0.375187	0.123525	0.379625	-0.004438
P	ph3	0.461970	0.199120	0.460304	0.001666
P	ph3ph+	0.431622	0.177402	0.435493	-0.003871
P	ph4+	0.418091	0.158223	0.424313	-0.006223
P	phcl2	0.478012	0.191442	0.474129	0.003883
P	pme3h+	0.447142	0.190459	0.448621	-0.001479
P	poh3h+	0.431011	0.167001	0.429280	0.001731
P	pphh3+	0.422571	0.162900	0.425931	-0.003360
P	p-sbetaine	0.437980	0.198606	0.448606	-0.010626
P	verkadeh+	0.472521	0.238576	0.488576	-0.016055
S	allylthiol	0.443474	0.211998	0.449145	-0.005671
S	diphenylsh+	0.390624	0.140349	0.394892	-0.004268
S	etsh2+	0.400411	0.146905	0.395044	0.005367
S	etsh	0.444463	0.214382	0.449585	-0.005121
S	h2s2	0.436570	0.184950	0.424855	0.011715
S	h2s	0.429694	0.206467	0.443384	-0.013690
S	hsoh	0.437938	0.202132	0.439223	-0.001285
S	isopentylthiol	0.447607	0.220069	0.456043	-0.008436
S	mesh	0.443412	0.211144	0.446565	-0.003152
S	nh2sh	0.443779	0.207725	0.443774	0.000004
S	nitrophenol	0.433827	0.179988	0.435886	-0.002059
S	o-nh2thiopenol	0.437640	0.204633	0.448381	-0.010741
S	phsh2+	0.389787	0.139083	0.391911	-0.002124
S	phsh	0.438371	0.202816	0.444290	-0.005919
S	s8h+	0.360310	0.115683	0.369747	-0.009438
S	scl2h+	0.359797	0.128463	0.374592	-0.014795
S	senol	0.436870	0.203782	0.443672	-0.006801
S	shpyridone	0.436175	0.201701	0.445757	-0.009582
S	sthfh+	0.409844	0.154024	0.402663	0.007180
S	tbutylsh	0.447163	0.220320	0.456369	-0.009205
S	thioac	0.438270	0.196116	0.445785	-0.007515
S	thioacetamidh+	0.417425	0.164474	0.412065	0.005360
S	thioacetonh+	0.408605	0.155542	0.407414	0.001192
S	thioessig	0.429063	0.174817	0.423240	0.005824
S	thionapthol	0.440106	0.201315	0.446364	-0.006258
S	thiuronium	0.419348	0.170448	0.413593	0.005755
Cl	br2clh	0.381402	0.158967	0.387806	-0.006403
Cl	brclh+	0.324728	0.067346	0.305718	0.019009
Cl	cl2h+	0.321884	0.097324	0.333836	-0.011951
Cl	clethenh+	0.342988	0.111484	0.353297	-0.010309
Cl	clh2+	0.346346	0.111824	0.348884	-0.002537
Cl	etclh+	0.356187	0.116316	0.357475	-0.001289
Cl	hclhf	0.393733	0.158526	0.395230	-0.001497
Cl	meclh+	0.353979	0.116803	0.357406	-0.003427
Cl	phclh+	0.340075	0.108151	0.350073	-0.009998

Free Energy Relationships

Table A6.3.: Statistical measures of different orders of FER, determined for r^{2}SCAN-3c/COSMO-RS on the PKA74 set.

	FER orders					
	LFER	QFER	CFER	(4)FER	(5)FER	(6)FER
MAD	2.061	1.939	1.856	1.675	1.603	1.603
RMSD	2.668	2.584	2.416	2.315	2.250	2.250
SD	2.685	2.600	2.431	2.330	2.263	2.263
R^2	0.956	0.959	0.964	0.967	0.969	0.969
BIC	531.108	530.256	523.663	521.067	520.758	525.164

Table A6.4.: CFER parameters for GFN2-xTB/ALPB with and without the energy correction E_{mod}^{TB} on the PKA74 set.

		FER parameters				statistics [pK_a]	
	method	c_0	c_1	c_2	c_3	MAD	RMSD
CFER	GFN2-xTB	-95.392771	2.318512	-0.018978	0.000068	3.13	4.30
CFER	GFN2-xTB+E_{mod}^{TB}	-1855.025277	26.075982	-0.12496355	0.00020571	2.79	3.68

Table A6.5.: CFER parameters determined for GFN2-xTB+E_{mod}^{TB}/ALPB, r^{2}SCAN-3c, B97-3c, B97-D/TZ, PBE0-D3/TZ, PW6B95-D3/TZ, and ωB97X-V/TZ levels of theory using data from the PKA74 set. All DFT methods employ COSMO-RS implicit solvation, GFN2-xTB uses ALPB implicit solvation. TZ is an abbreviation for the def2-TZVPP basis set.

	CFER parameters			
	c_0	c_1	c_2	c_3
GFN2-xTB	-1855.0252772	26.0759822	-0.1249636	0.0002057
r^{2}SCAN-3c	-1511.8899792	21.1100681	-0.1011999	0.0001683
B97-3c	-1835.3033945	25.4169227	-0.1201689	0.0001956
B97-D/TZ	-1852.5125079	25.7839522	-0.1226496	0.0002009
PBE0-D3/TZ	-1441.4834019	20.0872293	-0.0961330	0.0001597
PW6B95-D3/TZ	-1580.6043989	22.2170835	-0.1069510	0.0001777
ωB97X-V/TZ	-1796.5154741	25.4641832	-0.1231099	0.0002043

Table A6.6.: CFER parameters determined for GFN2-xTB+E_{mod}^{TB}/ALPB, r^{2}SCAN-3c, B97-3c, B97-D/TZ, PBE0-D3/TZ, PW6B95-D3/TZ, and ωB97X-V/TZ levels of theory using data from the TR224 set. All DFT methods employ COSMO-RS implicit solvation, GFN2-xTB uses ALPB implicit solvation. TZ is an abbreviation for the def2-TZVPP basis set.

	CFER parameters			
	c_0	c_1	c_2	c_3
GFN2-xTB	6702.3111485	-100.4483504	0.4988740	-0.0008201
r^{2}SCAN-3c	5014.2837220	-75.7090708	0.3781462	-0.0006239
B97-3c	3032.1086142	-45.1533848	0.2212100	-0.0003555
B97-D/TZ	7250.3854583	-106.5703614	0.5191041	-0.0008370
PBE0-D3/TZ	5655.4664440	-84.3407514	0.4163611	-0.0006795
PW6B95-D3/TZ	5013.0940026	-74.7015098	0.3682453	-0.0005996
ωB97X-V/TZ	-1852.6895447	25.5589506	-0.1193445	0.0001901

The PKA74 Set

Table A6.7.: Dissociation free energy for the PKA74 set for all levels of theory discussed in the manuscript. All DFT methods employ COSMO-RS.

		$\frac{\Delta G'_{diss}}{RT\ln(10)}$						
	exp. pK_a	GFN2-xTB	r^{2}SCAN-3c	B97-3c	B97-D	PBE0-D3	PW6B95-D3	ωB97X-V
aceton	20.00	212.20352	214.76880	217.32276	217.95139	214.99135	217.17341	217.87325
acetylaceton	9.00	195.88290	199.74523	202.21462	203.55419	199.80033	202.03999	203.99023
barbituric	4.00	189.09334	195.19923	198.15227	199.14465	195.39072	197.18324	198.74507
benzol	43.00	234.50254	235.82799	238.19299	238.40700	236.18575	238.10884	236.73326
bzh+	-24.00	181.47535	169.24948	172.97912	173.40195	169.34709	169.44482	169.62491
ch3cn	25.00	219.67854	216.56670	219.67577	219.83199	216.84529	219.70658	220.61503
chcl3	25.00	212.26211	217.91378	221.46449	222.48056	219.49572	223.62573	225.12038
cpdien	16.00	214.32850	204.20298	208.03149	208.59722	204.80820	206.52780	207.04647
cyclohexandion	5.00	193.87012	195.03204	197.36135	199.21603	194.91731	197.76590	199.63630
diphenylmethan	33.50	221.42982	224.64024	227.72958	228.23777	225.26239	228.18801	231.60312
ethan	50.00	242.81419	241.56928	244.05712	243.63445	243.07631	243.92646	243.24671
ethen	44.00	233.89682	232.51591	234.67055	234.61881	234.15880	234.72940	233.66125
ethin	25.00	218.48936	209.29619	211.57882	211.96208	210.08057	211.76499	210.47333
fluoren	23.00	215.58056	215.83960	219.36070	220.14240	216.38988	219.12034	221.31874
malonitril	11.20	200.92355	197.81855	200.69295	201.41488	197.71336	201.08317	203.20410
tbutyl+	-12.00	183.79697	182.60914	185.09901	185.39320	182.38892	182.35252	182.96778
toluol	41.00	228.03701	229.77670	233.32908	233.40761	230.50261	233.37531	234.56750
acetamid	15.00	212.77864	212.09974	214.39543	215.45646	212.41026	214.57754	214.08831
anilin	27.00	219.66524	218.21182	221.06173	221.71120	218.48169	221.45857	221.30201
ch3cnh+	-10.00	186.62335	178.26390	180.78428	181.53533	178.00898	179.66644	179.05868
diphenylamineh+	0.80	189.93869	189.10111	193.41511	193.24644	188.96276	190.55723	191.92577
guanidinium+	13.40	204.81388	208.42280	211.30658	211.91802	208.34904	210.76512	210.32000
hn3	4.70	196.93040	195.02320	197.48753	198.77380	193.16997	198.09892	198.44004
hydrazinium+	8.10	198.60457	199.85492	203.60415	203.32436	199.23569	201.53644	201.43695
imidazol	14.50	198.93214	200.10742	202.73512	203.39539	200.24171	201.82940	201.54800
imidazolh+	7.00	209.76073	206.51012	209.32085	210.35081	206.69316	209.08698	208.36953
isocyanic	3.90	195.43616	194.42710	197.04591	198.24342	193.82645	196.07387	195.19715
morpholineh+	8.50	197.55272	201.49350	204.95763	205.11424	201.21260	202.66967	202.65111
Naphthyridinium	3.40	197.03335	195.99162	198.77070	199.79041	196.13148	197.72451	197.28197
nh3	36.00	234.76382	226.74442	230.48856	229.71389	227.12123	229.86449	229.19075
nh4+	9.40	201.63088	201.90391	205.82940	204.82143	201.92648	203.34136	203.28099
nph3h+	-5.00	185.77282	184.05552	187.99155	188.64328	183.85070	185.33975	186.75997
phnh3+	4.60	194.66299	194.43714	198.88709	198.21552	194.62543	196.31647	196.60919
phtalimid	9.60	202.67650	200.98894	203.58549	205.45821	201.15931	203.67594	203.08803
pyridinium+	5.20	196.73601	198.81479	201.43017	202.06016	198.96258	200.13778	199.86478
pyrimidinium	1.30	190.10311	193.96579	196.46963	197.29926	193.85476	195.32419	195.19211
pyrrol	15.00	217.72524	209.83845	212.44663	213.59191	210.14841	212.51300	211.75736
sacharin	1.60	188.97227	192.44899	195.73433	197.59603	192.34721	194.61154	194.07873
acetamidh+	0.00	189.78396	189.89838	191.84264	192.86571	189.96835	191.34970	191.13705
acetonh+	-7.00	188.58811	185.56096	187.29736	188.23528	185.45734	186.17292	185.93249
ala	2.40	197.42156	195.96900	197.08868	199.37158	196.33711	198.21193	197.75204

