Computational Methods for Generating and Evaluating Three-Dimensional Molecular Structures

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

vorgelegt von
Christoph Plett
aus
Altenkirchen

Bonn 2025

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

Gutachter/Betreuer:	Prof. Dr. Stefan Grimme
Gutachter:	Prof. Dr. Thomas Bredow
Tag der Promotion:	16.01.2026
Erscheinungsjahr:	2026

Statement of Authorship

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 The ideas and work of others, whether published or unpublished, have been fully acknowledged and referenced in my thesis.

Publications

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

1. C. Plett and S. Grimme, *Automated and Efficient Generation of General Molecular Aggregate Structures*, Angew. Chem. Int. Ed. **62** (2023) e202214477, DOI: 10.1002/anie.202214477.

2. S. Spicher, C. Plett, P. Pracht, A. Hansen, and S. Grimme, *Automated Molecular Cluster Growing for Explicit Solvation by Efficient Force Field and Tight Binding Methods*, J. Chem. Theory Comput. **18** (2022) 3174, DOI: 10.1021/acs.jctc.2c00239.

3. C. Plett, S. Grimme, and A. Hansen, *Conformational energies of biomolecules in solution: Extending the MPCONF196 benchmark with explicit water molecules*, J. Comput. Chem. **45** (2024) 419, DOI: 10.1002/jcc.27248.

4. C. Plett, A. Katbashev, S. Ehlert, S. Grimme, and M. Bursch, *ONIOM meets xtb: efficient, accurate, and robust multi-layer simulations across the periodic table*, Phys. Chem. Chem. Phys. **25** (2023) 17860, DOI: 10.1039/d3cp02178e.

5. C. Plett, S. Grimme, and A. Hansen, *Toward Reliable Conformational Energies of Amino Acids and Dipeptides—The DipCONFS Benchmark and DipCONL Datasets*, J. Chem. Theory Comput. **20** (2024) 8329, DOI: 10.1021/acs.jctc.4c00801.

For the following articles significant contributions have been made.

6. C. Plett, M. Stahn, M. Bursch, J. M. Mewes, and S. Grimme, *Improving Quantum Chemical Solvation Models by Dynamic Radii Adjustment for Continuum Solvation (DRACO)*, J. Phys. Chem. Lett. **15** (2024) 2462, DOI: 10.1021/acs.jpclett.3c03551.

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

8. A. Katbashev, M. Stahn, T. Rose, V. Alizadeh, M. Friede, C. Plett, P. Steinbach, and S. Ehlert, *Overview on Building Blocks and Applications of Efficient and Robust Extended Tight Binding*, J. Phys. Chem. A **129** (2025), DOI: 10.1021/acs.jpca.4c08263.

9. A. Mandal, C. Maurer, C. Plett, K. R. K. Chandramohan, R. Fleischer, G. Schnakenburg, S. Grimme, and A. Bunescu, *Selective C-H Borylation of Polyaromatic Compounds Enabled by Metal-Arene π-Complexation*, J. Am. Chem. Soc. **147** (2025) 15281, DOI: 10.1021/JACS.5C00774.

10. L. Nelles-Ziegler, C. Plett, and S. Grimme, *Quantum Chemistry Based Simulation of Enantioseparation on Cyclodextrin- and Polysaccharide-Based Chiral Stationary Phases*, Chem. Eur. J. (2025), DOI: 10.1002/chem.202501398.

The following scientific presentations were given.

1. Poster on *Improving Quantum Chemical Solvation Models by Dynamic Radii Adjustment for Continuum Solvation (DRACO)*, 60th Symposium on Theoretical Chemistry (2024), Braunschweig.

2. Talk on *Automated and Efficient Generation of General Molecular Aggregate Structures*, GDCh-Wissenschaftsforum Chemie (2023), Leipzig.

3. Talk on *Modeling Explicit Solvation with the Quantum Cluster Growth Algorithm*, Ruhr-Universität Bochum (2022), Bochum.

4. Poster on *Efficient and Automated Interaction Site Screening for Modeling of Dimer and Aggregate Geometries*, 58th Symposium on Theoretical Chemistry (2022), Heidelberg.

5. Poster on *Automated Cluster Growing for Explicit Solvation by Efficient Force-Field and Tight-Binding Methods*, ACS Fall (2022), Chicago.

Abstract

Computational chemistry has become an essential tool for understanding molecular structures, dynamics, and reactivities, facilitating the discovery of new compounds and the investigation of complex molecular phenomena. A fundamental prerequisite for reliable simulations is the availability of accurate three-dimensional molecular structures. This includes knowledge about the relevant conformers, distinct energetically favorable spatial arrangements of atoms that interconvert by rotation around single bonds. While the lowest-energy conformer is generally the most representative structure of a system and thus of primary focus, higher-energy conformers can also be significantly populated at finite temperatures, thereby influencing a system's physical and chemical properties. For finding these structures, computational workflows have become increasingly important. They typically rely on tools that employ computationally efficient methods to generate an initial set of possibly relevant conformers and subsequent refinement with more accurate, but computationally demanding methods. As molecular systems increase in size, the number of possible conformers grows exponentially, eventually exceeding the practical limits of existing methods used for conformer generation. Besides the size, additional factors like interacting molecules complicate the conformer generation. For these cases, not only intramolecular flexibility, but also different intermolecular orientations have to be considered and the additional interactions complicate the computation of accurate energies. This thesis addresses the challenges of conformer generation and evaluation for large and multi-molecule systems by presenting automated tools for generating structures consisting of multiple molecules, assessing methods for the energy evaluation of conformers, and outlining efficient schemes for treating large molecular systems with high accuracy.

The first tool presented is the automated interaction site screening (aISS) algorithm. It allows the efficient docking of multiple molecules by searching for energetically favorable intermolecular orientations. The aISS algorithm is applicable to systems with elements up to an atomic number of 86, thereby significantly expanding the capabilities of existing docking tools mostly focusing on bioorganic molecules. For a set of chemically diverse monomers, the aISS tool identified intermolecular interaction sites of comparable energies as a common conformer generation tool by being computationally much less demanding. This allows the routine treatment of intermolecular interactions even for chemically diverse systems up to thousands of atoms. In addition, the aISS comes with features like site-specific docking, especially useful for mechanistic studies.

The second tool, named quantum cluster growth (QCG), is a hybrid cluster-continuum approach that expands the possibilities of docking tools like the aISS algorithm toward modeling solvation. Through sequential docking of solvent molecules to a solute while applying individually adjusted constraints, physically meaningful solute–solvent clusters are built. Additionally, QCG can automatically generate an ensemble of low-lying conformers from a constructed cluster and compute solvation free energies.

It is shown that QCG yields structures well-suited for microsolvation studies, vibrational spectra analysis, and dynamic simulations in solution.

To evaluate the performance of various computational methods in ranking the conformers generated, e.g., by QCG, the solvMPCONF196 benchmark set is presented. It provides highly accurate conformational energies for systems of biologically relevant solutes solvated by up to nine water molecules. Tested semiempirical quantum mechanical (SQM) methods and force fields (FFs) offer the efficiency required by structure generation tools like QCG, but lack accuracy compared to the tested density functional theory (DFT) or wave function theory (WFT) methods. However, DFT and WFT methods were computationally much more demanding, making them generally unfeasible for the generation of conformer ensembles but well suited for their refinement (i.e. energetic re-ranking). In this regard, the solvMPCONF196 benchmark set provides valuable insights for a reasonable method selection, balancing cost and accuracy.

For large systems like clusters with many solvent molecules where even efficient DFT methods are hardly applicable, this thesis presents an alternative approach for achieving highly accurate results: the implementation of an ONIOM (our own N-layered integrated molecular orbital and molecular mechanics) scheme in the *xtb* program suite. This approach enables a straightforward application of said embedding scheme using the efficient FF and SQM methods available in *xtb* combined with common DFT or WFT methods. The resulting multi-layer scheme greatly accelerates structure and energy refinements compared to purely DFT-based approaches while maintaining similar accuracy. Its application is not only demonstrated for solute–solvent clusters, but also for electronically challenging systems like metal-organic frameworks (MOFs) as well as systems too large for applying purely DFT or WFT methods.

Lastly, the DipCONFS benchmark and DipCONFL dataset combination is presented that supports the development of future methods for conformer evaluation. Both sets cover conformers of 17 amino acids and their 289 possible dipeptides. The DipCONFS benchmark set contains about 1,000 highly accurate data points that have been utilized to identify suitable reference methods for generating the larger DipCONFL dataset. With almost 30,000 accurate DFT data points, the DipCONFL can complement training sets for machine-learned interatomic potentials (MLIPs), thereby supporting the development of more robust models that potentially improve achievable cost–accuracy ratios.

Contents

CHAPTER 1

Introduction

Computational simulations have become an invaluable tool for modern chemical research[1,2] and nowadays support the development of transformative technologies across various fields such as drug discovery,[3–6] material science,[7,8] and catalysis.[9–11] They enable the virtual modeling of molecular structures, physicochemical properties, and atomistic processes, thereby elucidating experimental findings.[12–17] Further, as computational methods can simulate unknown compounds efficiently, they have become highly valuable for applications such as high-throughput screening.[18,19]

A fundamental prerequisite for reliable computational simulations of a molecular system is the accurate representation of its three-dimensional structure. This, in turn, requires knowledge about the atomic composition and bonding pattern that determine the possible spatial arrangements of the atoms, from which the most stable structures typically represent the real system.[20] In this context, assessing conformers is important. They are defined as minima on the potential energy surface (PES) that can be transformed via rotation around single bonds.[21,22] While all conformers of a molecule have the same molecular formula and connectivity, they can differ significantly in their chemical and physical properties. For example, only certain conformers of a drug may be biologically active, while others are inactive.[23–25] Thus, identifying and using the conformers in computational simulations that are predominant in the real system is of significant importance to avoid erroneous predictions.[26]

Identifying the relevant conformers can, in principle, be done with various experimental techniques such as nuclear magnetic resonance (NMR) spectroscopy and X-ray crystallography, but these methods are often resource-intensive and limited in scope.[27–29] Therefore, computational methods that allow efficient and flexible exploration of the conformational space have become invaluable.[30,31] This commonly involves two major aspects: generating all relevant conformers and evaluating their relative conformational energies. Thereby, finding the lowest-energy conformer is of major importance as it typically has the highest probability of being found in the real system. Further, depending on the system's temperature, other conformers close in energy can also be populated significantly and thus contribute to the observable property of the system.[32] Thus, the ensemble of conformers is usually targeted that is as complete as possible with regard to low-energy conformers.

A major challenge when generating conformers computationally is the exponential growth of the conformational space with the number of rotatable bonds.[33] This can be highlighted, for example, with alkanes: while *n*-propane (C_3H_8) possesses only a single unique staggered conformer, the number of stable conformers increases combinatorially with chain length, reaching on the order of hundreds of thousands for *n*-tridecane ($C_{13}H_{28}$).[34] Therefore, to retain computational feasibility, conformer

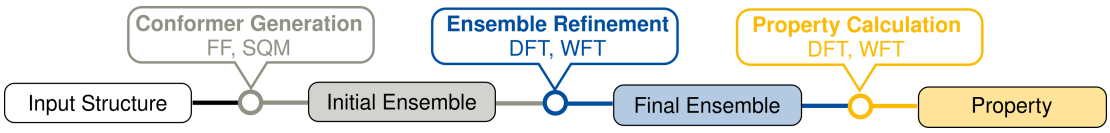

Figure 1.1: Schematic depiction of a typical computational workflow for predicting molecular properties from an initial structure via conformer generation and refinement.