Table A6.7.: Continue previous table...

	exp. pK_a	GFN2-xTB	r^{2}SCAN-3c	B97-3c	$\frac{\Delta G'_{diss}}{RT\ln(10)}$ B97-D	PBE0-D3	PW6B95-D3	ωB97X-V
ccl3cooh	0.65	187.66397	190.46348	192.35554	194.29844	191.41585	193.12826	192.98196
cf3cooh	0.00	186.75904	189.66818	191.27395	193.51122	190.42088	191.83530	191.64673
ch3cooh	4.80	184.54494	185.12829	187.01868	188.08384	185.02627	186.29746	186.15939
ch3cooh2+	-6.00	199.55347	196.06586	197.18986	199.47730	196.26689	198.31271	197.90082
dmsoh+	-1.50	188.61974	191.49020	194.08632	193.40534	192.11592	192.06835	192.04927
enol	10.50	207.00412	202.66913	203.80724	205.53057	203.07367	205.18904	205.01668
h2co3	3.60	192.76472	193.55052	194.93697	197.07089	193.56484	195.78770	195.39406
h2o	15.70	213.47525	208.53482	209.50865	211.88403	208.30854	211.12815	210.86365
h2o2	11.60	219.20978	214.85623	216.59402	217.72147	214.78288	217.42654	216.77278
h2po4-	7.20	199.97158	198.08996	199.29085	201.56372	197.11522	200.56114	200.02687
h2so4	-3.00	183.72398	183.41473	186.04267	187.21730	183.54725	184.69039	184.23503
h3o+	-1.74	189.34861	192.07450	195.05873	194.83849	191.66506	193.43670	193.38388
h3po4	2.20	192.14726	191.80256	193.69456	195.28014	192.62767	193.82087	193.39499
hclo3	-2.70	188.77803	185.54831	189.32549	189.67510	183.53572	184.82092	183.47234
hclo4	-10.00	178.47795	181.48394	185.12952	185.49504	179.81486	181.05272	180.13684
hco3-	10.30	205.70287	203.44705	203.46371	207.85270	203.50794	206.97226	206.25532
hcooh	3.80	195.06533	193.04378	194.11707	196.37165	194.72352	195.25132	195.06118
hno2	3.30	194.43467	193.44207	195.31018	197.79133	192.75863	196.12839	195.88125
hno3	-1.40	186.34188	187.38380	189.31676	191.75670	186.72517	188.78617	188.35846
hocl	7.50	203.83681	202.86709	205.87096	206.76133	203.94697	205.65656	205.49720
hso4-	1.90	195.93799	189.44101	191.36947	193.61575	188.38498	191.31376	190.70213
meoh	15.50	187.40273	187.74261	190.30081	190.76659	186.89417	188.90374	188.89700
meoh2+	-2.00	213.70884	215.02545	216.02343	217.21571	216.45186	217.12785	216.87814
mesulfons	-2.60	189.06169	185.58173	188.19856	189.13506	185.56030	186.59931	186.05905
odiphenol	9.50	204.13115	203.09149	204.41329	206.31205	203.28127	205.63722	205.49574
oxal1	1.20	194.36952	191.16753	192.65389	194.90959	191.12668	193.26832	192.88966
oxal2	4.20	199.62809	193.25085	194.26556	197.41446	193.20140	196.09355	195.61609
phboron	8.80	206.58483	203.44663	204.88763	206.40646	204.32205	205.81007	205.32084
phcooh	4.20	199.02042	196.08712	197.68422	199.79158	196.14091	198.31370	197.67026
phenol	10.00	207.89355	204.57893	205.91167	208.08181	204.86121	207.45840	207.17486
pikrin	0.25	187.79133	191.28660	193.56643	195.57454	190.93559	193.98405	194.70205
quadratic1	1.50	191.01450	190.33864	191.95712	194.09027	190.55319	192.80673	192.74606
quadratic2	3.40	199.90976	192.16528	193.33408	196.23422	191.78332	195.15793	194.98446
salicylat	2.75	196.21715	194.67003	196.31027	198.24702	194.68809	196.94799	196.20726
tbutanol	20.00	219.54687	217.87329	217.80392	220.27849	218.28955	219.77573	219.53119
thfh+	-2.10	184.67089	186.01943	188.49426	189.57540	184.68914	186.94628	186.86012
trifluorethanol	12.50	205.68129	210.18702	211.33660	213.50609	211.10963	212.76205	212.58870
tropolon	6.50	202.93731	199.48237	201.23362	203.41057	199.76824	202.44953	202.23078
uracil	9.40	204.22204	200.73761	203.29906	205.26579	201.41371	203.41862	202.90427
uronium+	0.10	189.02196	191.73113	193.87550	195.06031	191.99404	193.49317	193.10196
vitcmod	4.50	195.59710	195.91690	197.56262	199.67952	195.98809	198.52243	198.29509

Table A6.8.: Acid dissociation free energies calculated for the Ar-N subset of TR224. Various levels of theory are shown. All DFT methods employ COSMO-RS implicit solvation, GFN2-xTB uses ALPB implicit solvation.

Nr.	exp. pK_a	GFN2-xTB	r^{2}SCAN-3c	B97-3c	$\frac{\Delta G'_{diss}}{RT\ln(10)}$ B97-D	PBE0-D3	PW6B95-D3	ωB97X-V
1	1.10	190.115867	193.977038	196.483659	197.310060	195.599813	195.344158	195.212046
2	2.10	197.240853	195.015125	197.881165	198.583421	196.986500	196.692696	196.395971
3	2.84	195.249593	195.635663	198.370557	199.176163	197.424035	197.202207	196.943663
4	3.28	194.838811	196.365336	199.269024	200.128078	198.146915	197.853330	197.549057
5	3.39	198.441879	196.523354	199.497979	200.291161	198.473750	198.203529	197.571257
6	4.85	197.184139	198.165439	201.048686	201.854425	199.824570	199.679565	199.120046
7	4.86	196.859786	198.826531	201.529807	202.255870	200.273334	199.982081	199.434848
8	4.88	196.321834	198.297739	201.044074	201.783543	199.766898	199.493749	198.952970
9	5.05	197.541736	198.187766	201.148605	201.989145	199.837582	199.697936	199.050610
10	5.17	196.721605	198.776474	201.389449	202.018938	200.291350	200.095951	199.822849
11	5.60	198.673928	198.618401	201.642914	202.688140	200.524452	200.426506	199.873110
12	5.70	197.056020	199.012875	201.666119	202.366093	200.573698	200.405929	199.942684
13	5.82	197.433857	199.149922	201.919099	202.576136	200.763649	200.526754	200.072279
14	5.87	200.020069	200.894281	204.739112	204.935278	203.178968	202.838327	202.302087
15	5.97	198.757867	199.336643	202.092071	202.842371	200.984329	200.754089	200.265330
16	5.99	197.631772	199.581448	202.237806	202.958036	201.150456	200.906799	200.503295
17	6.00	198.646376	198.175052	201.108307	201.846684	200.175755	200.125761	199.693576
18	6.02	197.543085	199.715160	202.301734	203.003398	201.279155	201.094776	200.691385
19	6.45	198.693371	198.790957	201.644842	202.553870	200.998628	200.788467	200.433444
20	6.62	199.321102	200.454591	203.028456	203.776849	201.947559	201.724978	201.374327
21	7.05	199.899897	199.565994	202.067651	202.817369	201.246355	201.182517	200.786681
22	7.75	202.422789	201.220654	203.940811	204.660141	203.136484	203.018673	202.567536

Table A6.9.: Acid dissociation free energies calculated for the R-OH subset of TR224. Various levels of theory are shown. All DFT methods employ COSMO-RS implicit solvation, GFN2-xTB uses ALPB implicit solvation.

Nr.	exp. pK_a	GFN2-xTB	r^{2}SCAN-3c	B97-3c	$\frac{\Delta G'_{diss}}{RT\ln(10)}$ B97-D	PBE0-D3	PW6B95-D3	ωB97X-V
23	11.60	209.037950	212.153040	212.827748	215.350180	214.700359	214.272191	214.132604
24	12.02	203.360314	209.041596	210.596351	212.344186	212.344332	212.290220	212.446469
25	12.43	205.591053	210.084215	211.232873	213.415284	212.838167	212.649483	212.477988
26	13.55	213.958350	210.499430	211.715152	213.443423	212.844100	213.060691	212.711728
27	14.80	212.755883	213.819875	213.857787	216.375618	216.161466	216.214021	215.977647
28	14.90	206.779384	210.887787	212.107666	213.578032	212.839351	213.880197	213.329939
29	15.10	207.896957	210.749159	211.682885	213.226716	212.567223	213.378013	212.944860
30	15.40	213.753676	213.389716	214.388939	216.399404	215.775155	215.873375	215.720534
31	15.52	214.508838	213.791948	214.627761	216.764668	216.185507	216.340681	216.112463
32	15.54	213.739857	215.114006	216.115700	217.298355	217.059246	217.216412	216.959534
33	15.90	214.702139	215.426546	215.627909	217.537746	217.365908	217.460453	217.239901
34	16.10	214.632956	215.172641	215.331292	217.516034	217.238437	217.338072	217.048762
35	16.84	216.691498	216.185789	216.064294	218.679743	218.370962	218.326296	217.985604
36	17.00	219.416616	217.732738	217.666263	220.146252	219.760677	219.638594	219.403941
37	17.10	217.044434	216.417537	216.357809	218.637530	218.352796	218.357526	218.125359

Table A6.10.: Acid dissociation free energies calculated for the R-SH subset of TR224. Various levels of theory are shown. All DFT methods employ COSMO-RS implicit solvation, GFN2-xTB uses ALPB implicit solvation.

Nr.	exp. pK_a	GFN2-xTB	r^{2}SCAN-3c	B97-3c	$\frac{\Delta G'_{diss}}{RT\ln(10)}$ B97-D	PBE0-D3	PW6B95-D3	ωB97X-V
38	7.86	201.513422	197.204107	200.535311	203.149968	200.376856	201.282013	200.487229
39	7.95	200.744478	197.560197	201.032557	203.610363	200.981917	201.772230	201.146538
40	8.62	197.982347	199.789849	202.831003	205.398865	202.862854	203.910627	203.192934
41	9.38	199.795227	199.388646	202.641670	205.149047	202.546244	203.468691	202.636618
42	9.72	200.321726	199.403833	202.474235	205.071624	202.498721	203.288284	202.776414
43	9.85	203.505379	200.171414	203.137848	205.511524	202.784151	203.496539	203.073069
44	9.96	202.798602	199.514601	202.934889	205.440851	202.889813	203.859737	203.115982
45	10.27[a]	192.480106	195.384472	196.712235	198.982670	197.909141	197.695733	197.232550
46	10.33	202.177916	199.844077	203.075787	205.626446	203.226642	204.230034	203.539655
47	10.61	202.863657	200.188027	203.365549	205.854229	203.360012	204.306277	203.581392
48	10.67	202.560474	200.169139	203.484169	205.886284	203.351077	204.340194	203.448571
49	10.86	204.749428	200.931692	204.121264	206.458184	203.893087	204.787613	203.987366
50	11.05	206.692815	201.742469	205.029854	207.149297	204.527850	205.318580	204.514290
51	11.22	206.144958	202.106424	205.457969	207.503475	204.912029	205.715085	204.842760

[a] Alternative value found in the literature: 4.34

Table A6.11.: Acid dissociation free energies calculated for the R-NH_2 subset of TR224. Various levels of theory are shown. All DFT methods employ COSMO-RS implicit solvation, GFN2-xTB uses ALPB implicit solvation.

Nr.	exp. pK_a	GFN2-xTB	r^{2}SCAN-3c	B97-3c	$\frac{\Delta G'_{diss}}{RT\ln(10)}$ B97-D	PBE0-D3	PW6B95-D3	ωB97X-V
52	5.30	194.662725	195.110667	199.203321	198.511168	196.958901	196.647033	196.658606
53	9.34	200.736907	201.698589	205.778403	205.255248	203.562330	203.209171	203.054539
54	9.68	200.812020	201.875349	205.857766	205.373319	203.877945	203.545827	203.442054
55	9.80	202.076768	203.193613	207.069401	206.629577	204.941020	204.492905	204.287754
56	10.59	200.899498	203.124429	206.910425	206.458417	204.915422	204.536721	204.381489
57	10.60	200.942397	203.270006	207.041604	206.566662	205.049204	204.672601	204.540834
58	10.63	200.950249	203.857769	207.535744	206.913660	205.538895	205.199720	205.158102
59	10.68	202.832649	203.215633	207.199354	206.760641	205.015583	204.493889	204.248799
60	10.70	201.024693	203.361598	207.090920	206.576482	205.093298	204.717720	204.587262
61	11.23	201.674387	202.720452	206.812209	206.422595	204.656995	204.212218	203.901430

Table A6.12.: Acid dissociation free energies calculated for the R2-NH subset of TR224. Various levels of theory are shown. All DFT methods employ COSMO-RS implicit solvation, GFN2-xTB uses ALPB implicit solvation.

Nr.	exp. pK_a	GFN2-xTB	r^{2}SCAN-3c	B97-3c	$\frac{\Delta G'_{diss}}{RT\ln(10)}$ B97-D	PBE0-D3	PW6B95-D3	ωB97X-V
62	8.55	201.849709	202.414582	206.003067	205.920654	203.974036	203.623884	203.580596
63	10.54	201.227946	204.194754	207.610055	207.502593	205.651774	205.363595	205.331964
64	10.78	200.799674	204.185214	207.516458	207.324669	205.644248	205.396289	205.475774
65	11.00	200.836799	203.794234	207.348991	207.405619	205.280868	204.960347	204.751460
66	11.02	201.179214	203.867257	207.354168	207.319584	205.261270	204.981375	204.821812
67	11.22	200.934314	204.231320	207.783461	207.782240	205.819870	205.442119	205.359229
68	11.23	200.904716	203.697764	207.372827	207.507393	205.251683	204.971884	204.616073
69	11.27	200.789471	204.787758	208.244691	208.245968	206.310707	206.043943	205.956774

Table A6.13.: Acid dissociation free energies calculated for the R3-N subset of TR224. Various levels of theory are shown. All DFT methods employ COSMO-RS implicit solvation, GFN2-xTB uses ALPB implicit solvation.

Nr.	exp. pK_a	GFN2-xTB	r^{2}SCAN-3c	B97-3c	$\frac{\Delta G'_{diss}}{RT\ln(10)}$ B97-D	PBE0-D3	PW6B95-D3	ωB97X-V
70	9.69	202.133797	202.541872	205.824312	206.353399	203.694107	203.436760	203.403418
71	9.69	202.174824	202.607752	205.889439	206.400243	203.734311	203.472000	203.443435
72	9.80	201.855754	203.584649	206.441310	206.708106	204.692030	204.513669	204.784778
73	10.16	202.046195	203.667295	206.533912	206.921077	204.747218	204.512621	204.692042
74	10.75	202.087589	204.171933	207.322477	207.683954	205.212442	205.023919	204.942692

Table A6.14.: Acid dissociation free energies calculated for the R-COOH subset of TR224. Various levels of theory are shown. All DFT methods employ COSMO-RS implicit solvation, GFN2-xTB uses ALPB implicit solvation.