generation tools often use force fields (FFs) or semiempirical quantum mechanical (SQM) methods for generating and evaluating the possible conformers. While these methods provide the required efficiency, they often lack accuracy compared to computationally more demanding approaches like density functional theory (DFT) or wave function theory (WFT) methods. To make up this deficit, multi-level workflows are typically used.[35] They rely on SQM or FF methods for generating and assessing an initial ensemble of possibly relevant conformers, refine the conformational energies and structures with WFT or DFT methods, and finally calculate the properties employing highly accurate WFT or DFT methods to the populated conformers (Figure 1.1).[36–38]

In this context, treating multiple interacting molecules imposes various challenges, but becomes important in many applications. Non-covalent interactions (NCIs) are almost ubiquitous in chemistry and play a critical role in various fields like drug development,[39,40] material design,[41] and optimizing chemical reactions.[42] Further, they influence conformational energies and thus have an impact on the molecular structure, making their consideration often essential in computational simulations. However, they complicate both the conformer generation and their energetic evaluation. In addition to the intramolecular flexibility of the isolated components, their relative orientation must also be considered. This increases the number of possible stable structures and leads to generally very rich energy surfaces with numerous energy minima, complicating the generation of the relevant three-dimensional molecular structures.[43] Thus, despite the significant advances in computational conformer exploration, sampling systems with multiple interacting molecules remains challenging, even when optimizing common conformer search tools for treating NCIs.[44,45] Besides their influence on the conformational space, NCIs also complicate the computation of accurate conformational energies due to intermolecular contributions like electrostatic interactions between permanently polarized molecules, induction between permanent and induced multipoles, London dispersion interactions due to spontaneous electron fluctuation, and charge transfer between the molecules.[46] This increases the demands on methods for computing conformational energies and requires to test them beyond isolated molecules. Further, systems of interacting molecules can reach large sizes so that sufficiently accurate methods become unfeasible due to their usually rather large computational expense.[36]

Given the importance of NCIs, it is worthwhile to address these challenges to extend the capabilities of computational methods for finding physically meaningful orientations of interacting molecules and identifying the relevant conformers. This can become a key aspect for accurate computational workflows used to, e.g., predict protein-ligand affinities, the association of reactive molecules during a reaction mechanism, the modeling of solvation by including solvent molecules, and the investigation of solids (Figure 1.2).

Therefore, this thesis presents tools for exploring the conformational space of interacting molecules, data for assessing and training various methods for conformational energy computation with respect to NCIs and large systems, and schemes to combine different methods for achieving high accuracy even for large systems.

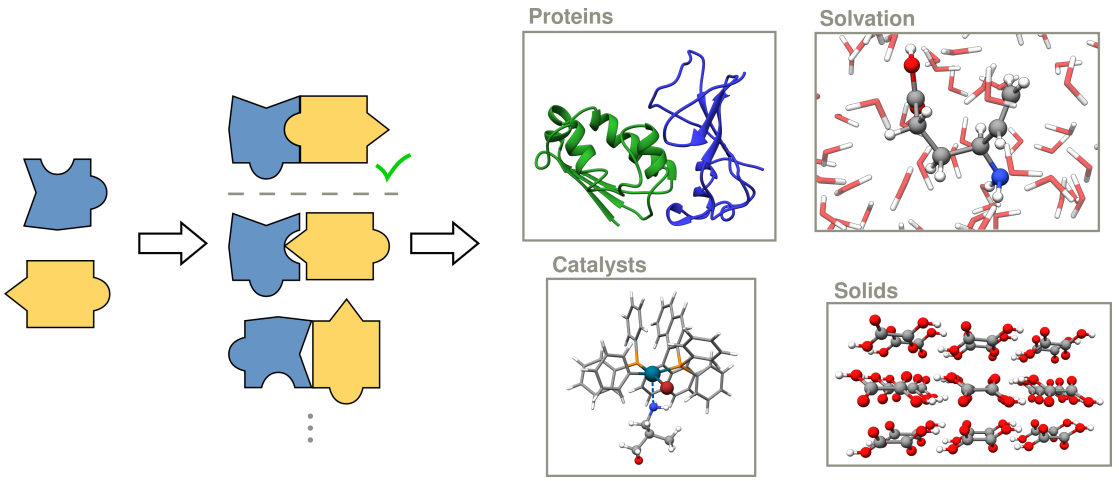

Figure 1.2: Illustration of the conformational complexity introduced by intermolecular interactions and the resulting challenge for identifying the optimal intermolecular arrangement. It is a key aspect to reach high accuracy in many computational simulations, e.g., for predicting protein-ligand affinities, the association of reactive molecules during a reaction mechanism, and the modeling of solvent molecules and solids.

Chapter 3 introduces a new efficient docking tool, called automated Interaction Site Screening (aISS), that automatically generates accurate structures of interacting molecules. It combines an intermolecular FF (xTB-IFF)[47] for structure screening with the efficient GFN1-xTB[48] or GFN2-xTB[49] SQM method, or the GFN-FF[50] FF for geometry optimizations. The aISS algorithm is robustly applicable to various molecules with elemental compositions across the periodic table up to radon. It thus significantly expands the capabilities of common docking tools that are often restricted to biomolecular systems.[51–53]

Chapter 4 extends this approach toward modeling solvation. This requires not only considering the interactions of individual molecules, but rather the solute interacting with a bulk of solvent molecules. Accounting for the resulting solvent effect is often essential for reliable simulations as it can significantly influence the conformational energies, thereby altering the structure and properties of the solute.[54–57] For example, the solvent has a significant influence on protein structures, redox potentials, and vibrational spectroscopy results, and plays a crucial role in the synthesis of compounds.[58–62] Facilitating the routine inclusion of solvent molecules in computational simulations, the Quantum Cluster Growth (QCG) hybrid cluster–continuum solvation approach[63–66] is presented that is capable of growing a solvent cluster around a solute by sequentially docking solvent molecules. Further, it can automatically generate an ensemble from this cluster and compute solvation free energies. Applied repulsive potentials ensure physically meaningful structures and the use of implicit solvent models accounts for the bulk solvent. Unlike using only implicit solvent models, the integration of solvent molecules with QCG can overcome limitations of common implicit solvent models such as their tendency to overestimate intramolecular interactions compared to intermolecular ones.[67,68] Further, the clusters of QCG provide an alternative approach to explicit solvent models that often require the simulation of much larger systems, limiting the application of accurate but expensive methods.[69]

While QCG relies, similarly to other conformer generation tools, on using efficient FF and SQM methods for generating the clusters, the resulting cluster sizes are typically treatable with more expensive but more accurate DFT or WFT methods. Therefore, it is important to understand the

limitations of the FF and SQM methods and to identify methods suitable for refining the ensembles. For this, typically benchmark sets are consulted that contain a representative set of structures for a given compound class along with highly accurate reference data.[70,71] While numerous benchmark sets exist covering the conformational energies of isolated molecules,[72–76] only a few include structures of interacting molecules.[77–80] For small clusters and aggregates, the available data are even more sparse, and especially in the context of explicit solvation, benchmark sets are hardly available, as the few existing ones mainly focus on pure water clusters.[81,82] To extend this field, the solvMPCONF196 benchmark set, presented in Chapter 5, includes conformers of biologically relevant molecules solvated with water molecules. It outlines the effects and the challenges arising from explicit solvation, and by assessing a variety of DFT, WFT, SQM, and FF methods against highly precise reference data, a reliable selection of these methods for refining solute–solvent clusters becomes possible.

However, the most accurate methods are usually computationally costly and thus not applicable to systems beyond a certain size. Aiming to extend this limit, Chapter 6 outlines an implementation of the ONIOM (our own N-layered integrated molecular orbital and molecular mechanics) scheme[83] in the *xtb* program suite.[84] This gives access to combining the GFN (geometries, frequencies, and non-covalent interactions) methods with DFT and WFT. The GFN methods are particularly well suited for an application within the ONIOM scheme due to their robust performance for a large variety of compounds with diverse elemental composition.[84,85] It is shown that the resulting multi-layer approach can provide accurate geometries and energies even for larger systems.

As an alternative to ONIOM schemes, machine-learning (ML) methods also show great potential for improving the cost-accuracy ratio for conformer evaluation. Generally, ML models have a wide range of application. For example, protein structure prediction has been revolutionized with AlphaFold[86] and DFT methods were improved and accelerated by ML.[87,88] In terms of efficiently computing conformational energies, machine-learned interatomic potentials (MLIPs) like AIMNet2[89] and the class of universal models for atoms (UMA)[90] are promising candidates as they map atomic structures to their potential energies and can offer significant speed advantages over DFT, WFT, and even SQM methods.[91] To achieve this, they incorporate no, or only a minor degree of physical descriptions, and compensate this by training a neural network on large datasets.[92] This causes their accuracy, range of applicability, and robustness to be highly dependent on the quality and diversity of the respective training data.[93] Therefore, diverse and well-balanced datasets are needed to train robust MLIPs and to evaluate them accurately.[94] To generate such data, often computational methods are applied to small, but representative systems.[95] Chapter 7 presents an extension to the available data for biologically important systems of amino acids and dipeptides. First, the DipCONFS benchmark set is presented, comprising nearly 1,000 conformers with highly accurate reference data. It is used to assess the performance of various DFT and WFT methods regarding their suitability as a reference method for large-scale datasets. Building upon these findings, the DipCONFS benchmark was extended to over 29,000 conformers with accurate electronic DFT properties forming the DipCONFL dataset. It offers valuable insights into data generation and extends the existing data, e.g., for training and validating ML models aimed at biomolecular conformational sampling.

Following this introduction, an overview is given on how energies of molecular structures can be computed and leveraged for structure generation and evaluation (Chapter 2). This is followed by a summary of the five previously mentioned projects in the Chapters 3 to 7 complemented by the respective publications in the appendices B to F. Finally, Chapter 8 summarizes the key findings of this work, places them in a broader scientific context, and outlines potential directions for future developments.

CHAPTER 2

Theoretical Background

2.1 Gibbs Free Energy

Generating and ranking conformers requires a way to map the three-dimensional structure of a molecule to its energy. In this regard, the Gibbs free energy is of fundamental importance as it relates theoretical results to observable phenomena. For typical reaction conditions of constant pressure and temperature, a system strives to adapt the state of lowest Gibbs free energy.[96] Thus, it can be used, e.g., to predict the most-populated conformers, the solubility of a molecule, and whether molecules associate or not.[35,97,98]

The Gibbs free energy G_{tot} includes the enthalpy H_{tot} and entropy S_{tot}:[99]

$$G_{\text{tot}}(T) = H_{\text{tot}}(T) - TS_{\text{tot}}(T). \tag{2.1}$$

Typically, $H_{\text{tot}}(T)$ is not computed directly, but is split into a temperature-dependent and independent part. In the gas phase, the temperature-independent part consists of the molecular energy E and the zero-point vibrational energy ($ZPVE$), which includes the residual energy at 0 K attributed to vibrational motions. The temperature-dependent part of the entropy ($H(0\ K \rightarrow T)$) stems from the population of different translational, vibrational, and rotational energy states of a system with increasing temperature. With this, the Gibbs free energy in the gas phase can be written as

$$G^{\text{gas}}_{\text{tot}}(T) = E + ZPVE + H(0\ K \rightarrow T) - TS_{\text{tot}}(T) = E + G^{\text{gas}}_{\text{corr.}}. \tag{2.2}$$

$G^{\text{gas}}_{\text{corr.}}$ includes the ZPVE and the enthalpy and entropy contributions arising from the temperature-dependent population of translational, vibrational, and rotational energy states (thermostatistical contributions). E is often obtained directly using QM methods, illustrated in Section 2.2, or using FFs as outlined in Section 2.3. The thermostatistical corrections $G^{\text{gas}}_{\text{corr.}}$ require additional approximations explained in Section 2.4. Additionally, when considering molecular processes in solution, solvent-dependent contributions also have to be accounted for. In terms of energy, they are usually included with the solvation free energy δG_{solv} that is used to calculate the Gibbs free energy in solution ($G^{\text{solu}}_{\text{tot}}$):

$$G^{\text{solu}}_{\text{tot}}(T) = E + G^{\text{gas}}_{\text{corr.}}(T) + \delta G_{\text{solv}}. \tag{2.3}$$

δG_{solv} is usually computed with additional methods outlined in Section 2.5. Finally, how methods

for computing the Gibbs free energy can then be utilized to generate conformers is illustrated in Section 2.6.