Nr.	exp. pK_a	GFN2-xTB	r^{2}SCAN-3c	B97-3c	$\frac{\Delta G'_{diss}}{RT\ln(10)}$ B97-D	PBE0-D3	PW6B95-D3	ωB97X-V
75	-0.26	186.596023	189.584151	191.190858	193.438633	192.165648	191.749687	191.540387
76	0.65	187.217078	190.335127	192.228814	194.172398	193.221878	192.979023	192.817915
77	1.24	190.952639	191.503693	192.985112	195.283318	194.107245	193.765296	193.555324
78	1.30	190.268271	191.976924	193.714282	195.778907	194.686207	194.450449	194.185869
79	2.44	192.963036	193.677885	194.995400	196.791395	195.997114	195.827334	195.563628
80	2.66	195.186535	193.077929	194.341239	196.614303	195.739689	195.460850	195.159831
81	2.80	194.262459	193.903271	195.312008	197.435154	196.416501	196.224959	195.876856
82	2.81	194.102665	193.373608	194.826289	197.006985	196.254577	196.055042	195.924226
83	2.86	197.564805	194.943041	196.150567	198.250764	197.456045	197.350669	196.991985
84	2.86	192.344778	193.244529	194.886621	196.894354	195.988088	195.813211	195.688340
85	3.07	193.690388	194.165437	195.636043	197.808207	196.685958	196.357132	196.056207
86	3.53	194.041093	195.264270	196.336941	198.326374	197.743830	197.543932	197.420947
87	3.54	196.759069	193.806672	195.010781	197.457182	196.447698	196.257871	195.905411
88	3.75	195.042475	193.364806	194.416025	196.526145	195.620690	195.457764	195.309873
89	3.83	195.756453	194.341207	195.508787	197.436785	196.531545	196.279217	196.069921
90	3.87	196.412473	194.721480	195.883504	197.849754	196.746653	196.433137	196.264253
91	4.10	196.551424	195.598476	196.950613	199.115460	198.146830	197.897942	197.580953
92	4.26	198.818863	195.592421	196.923091	199.206608	198.028008	197.858047	197.351410
93	4.31	197.653624	196.159828	197.574794	199.764081	198.813777	198.639094	198.200909
94	4.35	197.805492	195.693588	197.072664	199.315251	198.299100	198.126832	197.741088
95	4.52	197.440755	196.064702	197.388578	199.621380	198.604710	198.403365	197.955825
96	4.76	198.994457	196.037076	197.172787	199.461770	198.478741	198.302242	197.890309
97	4.82	198.769803	196.466089	197.530190	199.815678	198.892160	198.714255	198.219475
98	4.87	199.450318	196.481917	197.638155	199.917711	198.963637	198.761792	198.321672
99	4.90	199.103301	196.892194	197.947430	200.239982	199.213837	198.983822	198.478865
100	5.05	199.595908	197.125912	198.416052	200.511514	199.514616	199.247195	198.814270

Table A6.15.: Acid dissociation free energies calculated for the Ph-SH subset of TR224. Various levels of theory are shown. All DFT methods employ COSMO-RS implicit solvation, GFN2-xTB uses ALPB implicit solvation.

Nr.	exp. pK_a	GFN2-xTB	r^{2}SCAN-3c	B97-3c	$\frac{\Delta G'_{diss}}{RT\ln(10)}$ B97-D	PBE0-D3	PW6B95-D3	ωB97X-V
101	4.72	197.033546	195.059912	198.265454	200.734142	198.561857	199.215724	199.156030
102	5.24	199.249799	195.553563	198.980015	201.498759	198.747066	199.508298	198.801346
103	5.30	197.301170	194.619619	198.036210	200.627683	197.829910	198.563839	197.791704
104	5.33	199.401909	195.891518	199.266352	201.755253	199.193037	199.955242	199.373430
105	5.78	198.849278	196.596119	200.028000	202.597159	199.855303	200.665167	199.809384
106	6.02	197.229989	196.910394	200.321703	202.841144	200.090615	200.926758	200.040078
107	6.14	199.716082	196.891931	200.293470	202.860053	200.084718	200.903387	200.023331
108	6.39	200.415895	197.056654	200.511001	203.122202	200.257067	201.114626	200.183583
109	6.61	200.970487	197.213209	200.621832	203.153199	200.347686	201.247112	200.342006
110	6.64	201.960016	198.309556	201.698159	204.247113	201.338604	202.174312	201.297490
111	6.66	201.461880	197.484687	200.949991	203.505544	200.661265	201.573652	200.636302
112	6.78	201.145213	197.339990	200.758663	203.336409	200.500669	201.383909	200.474405
113	6.82	201.361864	197.420360	200.848347	203.399052	200.567312	201.479556	200.550574

Table A6.16.: Acid dissociation free energies calculated for the Ph-OH subset of TR224. Various levels of theory are shown. All DFT methods employ COSMO-RS implicit solvation, GFN2-xTB uses ALPB implicit solvation.

Nr.	exp. pK_a	$\frac{\Delta G'_{diss}}{RT\ln(10)}$ GFN2-xTB	r^{2}SCAN-3c	B97-3c	B97-D	PBE0-D3	PW6B95-D3	ωB97X-V
114	6.79	200.909051	202.429487	204.152676	206.168918	205.486801	205.163560	205.060773
115	7.14	198.053483	198.836150	200.078007	201.997337	201.836901	201.733325	202.784103
116	7.23	198.102956	200.309466	202.169679	204.303421	203.573265	203.033084	203.460629
117	7.66	201.225078	200.399777	201.745554	203.623295	203.210437	203.289895	203.570396
118	7.95	202.430834	201.222793	202.643440	204.472998	203.854117	203.996638	204.025344
119	8.00	202.439794	203.354571	204.590906	206.544792	206.030941	206.147983	206.185827
120	8.35	200.314357	202.481770	203.742129	205.730118	205.162275	205.171284	205.291321
121	8.41	202.800699	201.983561	203.378937	205.364359	204.634294	204.716246	204.776999
122	8.47	202.901739	201.940665	203.326745	205.273611	204.614219	204.708433	204.755345
123	8.47	202.914083	202.054610	203.466496	205.454791	204.720244	204.809159	204.835410
124	8.48	200.732351	203.174012	204.787002	206.754617	205.918513	206.102403	205.846497
125	8.50	202.850103	202.178594	203.572808	205.563417	204.848121	204.905926	204.969904
126	8.61	201.455197	202.791743	204.152594	206.025486	205.419147	205.551595	205.495102
127	8.81	201.597007	203.270042	204.673838	206.866072	206.067836	206.172665	205.952450
128	9.02	203.784992	203.677294	205.053729	207.110087	206.394365	206.542248	206.326912
129	9.28	203.992254	204.014599	205.405886	207.384488	206.592663	206.734899	206.417153
130	9.38	203.913365	204.453703	205.760269	207.796545	207.020272	207.181536	206.896255
131	9.39	202.892311	201.999086	203.391296	205.338959	204.753884	204.816511	204.846531
132	9.44	206.139898	204.530905	205.981492	208.104759	207.115323	207.355594	206.901685
133	9.48	203.951068	203.067744	204.317215	206.174159	205.347011	205.492717	205.414120
134	9.51	207.136961	204.675435	206.060637	208.128706	207.258563	207.435002	207.397142
135	9.59	206.458170	204.815467	206.259067	208.454417	207.547574	207.786430	207.311187
136	9.65	206.185014	204.195202	205.671254	207.826849	206.829987	207.063988	206.615375
137	9.82	206.908367	204.360411	205.623428	207.736508	207.049218	207.288901	207.084507
138	9.83	206.280805	204.400741	205.741758	207.914112	207.225448	207.493145	207.101374
139	9.90	206.994444	204.624572	206.064881	208.257478	207.347454	207.676699	207.214600
140	9.92	203.335681	203.763221	205.126796	206.917016	205.854349	206.200276	205.905730
141	9.93	206.873409	205.405143	207.062397	208.951355	207.994706	208.222821	208.043097
142	9.93	207.177091	203.981825	205.403351	207.770903	206.908727	207.165891	206.855512
143	9.94	203.319712	203.950759	205.250647	207.300027	206.560739	206.654343	206.543176
144	9.95	204.033210	205.189412	206.363167	208.491469	207.793910	207.920622	207.707362
145	9.96	205.767361	205.888688	207.114265	209.308363	208.587830	208.815202	208.491318
146	9.98	207.871925	204.563042	205.912840	208.064297	207.175833	207.467089	207.168639
147	10.00	208.321447	205.102586	206.434088	208.649647	207.750016	208.030858	207.731961
148	10.08	208.157009	204.742900	206.194117	208.366908	207.418359	207.718866	207.327010
149	10.19	207.458478	205.455379	206.760528	208.965126	208.096394	208.378681	208.085396
150	10.20	206.108328	205.593426	206.814740	209.087135	208.292721	208.541468	207.971787
151	10.20	207.149732	205.619428	207.149234	209.035823	208.078584	208.314010	208.120948
152	10.28	207.359504	205.318367	206.773883	208.658450	207.794087	208.053897	207.791171

Table A6.17.: Acid dissociation free energies calculated for the Ph-NH_2 subset of TR224. Various levels of theory are shown. All DFT methods employ COSMO-RS implicit solvation, GFN2-xTB uses ALPB implicit solvation.