Equations in this chapter are in atomic units. Ref 99 and 100 were taken as sources for most of the content in this Chapter.

2.2 Electronic Structure Theory

2.2.1 General Concepts

Electronic structure theory methods rely on treating the electrons and nuclei of molecular systems explicitly to compute the molecular energy E. This requires an appropriate mathematical description of the particles and a way to compute the energy from this description. For describing a system composed of N_{elec} electrons and N_{nuc} nuclei, a wavefunction $\Psi(\mathbf{r}, \mathbf{R}, t)$ is typically used that depends on all the positions of the electrons ($\mathbf{r} \equiv r_1, r_2, ..., r_{N_{elec}}$) and nuclei ($\mathbf{R} \equiv R_1, R_2, ..., R_{N_{nuc}}$), as well as the time t relative to a reference point t_0. The positions of the electrons r_i and nuclei R_i denote a vector uniquely defining the position, e.g., by the three coordinates x_i, y_i, and z_i in Cartesian space. The energy E of a wavefunction can be computed with the Hamiltonian $\hat{H}$. When no external field is acting on the system, its energy and thus the Hamiltonian is not time-dependent. In this case, the time-dependent part of the wavefunction is usually separated and not considered further, leading to the time-independent, non-relativistic Schrödinger equation:

$$\hat{H}\Psi(\mathbf{r}, \mathbf{R}) = E\Psi(\mathbf{r}, \mathbf{R}). \tag{2.4}$$

While defining an appropriate wavefunction is very complex and will be explained later, the time-independent Hamiltonian can be directly deduced from classical mechanics as a sum of the kinetic $\hat{T}$ and potential $\hat{V}$ energy operators:

$$\hat{H} = \hat{T}(\mathbf{r}, \mathbf{R}) + \hat{V}(\mathbf{r}, \mathbf{R}). \tag{2.5}$$

For molecular systems consisting of N_{elec} electrons at position r_i and N_{nuc} nuclei at position R_i with the mass M_i, the kinetic energy operator is given by a sum over all nuclei $\hat{T}_{\text{n}}(\mathbf{R})$ and electrons $\hat{T}_{\text{e}}(\mathbf{r})$:

$$\hat{T}(\mathbf{r}, \mathbf{R}) = \underbrace{-\sum_{i=1}^{N_{\text{nuc}}} \frac{1}{2M_i}\nabla_i^2}_{\hat{T}_{\text{n}}(\mathbf{R})} \underbrace{-\sum_{i=1}^{N_{\text{elec}}} \frac{1}{2}\nabla_i^2}_{\hat{T}_{\text{e}}(\mathbf{r})}, \tag{2.6}$$

with

$$\nabla_i^2 = \left(\frac{\partial^2}{\partial x_i^2} + \frac{\partial^2}{\partial y_i^2} + \frac{\partial^2}{\partial z_i^2}\right). \tag{2.7}$$

The potential energy can be expressed with the Coulomb law, describing the potential energy of interacting point charges. Therefore, a sum over all pairwise interactions is used, including the nuclei-nuclei repulsion ($\hat{V}_{\text{nn}}(\mathbf{R})$), and electron-electron repulsion ($\hat{V}_{\text{ee}}(\mathbf{r})$) as well as the nuclei-electron

attraction ($\hat{V}_{\mathrm{ne}}(\mathbf{r},\mathbf{R})$):

$$\hat{V}(\mathbf{r},\mathbf{R}) = \underbrace{\sum_{i}^{N_{\mathrm{nuc}}-1}\sum_{j=i+1}^{N_{\mathrm{nuc}}}\frac{Z_i Z_j}{|R_i - R_j|}}_{\hat{V}_{\mathrm{nn}}(\mathbf{R})} + \underbrace{\sum_{i}^{N_{\mathrm{elec}}-1}\sum_{j=i+1}^{N_{\mathrm{elec}}}\frac{1}{|r_i - r_j|}}_{\hat{V}_{\mathrm{ee}}(\mathbf{r})} \underbrace{-\sum_{i}^{N_{\mathrm{elec}}}\sum_{j}^{N_{\mathrm{nuc}}}\frac{Z_j}{|r_i - R_j|}}_{\hat{V}_{\mathrm{ne}}(\mathbf{r},\mathbf{R})}. \tag{2.8}$$

With the Hamiltonian defined, the two remaining components required for computing the energy of the system are the wavefunction $\Psi(\mathbf{r},\mathbf{R})$ and the positions of the electrons $\mathbf{r}$ and nuclei $\mathbf{R}$. For determining these, often the Born–Oppenheimer approximation, separating the dependence of the electron positions from the movements of the nuclei, is applied.[101] It is based on the fact that a nuclei is much heavier than an electron (a single proton is about 1838 times heavier than an electron) and thus an electron is much faster than the nuclei. Therefore, when evaluating the electronic energy contributions, the nuclei positions are approximately fixed. Based on this approximation and the electronic Hamiltonian

$$\hat{H}_{\mathrm{e}} = \hat{T}_{\mathrm{e}}(\mathbf{r}) + \hat{V}_{\mathrm{ee}}(\mathbf{r}) + \hat{V}_{\mathrm{ne}}(\mathbf{r},\mathbf{R}), \tag{2.9}$$

the wavefunction is split into a product of an electron ($\Psi(\mathbf{r})$) and nuclei-dependent ($\Psi(\mathbf{R})$) part. While the wavefunction of the nuclei is typically not determined explicitly, the electronic wavefunction can be solved approximately with the electronic Schrödinger equation:

$$\hat{H}_{\mathrm{e}}\Psi(\mathbf{r}) = E_{\mathrm{e}}\Psi(\mathbf{r}). \tag{2.10}$$

Independent of the nuclear motion, this equation provides a connection between the electron positions and their energy E_{e} within the Born–Oppenheimer approximation. It can be used to determine the electron positions by considering that electrons adapt the state of lowest energy possible. In principle, this is done by altering the electron positions of an initial guess until the lowest electronic energy results. This state of lowest energy is the electronic ground state $\Psi_0(\mathbf{r})$. $\hat{V}_{\mathrm{nn}}$ is calculated separately via Coulomb law (Equation 2.8) and added to E_{e} to yield the molecular energy E.

With the energy expression and a way to determine the electron positions at hand, the final but very crucial question is left of how the wavefunction should be defined. Based on different ways to answer this question, various methods like Hartree-Fock or coupled cluster have evolved, which are outlined in the following chapter. As testing every possible wavefunction is practically impossible for almost any molecular system, they are usually limited to varying only a few parameters c_Ψ. This restriction prevents finding the exact wavefunction in almost all cases, causing the computed energies of the ground state ($\tilde{E}_0$) to be higher than or equal to the exact energy E_0 according to the variational principle:

$$\tilde{E}_0 \geq E_0. \tag{2.11}$$

2.2.2 Hartree-Fock Theory

A common form of the multi-electron wavefunction $\Psi(\mathbf{r})$ is a linear combination of single-electron wavefunctions $\psi(i)$. In the Hartree-Fock (HF) method, this linear combination has the form of a Slater

determinant Ψ^{SD}:

$$\Psi(\mathbf{r}) \approx \Psi^{\mathrm{SD}} = \frac{1}{\sqrt{N_{\mathrm{elec}}!}} \begin{vmatrix} \psi_1(1) & \psi_2(1) & \cdots & \psi_{N_{\mathrm{elec}}}(1) \\ \psi_1(2) & \psi_2(2) & \cdots & \psi_{N_{\mathrm{elec}}}(2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(N_{\mathrm{elec}}) & \psi_2(N_{\mathrm{elec}}) & \cdots & \psi_{N_{\mathrm{elec}}}(N_{\mathrm{elec}}) \end{vmatrix}. \tag{2.12}$$

Here, $i = 1, 2, ..., N_{\mathrm{elec}}$ summarizes the spatial and spin coordinates of an electron i, and $\psi_k(i)$ denotes the single-electron wavefunctions, that are called molecular orbitals (MOs) when describing molecules. The Slater determinant fulfills several important requirements for a physically appropriate description: The electrons are indistinguishable, changing two electrons leads to a change in sign, resembling their antisymmetry, and also, the Pauli principle is fulfilled as using two identical single-electron orbitals in the Slater determinant would result in a vanishing wavefunction.

The single-electron wavefunctions consist of a spatial (χ_k) and a spin function (σ_k):

$$\psi_k(i) = \chi_k(i) \cdot \sigma_k. \tag{2.13}$$

The electronic spin functions are α or β that are orthonormal:

$$\begin{aligned} \langle\alpha|\alpha\rangle = \langle\beta|\beta\rangle &= 1 \\ \langle\alpha|\beta\rangle &= 0 \end{aligned} \tag{2.14}$$

The spatial orbitals are represented by a linear combination of atomic orbitals (LCAO), with the atomic orbitals (AOs) represented as basis functions ϕ_μ:

$$\chi_k(i) = \sum_{\mu=1}^{N_{\mathrm{BF}}} c_{\mu k} \phi_\mu, \tag{2.15}$$

with N_{BF} denoting the total number of basis functions. The basis functions contribute a certain amount to each MO χ_k defined by the coefficient $c_{\mu k}$. While the number and shape of the basis functions used are defined by the basis set introduced in section 2.2.3, the coefficients are optimized during the Hartree-Fock procedure. Therefore, it is important to know how the energy of a Slater determinant can be computed.

When using such an approximated wavefunction, it will likely not be an eigenstate of the Hamiltonian. Thus, the expectation value of the energy is computed by integration over all coordinates:

$$E_{\mathrm{e}} = \int \Psi^{SD^*} \hat{H}_{\mathrm{e}} \Psi^{SD} dr = \langle \Psi^{SD} | \hat{H}_{\mathrm{e}} | \Psi^{SD} \rangle. \tag{2.16}$$

The electronic Hamiltonian can be defined as a sum of the one-electron operators $\hat{T}_{\mathrm{e}}$ and $\hat{V}_{\mathrm{ne}}$, $\hat{h}$, and the two electron operator $\hat{g}$:

$$\hat{H}_{\mathrm{e}} = \sum_{i=1}^{N_{\mathrm{elec}}} \hat{T}_{\mathrm{e}}(r_i) + \sum_{i=1}^{N_{\mathrm{elec}}} \sum_{j=1}^{N_{\mathrm{nuc}}} \hat{V}_{\mathrm{ne}}(r_i, R_j) + \sum_{i=1}^{N_{\mathrm{elec}}} \sum_{j=i+1}^{N_{\mathrm{elec}}} \hat{V}_{\mathrm{ee}}(r_i, r_j) = \sum_{i=1}^{N_{\mathrm{elec}}} \hat{h}_i + \sum_{i=1}^{N_{\mathrm{elec}}} \sum_{j=i+1}^{N_{\mathrm{elec}}} \hat{g}_{ij}, \tag{2.17}$$

with

$$\hat{h}_i = \hat{T}_\mathrm{e}(r_i) + \sum_{j=1}^{N_\mathrm{nuc}} \hat{V}_\mathrm{ne}(r_i, R_j) = -\frac{1}{2}\nabla_i^2 - \sum_{j=1}^{N_\mathrm{nuc}} \frac{Z_j}{|r_i - R_j|}, \tag{2.18}$$

and

$$\hat{g}_{ij} = \hat{V}_\mathrm{ee}(r_i, r_j) = \frac{1}{|r_i - r_j|}. \tag{2.19}$$