Nr.	exp. pK_a	$\frac{\Delta G'_{diss}}{RT\ln(10)}$ GFN2-xTB	r^{2}SCAN-3c	B97-3c	B97-D	PBE0-D3	PW6B95-D3	ωB97X-V
153	0.28	186.352434	185.918123	190.501609	190.092091	188.573829	188.203735	189.819965
154	0.98	187.614365	187.339686	191.611497	191.136015	190.222450	190.032994	192.376690
155	2.04	191.657059	191.023414	195.888287	195.319056	193.530950	193.067341	193.933827
156	2.10	191.858795	191.428060	196.298579	195.792173	193.868773	193.466474	194.206269
157	2.16	191.510344	191.022385	195.875417	195.355095	193.488552	193.089035	193.860989
158	2.30	190.689464	189.693007	194.277294	193.496736	191.859737	191.742566	192.929457
159	2.32	190.415432	189.828727	194.399550	193.618925	192.065918	191.934243	193.117128
160	2.38	190.842596	189.956817	194.560882	193.827607	192.123648	192.001087	193.131425
161	2.45	189.554222	190.219545	194.654795	194.078453	192.443135	192.204487	193.122634
162	2.62	194.690857	190.812244	195.479375	195.062860	193.158640	192.930137	193.253127
163	2.96	191.232249	192.055582	196.542434	195.953707	194.152998	193.881475	194.302432
164	3.05	192.691402	192.564900	196.999123	196.455229	194.658868	194.467920	195.098420
165	3.32	194.033200	192.464149	196.990180	196.456369	194.591375	194.415943	194.841082
166	3.38	191.820806	192.645312	197.168956	196.547337	194.668107	194.450409	194.783292
167	3.56	192.531694	192.728222	197.163498	196.618027	194.804096	194.599016	195.223503
168	3.78	194.762281	193.451676	198.140076	197.642821	195.567227	195.355253	195.626544
169	3.81	194.220926	193.165135	197.593920	197.046057	195.264519	195.091429	195.431216
170	4.05	194.298369	193.848024	198.415374	197.923407	195.873345	195.724966	195.801258
171	4.17	194.953193	194.430868	198.952124	198.413019	196.380197	196.175725	196.262374
172	4.17	194.684914	194.278093	198.797094	198.335221	196.237144	196.004869	196.118692
173	4.20	194.753281	194.150250	198.693086	198.178903	196.095540	195.873846	196.008273
174	4.36	194.775272	194.044623	198.533093	197.931533	195.964487	195.791857	196.169753
175	4.38	195.822700	194.176737	198.678079	198.143214	196.193666	195.971592	196.275742
176	4.40	194.378135	194.323279	198.646092	198.299945	196.328950	196.274743	195.960799
177	4.47	196.700665	194.654769	199.115231	198.724586	196.756076	196.508005	196.823669
178	4.49	196.405521	194.523212	198.991885	198.549811	196.701224	196.455912	196.719847
179	4.52	193.037499	194.317726	198.594989	198.085386	196.366146	196.104195	196.454543
180	4.64	195.957433	194.502122	198.974994	198.420989	196.526243	196.363521	196.621676
181	4.67	195.089093	194.691317	199.156655	198.563246	196.642274	196.497927	196.725906
182	4.72	196.971706	195.059484	199.515060	199.076705	197.182708	196.905277	197.162978
183	5.07	195.168492	195.176196	199.518231	198.940599	197.144803	196.989605	197.237486
184	5.25	195.691837	196.133112	200.435560	200.027464	198.044452	197.771375	198.047211
185	5.29	195.488759	196.002540	200.296759	199.873124	197.916370	197.629734	197.930709
186	5.50	195.497038	196.214509	200.477383	200.044710	198.220842	197.941816	198.126742

Table A6.18.: Acid dissociation free energies calculated for the Ph-COOH subset of TR224. Various levels of theory are shown. All DFT methods employ COSMO-RS implicit solvation, GFN2-xTB uses ALPB implicit solvation.

Nr.	exp. pK_a	GFN2-xTB	r^{2}SCAN-3c	B97-3c	$\frac{\Delta G'_{diss}}{RT\ln(10)}$ B97-D	PBE0-D3	PW6B95-D3	ωB97X-V
187	2.17	195.593567	192.544080	194.131094	196.380730	195.259439	195.007026	194.847357
188	2.94	195.895702	194.479705	195.960207	198.027235	197.128882	197.124409	196.240295
189	2.95	195.535872	192.550045	194.496704	195.378276	193.762134	195.139789	194.659183
190	2.98	197.099556	194.635076	196.275048	198.212432	196.857458	196.905173	196.162863
191	3.27	196.440701	195.249445	196.545804	198.654662	197.819287	197.692827	196.866082
192	3.45	196.986370	194.750711	196.392409	198.394074	197.156717	196.943227	196.468812
193	3.46	197.899793	195.596055	197.201981	199.231243	198.111047	197.939219	197.366377
194	3.51	197.735929	195.161528	196.816027	198.863780	197.601974	197.401534	196.825259
195	3.54	197.999338	195.409908	197.044377	199.104897	197.823876	197.614199	197.061466
196	3.77	198.091297	196.174383	197.357330	199.462400	198.508515	198.560171	197.610272
197	3.83	198.036197	195.533359	197.172165	199.274165	197.984870	197.776183	197.187298
198	3.87	197.908003	195.655625	197.264981	199.342779	198.052238	197.838347	197.211545
199	3.91	198.032368	195.904056	197.068309	199.109113	198.344541	198.451497	197.348222
200	3.99	198.729355	195.740632	197.363103	199.485730	198.166584	197.967542	197.292710
201	4.08	198.641259	196.229263	197.855069	199.980228	198.638817	198.428678	197.707353
202	4.09	198.361167	195.494009	196.621369	198.882721	198.042130	198.077713	196.825011
203	4.09	198.685519	196.067369	197.698475	199.820042	198.471160	198.262409	197.540213
204	4.14	198.520598	196.021934	197.596123	199.710525	198.446139	198.239135	197.576068
205	4.17	198.571529	196.062118	197.691920	199.823306	198.456048	198.253627	197.525179
206	4.21	198.001872	195.205792	196.353425	198.614193	197.767267	197.830152	196.474136
207	4.24	199.191224	196.340598	197.948618	200.066568	198.745040	198.567333	197.876445
208	4.34	199.100048	196.484473	198.044803	200.188864	198.864235	198.677088	197.986557
209	4.35	199.034183	196.381257	198.007029	200.104997	198.792321	198.611338	197.894593
210	4.35	199.051093	196.525333	198.106532	200.226503	198.928613	198.759182	198.032216
211	4.45	199.357707	196.916548	198.480738	200.682125	199.255111	199.039896	198.270377
212	4.47	199.272296	196.758994	198.331906	200.515455	199.129737	198.924348	198.146680
213	4.58	199.251745	196.945246	198.494937	200.688812	199.314401	199.096561	198.315641
214	4.92	197.224953	194.686107	196.360924	198.378938	197.074146	196.834805	196.394891

Table A6.19.: Acid dissociation free energies calculated for the R-C-H subset of TR224. Various levels of theory are shown. All DFT methods employ COSMO-RS implicit solvation, GFN2-xTB uses ALPB implicit solvation.

Nr.	exp. pK_a	GFN2-xTB	r^{2}SCAN-3c	B97-3c	$\frac{\Delta G'_{diss}}{RT\ln(10)}$ B97-D	PBE0-D3	PW6B95-D3	ωB97X-V
215	9.90	201.812041	203.531214	206.613864	207.523152	205.575771	205.944535	209.056243
216	10.40	199.331032	204.162891	206.424802	207.738451	206.655999	206.776807	209.759801
217	10.50	196.584359	204.004173	207.327558	208.412228	206.096227	206.579712	209.494424
218	11.50	198.202156	203.758310	206.988746	208.067598	205.839274	206.195696	209.508460
219	13.10	201.085907	205.600759	208.132871	209.079514	207.965298	208.390748	210.695094
220	14.80	210.016244	208.569445	210.609520	211.853374	210.759149	211.085392	213.417162
221	16.70	206.453168	209.447682	212.004739	212.732900	211.711031	212.118604	213.081832
222	18.30	210.682116	213.011765	215.893633	216.756786	215.176947	215.576407	215.972555
223	19.20	212.059500	214.049231	216.946529	217.760383	216.091307	216.488731	216.838420
224	19.30	211.829018	213.164619	215.708822	216.558276	215.306788	215.615683	216.385332

Non-rigid Molecules

Table A6.20.: CFER parameters determined for r^{2}SCAN-3c/COSMO-RS, GFN2-xTB+E_{mod}^{TB}/ALPB and B97-3c/COSMO-RS from the drug and SAMPL6 set.

	CFER parameters			
	c_1	c_2	c_3	c_4
r^{2}SCAN-3c	37659.112434	-568.698226	2.858687	-0.004782
GFN2-xTB	12284.084912	-186.207085	0.938732	-0.001573
B97-3c	34751.943849	-515.624632	2.546401	-0.004185

Drug Benchmark Set

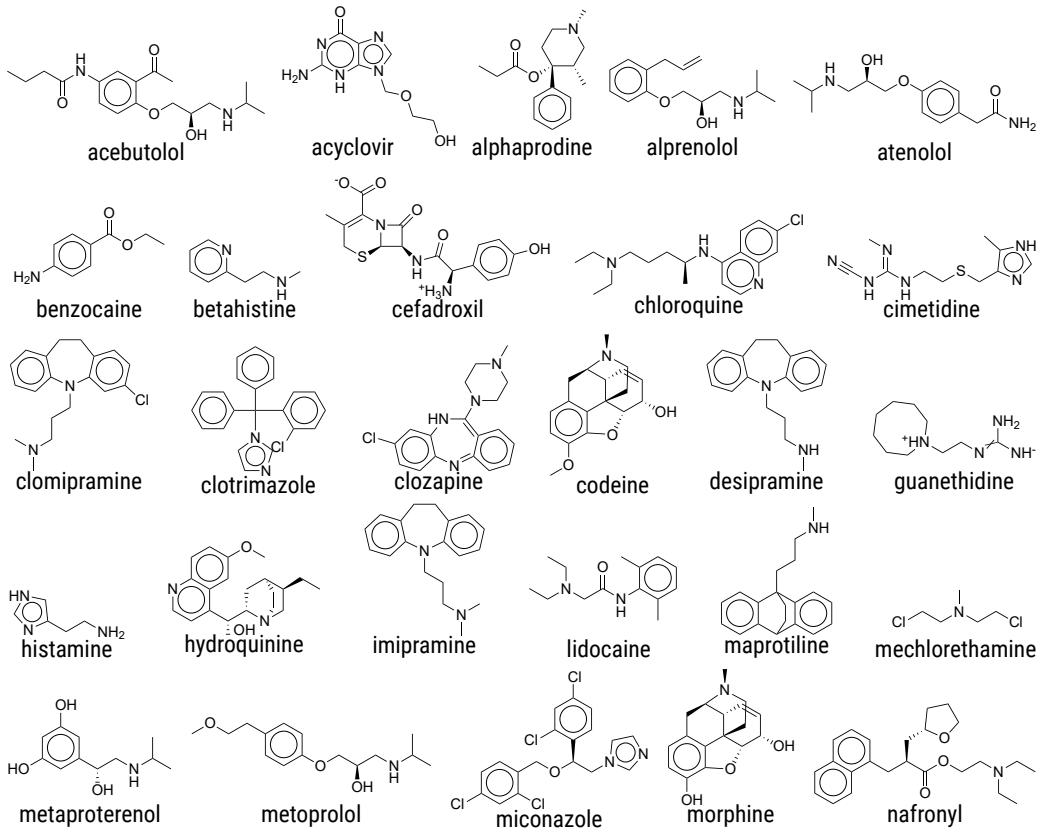

Figure A6.1.: Investigated drug molecules in their neutral protonation state (pt. 1)