With this, it can be shown that most combinations of the single-electron wavefunctions ψ_k become zero due to the orthonormality of the spin orbitals. Only the following sums over the MOs k and l are non-zero in Equation 2.16:

$$E_\mathrm{e} = \sum_{k=1}^{N_\mathrm{elec}} \langle \psi_k(i)|\hat{h}_i|\psi_k(i)\rangle + \sum_{k=1}^{N_\mathrm{elec}} \sum_{l=k+1}^{N_\mathrm{elec}} \langle \psi_k(i)\psi_l(j)|\hat{g}_{ij}|\psi_k(i)\psi_l(j)\rangle - \sum_{k=1}^{N_\mathrm{elec}} \sum_{l=k+1}^{N_\mathrm{elec}} \langle \psi_k(i)\psi_l(j)|\hat{g}_{ij}|\psi_l(i)\psi_k(j)\rangle = \sum_{k=1}^{N_\mathrm{elec}} I_k + \sum_{k=1}^{N_\mathrm{elec}} \sum_{l=k+1}^{N_\mathrm{elec}} \left(J_{kl} - K_{kl}\right). \tag{2.20}$$

I_k corresponds to the mean kinetic and potential energy of an electron i occupying the single-electron MO ψ_k that is influenced by the field of the nuclei. The Coulomb integral J_{kl} is analogous to the Coulomb repulsion between two electrons i and j occupying the MOs ψ_k and ψ_l. The exchange integral K_{kl} does not have a classical analog and is only non-zero when the two MOs ψ_k and ψ_l have the same spin orbital.

For closed shell cases, often the spatial orbitals are assumed to be doubly occupied so that each spatial orbital is occupied by an electron with α and β spin. This leads to restricted Hartree Fock (RHF), where, different from unrestricted Hartree Fock (UHF), only half the orbitals have to be treated explicitly due to the spin orthonormality. Thus, the energy can be computed as:

$$E_\mathrm{e}^\mathrm{RHF} = 2 \sum_{k=1}^{N_\mathrm{elec}/2} I_k + \sum_{k=1}^{N_\mathrm{elec}/2} \sum_{l=1}^{N_\mathrm{elec}/2} \left(2J_{kl} - K_{kl}\right). \tag{2.21}$$

To finally compute the energy ϵ_k of an MO ψ_k, the single-electron Coulomb and exchange operators $\hat{J}_l$ and $\hat{K}_l$ are defined to act on a wavefunction ψ_k according to:

$$\hat{J}_l(i)\psi_k(i)\rangle = \langle \psi_l(j)|\hat{g}_{ij}|\psi_l(j)\rangle|\psi_k(i)\rangle, \tag{2.22}$$

and

$$\hat{K}_l(i)\psi_k(i)\rangle = \langle \psi_l(j)|\hat{g}_{ij}|\psi_k(j)\rangle|\psi_l(i)\rangle. \tag{2.23}$$

With this, the Fock operator can be formed

$$\hat{f}(i) = \hat{h}(i) + \sum_{j}^{N_\mathrm{elec}} \left(\hat{J}_j(i) - \hat{K}_j(i)\right), \tag{2.24}$$

that can be used to compute the MO energies ϵ_k

$$f(i)\psi_k(i) = \epsilon_k\psi_k(i). \tag{2.25}$$

Accordingly, the MOs are represented as a set of eigenvectors of the Fock operator with the respective MO energies as eigenvalues. As the MOs are already required for constructing the Fock operator, both must be optimized together.

Inserting the LCAO approach into Equation 2.25 leads to the Roothaan–Hall equations:

$$f(i)\sum_{\mu=1}^{N_{\mathrm{BF}}} c_{\mu k}\phi_\mu = \epsilon_k \sum_{\mu=1}^{N_{\mathrm{BF}}} c_{\mu k}\phi_\mu, \tag{2.26}$$

which can be written with the individual elements arranged in matrices:

$$\mathbf{FC} = \mathbf{SC}\boldsymbol{\epsilon}. \tag{2.27}$$

$\mathbf{F}$ contains all Fock operators $f(i)$ and $\boldsymbol{\epsilon}$ the orbital energies of the occupied molecular orbital. $\mathbf{S}$ denotes the matrix containing the overlap elements of the basis functions, and $\mathbf{C}$ the coefficients $c_{\mu k}$. This equation is solved iteratively to determine the optimal coefficients and thus the optimal MOs. An initial guess for the coefficients is used to form $\mathbf{F}$. Diagonalizing this matrix yields new coefficients $\mathbf{C}$ with typically lower MO energies. These are used to form a new matrix $\mathbf{F}$, which is again diagonalized. This is repeated until the total energy of the system converges. With this procedure, called the Self-Consistent Field (SCF) method, the Slater determinant with the lowest energy possible for the introduced constraints of the electronic ground state is found. As this procedure can introduce significant computational costs, usually approximations like the resolution-of-the-identity (RI) approximation or the numerical chain-of-sphere integration for the exchange integrals (COSX) are used to accelerate the HF method.[102,103]

2.2.3 Basis Sets

The basis functions chosen to form the MOs via the LCAO approach (Equation 2.15) are crucial for the accuracy of approaches like Hartree Fock. In principle, the number and shape of the basis functions ϕ_μ could be chosen to resemble those of the isolated atoms, e.g., one 1s orbital for the hydrogen atom. Therefore, a single Slater function ($Y_{l_Q}^{m_Q}(r)r^{n_Q-1}e^{-\zeta|r|}$) could be used, where $Y_{l_Q}^{m_Q}(r)$ is determined by the type and thus the shape of the orbital, n_Q is the quantum number of the respective orbital and ζ an exponent that determines the spatial extent of the orbital. However, this is usually a poor approximation for molecular systems, as, for example, the hydrogen atoms in water (H_2O) have a lower electron density compared to the isolated atom, which would require smaller 1s orbitals to form the MOs. Further, the spherical 1s orbitals deform upon bonding, which also must be considered.