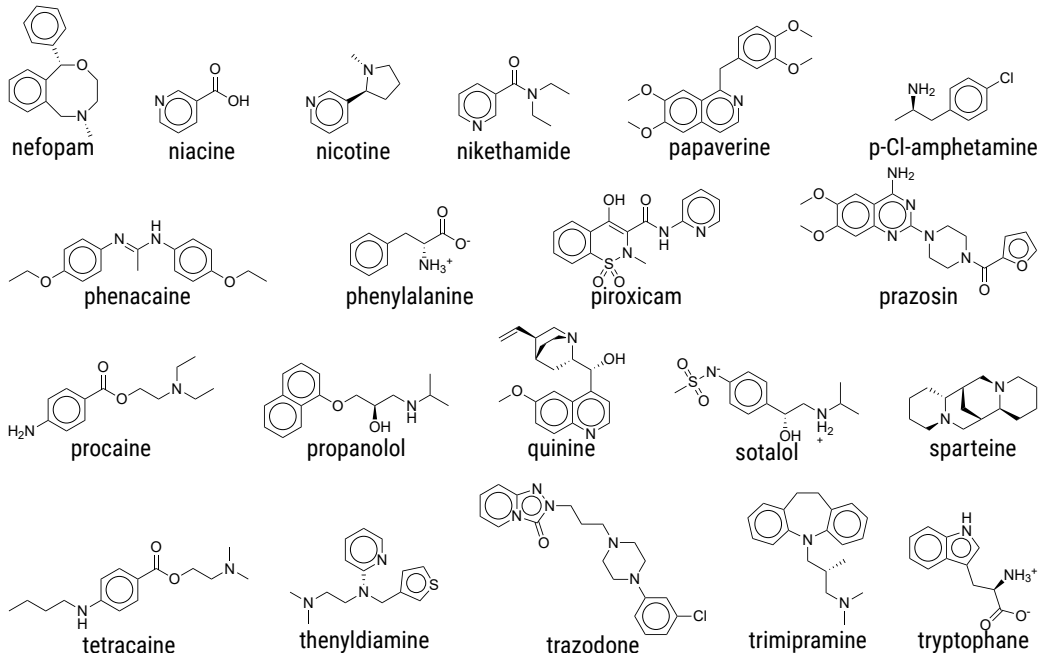

Figure A6.2.: Investigated drug molecules in their neutral protonation state (pt. 2)

Table A6.21.: Acid dissociation free energies for the flexible drug benchmark set at the r^{2}SCAN-3c/COSMO-RS, GFN2-xTB+E_{mod}^{TB}/ALPB and B97-3c/COSMO-RS level.

			$\frac{\Delta G'_{diss}}{RT\ln(10)}$		
		exp. pK_a	r^{2}SCAN-3c	GFN2-xTB	B97-3c
1	acebutolol	9.50	203.998707	205.860197	207.372498
2	acyclovir	2.20	193.537952	194.735579	196.478811
3	alphaprodine	8.70	201.323400	200.604678	204.752763
4	alprenolol	9.60	203.023857	202.389308	206.796334
5	atenolol	9.60	202.228354	202.096999	205.836038
6	benzocaine	2.50	190.092551	190.332008	194.721840
7	betahistine	10.00	203.299444	205.772887	206.920214
8	betahistine+	3.90	195.966529	191.969840	198.939911
9	cefadroxil-	7.00	198.919657	196.614623	—
10	chloroquine	10.60	202.416045	205.017334	205.689354
11	cimetidine0	6.80	207.180120[a]	198.776353	209.662065[a]
12	clomipramine	9.40	202.024218	202.836156	205.279785
13	clotrimazole	5.80	198.189084	197.644979	201.090956
14	clozapine	7.50	197.981240	198.344544	201.161951
15	clozapine+	3.90	197.786666	199.394335	201.059971
16	codeine	8.10	200.208203	199.190383	203.789814
17	desipramine	10.30	202.842364	200.848801	206.500956
18	guanethidine	11.40	206.652562	207.892505	209.163574
19	histamine	9.70	202.509310	205.919226	206.529420
20	hydroquinine	9.10	202.078423	200.093147	205.810353
21	hydroquinine+	4.10	197.535923	197.187386	200.556105
22	imipramine	9.60	202.094964	202.680539	205.520923
23	labetalol	9.30	201.255348	202.119568	205.264672
24	lidocaine	7.90	199.315283	200.462297	202.126331
25	maprotiline	10.30	202.912397	201.173762	206.569462
26	mechlorethamine	6.40	199.659045	196.842763	203.082943
27	metaproterenol	9.90	202.157590	202.727757	205.888678
28	metoprolol	9.60	202.354476	202.324562	205.981459
29	miconazole	6.40	198.888428	197.999825	201.487964
30	morphine	8.20	200.089781	199.087744	203.698130
31	nafronyl	9.10	201.368051	203.486634	204.838803
32	nefopam	8.50	201.101947	199.896133	204.512113
33	niacine-	4.80	198.139303	196.226316	200.727399
34	nicotine	8.10	199.950574	199.217766	203.361273
35	nicotine+	4.80	196.049681	194.425384	198.801962
36	nikethamide	3.50	196.261462	194.280926	199.013279
37	papaverine	8.07	200.302895	200.118983	203.176066
38	p-Cl-amphetamine	9.90	201.462146	202.605859	205.622131
39	phenacaine	9.30	203.342334	201.367618	206.335734
40	phenylalanine-	8.90	200.257527	201.986584	204.182521
41	piroxicam	5.30	196.269551	196.566943	199.272837
42	prazosin	7.00	200.412901	195.957535	203.652849
43	procaine	9.10	202.487330	202.545609	205.990229
44	procaine+	2.00	189.181533	188.538912	193.668393
45	propanolol	9.60	202.004559	206.336238	205.704995
46	quinine	8.50	201.399294	199.821850	205.092518
47	sotalol	9.30	201.737329	203.085999	205.437554
48	sparteine	12.00	207.046798	208.899822	210.762596
49	tetracaine	8.50	202.331854	201.834995	205.640921
50	thenyldiamine	8.90	203.381151	203.379449	207.017171
51	trazodone	6.80	198.663543	204.212586	203.128758
52	trimipramine	9.40	201.964598	198.617859	205.630875
53	tryptophan-	9.10	201.749066	202.988346	205.649437

[a] neglected outlier

Table A6.22.: Minimum and maximum acid dissociation constants calculated for the flexible drug benchmark set at the B97-3c/COSMO-RS, r^{2}SCAN-3c/COSMO-RS and GFN2-xTB+E_{mod}^{TB}/ALPB level. All pK_a values in this table were obtained using the CFER parameters fitted on the TR224 set.

			r^{2}SCAN-3c		GFN2-xTB		B97-3c	
		exp. pK_a	min. pK_a	max. pK_a	min. pK_a	max. pK_a	min. pK_a	max. pK_a
1	acebutolol	9.50	8.93	11.60	7.26	14.13	9.45	12.57
2	acyclovir	2.20	2.42	4.53	1.75	6.96	2.43	4.84
3	alphaprodine	8.70	7.63	9.08	3.84	11.07	8.47	9.51
4	alprenolol	9.60	8.22	10.85	5.47	12.19	9.06	12.04
5	atenolol	9.60	7.71	10.31	4.75	12.31	8.57	11.35
6	benzocaine	2.50	1.33	2.00	0.94	1.94	2.06	2.79
7	betahistine	10.00	9.49	10.42	10.62	13.06	10.31	11.26
8	betahistine+	3.90	4.07	4.94	1.28	2.70	4.26	5.18
9	cefadroxil	7.00	6.28	7.56	1.79	10.71	—	—
10	chloroquine	10.60	7.79	10.39	7.35	13.85	8.25	10.98
11	cimetidine	6.80	11.32[a]	13.06[a]	3.52	9.82	11.6[a]	13.39[a]
12	clomipramine	9.40	7.65	10.05	5.84	12.87	8.18	10.47
13	clotrimazole	5.80	5.46	6.74	4.61	6.83	5.65	6.93
14	clozapine	7.50	5.74	6.52	3.70	8.57	6.09	6.89
15	clozapine+	3.90	4.78	6.31	4.07	9.03	5.30	6.80
16	codeine	8.10	6.81	8.66	4.27	9.15	7.49	9.38
17	desipramine	10.30	7.92	10.40	4.72	11.83	8.78	11.40
18	guanethidine	11.40	10.68	12.96	8.67	14.89	10.78	13.69
19	histamine	9.70	8.86	9.89	10.97	12.81	9.96	11.26
20	hydroquinine	9.10	7.64	10.11	4.10	10.42	8.67	11.11
21	hydroquinine+	4.10	4.73	6.96	2.59	9.40	4.97	7.13
22	imipramine	9.60	7.80	9.81	5.77	12.95	8.33	10.42
23	labetalol	9.30	7.24	9.63	5.56	12.06	8.16	10.91
24	lidocaine	7.90	6.10	7.86	3.99	10.55	6.27	8.10
25	maprotiline	10.30	7.98	10.76	4.32	11.78	8.84	11.59
26	mechlorethamine	6.40	5.95	8.23	2.21	6.11	6.45	8.62
27	metaproterenol	9.90	7.63	10.10	5.53	12.01	8.54	11.31
28	metoprolol	9.60	7.81	10.39	5.03	12.12	8.53	11.62
29	miconazole	6.40	5.75	7.77	3.37	11.44	5.30	8.08
30	morphine	8.20	6.96	8.57	4.17	8.95	7.61	9.26
31	nafronyl	9.10	7.29	9.67	6.36	12.33	7.92	10.39
32	nefopam	8.50	7.37	9.27	4.15	9.68	7.90	10.21
33	niacine	4.80	6.18	6.18	4.57	4.57	6.16	6.16
34	nicotine	8.10	7.09	8.85	4.33	9.26	7.65	9.58
35	nicotine+	4.80	3.61	5.18	2.17	5.57	3.53	5.27
36	niketham ide	3.50	4.22	5.69	1.59	6.46	4.30	5.78
37	papaverine	8.07	7.19	8.74	4.91	10.67	7.48	8.91
38	p-Cl-amphetamine	9.90	7.74	9.10	5.98	12.26	8.91	10.39
39	phenacaine	9.30	8.34	10.76	5.93	10.65	8.74	11.58
40	phenylalanine	8.90	7.26	8.54	6.04	10.38	8.19	9.52
41	piroxicam	5.30	4.92	5.68	2.50	7.23	5.16	6.15
42	prazosin	7.00	6.85	8.38	1.77	9.38	7.29	8.94
43	procaine	9.10	7.93	10.54	4.73	12.42	8.42	11.24
44	procaine+	2.00	0.58	2.12	0.93	3.40	0.86	3.09
45	propanolol	9.60	7.27	10.27	4.85	14.01	8.13	11.28
46	quinine	8.50	7.16	9.68	4.14	9.84	8.00	10.67
47	sotalol	9.30	7.36	10.04	5.42	12.29	8.11	11.25
48	sparteine	12.00	12.24	12.24	10.61	15.57	13.33	13.33
49	tetracaine	8.50	7.71	10.17	4.55	10.70	8.12	10.70
50	thenyldiamine	8.90	8.63	10.78	5.85	12.07	9.50	11.73
51	trazodone	6.80	5.46	7.73	5.80	13.21	6.32	9.08
52	trimipramine	9.40	7.88	9.98	3.47	13.01	8.41	10.91
53	tryptophane	9.10	7.69	9.82	6.88	12.45	8.61	10.98
	MAD	—	1.23	0.92	3.25	2.58	0.82	1.57
	RMSD	—	1.40	1.13	3.57	2.92	0.96	1.73

[a] neglected outlier

SAMPL6 Benchmark Set

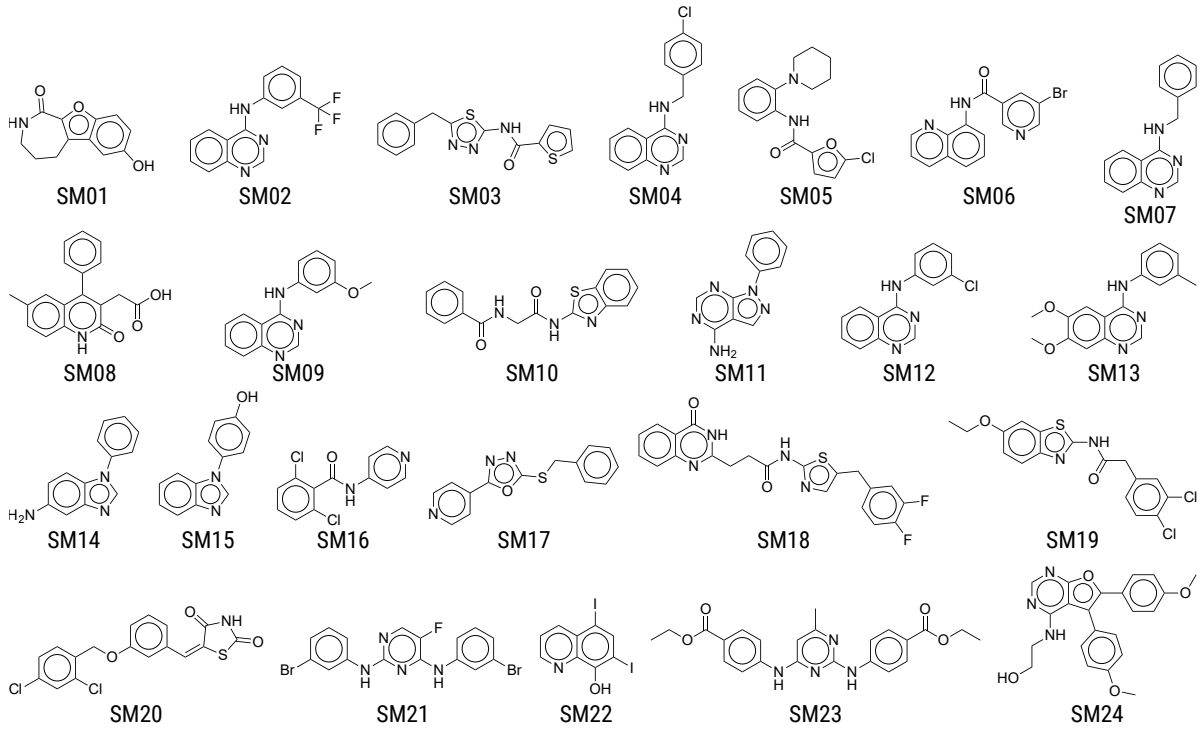

Figure A6.3.: Investigated SAMPL6 molecules in their neutral protonation state.

Table A6.23.: Acid dissociation free energies for the SAMPL6 benchmark set at the r^{2}SCAN-3c/COSMO-RS, GFN2-xTB+E_{mod}^{TB}/ALPB and B97-3c/COSMO-RS level.

			$\frac{\Delta G'_{diss}}{RT\ln(10)}$		
		exp. pK_a	r^{2}SCAN-3c	GFN2-xTB	B97-3c
1	SM01-	9.53	204.873812	205.840476	206.124628
2	SM02	5.03	198.268219	196.235418	201.237057
3	SM03-	7.02	199.798406	192.099444	202.804185
4	SM04	6.02	198.914314	198.113983	201.700387
5	SM05	4.59	195.792115	187.282571	199.433817
6	SM06	3.03	193.373074	193.566269	196.514900
7	SM06-	11.74	207.109076	206.161299	209.349298
8	SM07	6.08	199.083159	197.851320	201.993863
9	SM08-	4.22	195.583327	193.591560	197.612225
10	SM09	5.37	198.589074	196.967774	201.548259
11	SM10-	9.02	201.601089	192.243970	204.497921
12	SM11	3.89	195.432361	192.575935	198.262989
13	SM12	5.28	198.181950	197.073423	201.130245
14	SM13	5.77	199.453897	197.661042	202.278515
15	SM14+	2.58	190.727305	189.962271	195.137044
16	SM14	5.30	198.753585	200.324025	201.776296
17	SM15	4.70	197.245681	195.917449	200.248178
18	SM15-	8.94	204.140073	205.221013	205.506767
19	SM16	5.37	198.551263	198.524350	201.257822
20	SM16-	10.65	204.243065	199.121457	206.322877
21	SM17	3.16	196.602559	195.948580	199.256822
22	SM18	2.15	192.568925	191.328006	196.596531
23	SM18_2	9.58	203.343184	203.755478	204.765284
24	SM18-	11.02	203.901229	193.945662	206.037370
25	SM19-	9.56	203.134766	196.349667	205.916339
26	SM20-	5.70	198.744610	188.908059	201.788779
27	SM21	3.86	195.570797	189.335496	198.671954
28	SM22	2.40	193.626473	187.303665	196.667554
29	SM22-	7.43	201.186391	200.328920	203.166492
30	SM23	4.52	196.821424	195.738072	199.830464
31	SM24	2.60	193.653266	192.941149	195.872067

Table A6.24.: Minimum and maximum acid dissociation constants calculated for the SAMPL6 benchmark set at the B97-3c/COSMO-RS, r^{2}SCAN-3c/COSMO-RS and GFN2-xTB+E_{mod}^{TB}/ALPB level. All pK_a values in this table were obtained using the CFER parameters fitted on the TR224 set.

			r^{2}SCAN-3c		GFN2-xTB		B97-3c	
		exp. pK_a	min. pK_a	max. pK_a	min. pK_a	max. pK_a	min. pK_a	max. pK_a
1	SM01-	9.53	10.78	10.82	9.82	13.95	9.96	9.99
2	SM02	5.03	6.13	7.33	2.77	5.93	6.24	7.33
3	SM03-	7.02	6.32	8.01	1.64	3.24	6.52	8.62
4	SM04	6.02	6.37	7.01	3.73	8.58	6.38	7.26
5	SM05	4.59	4.17	5.54	0.93	2.06	4.76	6.28
6	SM06	3.03	2.54	3.46	1.41	4.30	2.93	3.66
7	SM06-	11.74	11.97	13.17	9.82	12.87	12.01	14.05
8	SM07	6.08	6.46	7.06	3.69	8.02	6.73	7.37
9	SM08-	4.22	4.21	4.76	1.64	3.43	3.37	4.37
10	SM09	5.37	5.47	7.39	3.15	7.20	5.97	7.41
11	SM10-	9.02	7.83	9.37	1.28	6.05	7.96	9.54
12	SM11	3.89	4.44	4.44	2.61	2.61	4.53	4.53
13	SM12	5.28	5.96	7.15	3.34	7.31	6.20	7.14
14	SM13	5.77	6.57	7.42	3.28	7.58	7.00	7.54
15	SM14+	2.58	1.86	1.93	1.18	2.12	2.59	2.69
16	SM14	5.30	6.50	7.02	6.34	8.38	6.78	7.23
17	SM15	4.70	5.28	5.90	3.60	5.28	5.54	6.11
18	SM15-	8.94	10.05	10.38	9.58	10.75	9.32	9.61
19	SM16	5.37	6.35	7.57	5.36	8.32	6.38	7.21
20	SM16-	10.65	9.19	11.48	4.75	10.02	9.41	11.78
21	SM17	3.16	4.65	5.56	2.65	6.36	4.78	5.56
22	SM18	2.15	2.13	3.68	0.93	5.50	2.18	4.42
23	SM18_2	9.58	8.78	10.61	5.07	12.23	7.97	10.61
24	SM18-	11.02	9.16	10.98	1.08	7.94	8.85	11.00
25	SM19-	9.56	8.34	10.77	1.99	6.79	8.50	10.98
26	SM20-	5.70	5.56	7.50	0.93	3.36	5.90	7.89
27	SM21	3.86	3.98	5.52	0.94	2.59	4.12	5.94
28	SM22	2.40	3.38	3.38	1.02	1.02	3.54	3.54
29	SM22-	7.43	8.26	8.26	7.27	7.27	7.85	7.85
30	SM23	4.52	4.36	6.59	2.68	7.24	4.57	6.91
31	SM24	2.60	2.66	4.12	1.20	5.33	2.32	4.04
	MAD	—	0.71	1.27	2.72	2.08	0.73	1.37
	RMSD	—	0.86	1.41	3.61	2.33	0.91	1.55

A7. List of 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:

$$\overline{x} = \frac{1}{n}\sum_{i}^{n} x_i \tag{A7.1}$$

- Mean deviation (MD):

$$MD = \frac{1}{n}\sum_{i}^{n}(x_i - r_i) \tag{A7.2}$$

- Mean absolute deviation (MAD):

$$MAD = \frac{1}{n}\sum_{i}^{n}|x_i - r_i| \tag{A7.3}$$

- Standard deviation (SD):

$$SD = \sqrt{\frac{\sum_i^n |(x_i - r_i) - MD|^2}{n-1}} \tag{A7.4}$$

- Root-mean-square deviation (RSMD):

$$RMSD = \sqrt{\frac{\sum_i^n |x_i - r_i|^2}{n}} \tag{A7.5}$$

- Residual sum of squares (RSS):

$$RSS = \sum_i (y_i - f(x_i))^2 \tag{A7.6}$$

- Bayesian information criterion (BIC) with k fit parameters:

$$BIC = n\ln(RSS/n) + k\ln(n) \tag{A7.7}$$

- Match score (r_{match}):

$$r_{match} = \frac{\left(\sum_i^n r_i x_i\right)^2}{\left(\sum_i^n r_i^2\right)\left(\sum_i^n x_i^2\right)} \tag{A7.8}$$

- Euclidian norm (r_{euclid}):

$$r_{euclid} = \left(1.0 + \frac{\sum_i^n (r_i - x_i)^2}{\sum_i^n (x_i)^2}\right)^{-1} \tag{A7.9}$$

- Pearson correlation coefficient ($r_{pearson}$ or ρ):

$$r_{pearson} = \frac{\sum_i^n (r_i - \bar{r})(x_i - \bar{x})}{\sqrt{\sum_i^n (r_i - \bar{r})^2}\sqrt{\sum_i^n (x_i - \bar{x})^2}} \tag{A7.10}$$

- Spearman correlation coefficient ($r_{spearman}$):

$$r_{spearman} = 1.0 - \frac{6 \sum_i^n (rg(r_i) - rg(x_i))^2}{k\,(n^2 - 1)} \tag{A7.11}$$

A8. List of Abbreviations

AES	Anisotropic electrostatics
AIMD	*ab initio* molecular dynamics
ALPB	Analytical linearized Poisson–Boltzmann implicit solvation model
AMAX	Absolute maximum deviation
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
CCDC	Cambridge Crystallographic Data Centre
CCE	Clustered conformer ensemble
CE	Conformer ensemble
CFER	Cubic free energy relationship
CI	Configuration interaction
CN	Coordination number
COSMO	Conductor-like screening model
COSMO-RS	Conductor-like screening model for real solvents
CPCM	Conductor-like polarizable continuum solvation 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
DOF	Degrees of freedom
ECP	Effective core potential
EDA	Energy decomposition analysis
EEQ	Electronegativity equilibrium
EHT	Extended Hückel theory
EN	Electronegativity
FER	Free energy relationship
FF	Force field
FT-IR	Fourier-transform infrared
FWHM	Full-width at half maximum
GC	Genetic crossing
GBSA	Generalized Born, solvent accessible surface area
GFN	Geometries, frequencies, non-covalent interactions (referring to GFNn–xTB)

GFS	Global frequency scaling (factor)
gCP	Geometrical counterpoise
GGA	Generalized gradient approximation
GTO	Gaussian type orbital
HB	Hydrogen bond
HF	Hartree–Fock
HO	Harmonic oscillator
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
LFER	Linear free energy relationship
LMO	Localized molecular orbital
LP	Lone-pair
LDA, LSDA	Local (spin) density approximation
LUMO	Lowest unoccupied molecular orbital
MAD	Mean absolute deviation
MD	Molecular dynamics
MIE	Mutual information expansion
MIST	Maximum information spanning tree
MNDO	Modified neglect of diatomic overlap
MO	Molecular orbital
MP	Møller-Plesset
msRRHO	Modified and scaled RRHO
MSFS	Molecule specific frequency scaling (factor)
MTD	Metadynamics
NDDO	Neglect of diatomic differential overlap
NMR	Nuclear magnetic resonance (spectroscopy)
PCA	Principle component analysis
PM	Parametric method (referring to PMx methods)
PES	Potential energy surface
QFER	Quadratic free energy relationship
QM	Quantum mechanical
RI	Resolution-of-identity
RMSD	Root-mean-square deviation
RRHO	Rigid-rotor harmonic-oscillator approximation
RPA	Random phase approximation
SCC	Self-consistent charge
SCF	Self-consistent field
SD	Standard deviation
SIE	Self-interaction error
SPE	Single point energy
SQM	Semiempirical quantum mechanical
SRB	Short-range bond
STO	Slater type orbital
TM	Transition metal
TS	Transition state
UEG	Uniform electron gas
UV	Ultraviolet
VIS	Visible
WBO	Wiberg bond order
WFT	Wave function theory
xTB	Extended tight-binding
ZPVE	Zero-point vibrational energy

Acknowledgment

Finally, I would like to express my gratitude to the people that made this thesis possible. In particular, I wish to thank my supervisor Prof. Dr. Stefan Grimme for the great liberty and opportunity to develop own ideas and work on interesting projects. His great scientific advise and financial support, which allowed me to present my work at many occasions, enabled me to pursue a scientific future.
I am grateful to Prof. Dr. Thomas Bredow for being my second referee and his scientific advise as my mentor in the Bonn International Graduate School of Chemistry (BIGS Chemistry) program.
I also thank Sebastian Spicher, Fabian Bohle, Prof. Dr. Christoph Bannwarth and Dr. Andreas Hansen for proof-reading parts of this thesis.
For the welcoming and scientifically productive atmosphere I thank all current and former colleagues Dr. Christoph Bauer, Dr. Gerit Brandenburg, Markus Bursch, Dr. Eike Caldeweyher, Sebastian Dohm, Sebastian Ehlert, Thomas Gasevic, Johannes Gorges, Dr. Andreas Hansen, Christian Hölzer, Julia Kohn, Jeroen Koopman, Dr. Jan Mewes, Marcel Müller, Hagen Neugebauer, Jana Pisarek, Thomas Rose, Sarah Schmitz, Dr. Jakob Seibert, Marcel Stahn, Julius Stückrath, and Dr. Marc de Wergifosse.
Furthermore, I wish to thank all of my collaborators outside the Grimme group that I had the pleasure to work with. In the context of this thesis, I would like to express my gratitude to Prof. Dr. David Grant and his group, and Dr. Rainer Wilcken, Dr. Anikó Udvarhelyi and Dr. Stephane Rodde for the fruitful collaboration.
A special thanks goes to Prof. Dr. Christoph Bannwarth, Fabian Bohle, Dr. Eike Caldeweyher, and Sebastian Spicher, who I value as good friends and colleagues and with whom I had many fruitful scientific and non-scientific discussions about everything and anything. Last but not least, thanks go to my close friends Hptm. Martin Feld and Jan Lüttgenau, my sister Lisa and my parents Michael and Petra for their unconditional support over the entire period of my academic studies.