In principle, this could be accounted for by using an infinite number of differently shaped basis functions per atom, for which the LCAO approach becomes exact. However, the computational costs of Hartree Fock scale formally with $O(N_{\mathrm{BF}}^4)$ due to the two-electron integrals, so that a set of basis functions should be chosen that is as small as possible by providing enough flexibility to accurately cover a variety of molecular systems. This led to the development of basis sets that contain a specific set of basis functions per element with defined shape. The number of basis functions used per element

to form the MOs is commonly denoted by ζ. For example, a single-ζ basis set would have a single basis function per hydrogen atom, while a double-ζ (DZ) basis set would provide two basis functions with different sizes. This can be continued with triple-ζ (TZ), quadruple-ζ (QZ), and so on. Additionally, basis functions with higher angular momentum can be included to describe the deformation of the AOs, e.g., upon bonding and polarization. Finally, also diffuse functions with a large spatial extent can be included to describe delocalized electron densities, e.g., for anions.

Two classes of basis sets are mainly used in this work. Dunning's basis sets, denoted by cc-pVx_D ($x_D = DZ, TZ, ...$) or aug-cc-pVx_D if diffuse functions are included, were optimized for correlated methods.[104] The Ahlrichs' basis sets, denoted by def2-x_AVP ($x_A = S, TZ, QZ, ...$), are reliable choices for most DFT applications.[105] Respective sets with polarization functions are denoted as def2-x_AVPP and additional diffuse functions by def2-x_AVPPD.[106]

Basis sets of both types further replace core electrons in heavier atoms with effective core potentials. They mimic the core electrons without the need to explicitly treat them and often account partially for effects such as relativity. Further, they rely on using Gaussian-shaped basis functions ($Y_{l_Q}^{m_Q}(r)e^{-\zeta r^2}$), even though appropriate basis functions are usually shaped similarly to a Slater function. However, they have technical advantages like the Gaussian product theorem, and by contracting multiple primitive Gaussian-type orbitals (PGTOs) to form a contracted Gaussian-type orbital (CGTO), appropriate basis functions can be obtained:

$$\phi_\mu = \sum_{k_{PGTO}=1}^{N_{PGTO}} \iota_{k_{PGTO}} \phi_{\mu k_{PGTO}}. \tag{2.28}$$

The number of PGTOs N_{PGTO}, the shape of the PGTOs $\phi_{\mu k_{PGTO}}$, and the coefficient $\iota_{k_{PGTO}}$ which define the contribution of each PGTO to the final basis function, are specified with the basis sets.

The choice of basis set sizes strongly depends on the targeted accuracy. While larger sets like quadruple-ζ almost reach the basis set limit, they lead to high computational costs. On the other hand, smaller-sized sets lack accuracy due to the basis set incompleteness error (BSIE) and the basis set superposition error (BSSE). While the BSIE arises due to the lack of basis functions, the BSSE is an artificial energy lowering as the lack of basis functions is compensated by the unphysical use of basis functions, e.g., of another molecule. Therefore, the BSSE is most pronounced for interacting fragments, but can also occur for single molecules.[107] As the BSSE can lead to significant errors when smaller basis sets are used,[108] schemes like the Boys–Bernardi counterpoise corrections or the geometric Counter–Poise (gCP) correction were developed aiming to reduce the BSSE.[109,110]

2.2.4 Electron Correlation

In Hartree Fock, electrons move in the mean field of all the other electrons, neglecting the correlated movements of individual electrons. As a result, the Hartree–Fock ground-state energy E_0^{HF} does not account for the correlation energy E_{corr}:

$$E_{\mathrm{corr}} = E_{\mathrm{exact}} - E_0^{\mathrm{HF}}. \tag{2.29}$$

To improve on this, excited determinants can be used in addition to Ψ_0^{SD}. These additional determinants are similar to Ψ_0^{SD}, but replace occupied MOs by virtual orbitals. Covering all possible excited Slater

determinants leads to the full configuration interaction (FCI) method

$$\Psi^{\mathrm{FCI}} = a_0^{\mathrm{CI}}\Psi_0^{\mathrm{SD}} + \sum_S a_S^{\mathrm{CI}}\Psi_S^{\mathrm{SD}} + \sum_D a_D^{\mathrm{CI}}\Psi_D^{\mathrm{SD}} + \ \ldots, \tag{2.30}$$

where $\Psi_{0/S/D}^{\mathrm{SD}}$ denotes the reference determinant (0), the singly excited determinants (S), and the doubly excited determinants (D). $a_{0/S/D}^{\mathrm{CI}}$ are the coefficients determining how much the respective Slater determinant contributes to the total configuration-interaction Wavefunction Ψ^{FCI}. These coefficients are determined variationally. The FCI method is computationally extremely expensive in terms of basis functions ($O(N_{\mathrm{BF}}!)$ and thus usually not applicable to molecules containing more than a few atoms. To reduce the expense, it can be truncated after a few excitations, but this leads to methods not being size consistent, potentially causing larger errors.[111]

An alternative, size-consistent approach is coupled cluster. Here, the wavefunction Ψ^{CC} is defined with an excitation operator $\hat{T}$

$$\Psi^{\mathrm{CC}} = e^{\hat{T}}\Psi_0^{\mathrm{SD}}, \tag{2.31}$$

that contains all excitation operators

$$\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 + \ \ldots\ + \hat{T}_{N_{\mathrm{elec}}}. \tag{2.32}$$

These excitation operators generate all the respective excitations of the ground-state Slater determinant. To this, a Taylor expansion is applied

$$e^{\hat{T}} = 1 + \hat{T} + \frac{1}{2}\hat{T}^2 + \frac{1}{6}\hat{T}^3 + \ \ldots, \tag{2.33}$$

that is truncated after a certain order. This has the advantage that truncation after, e.g., $\hat{T}_2$ yields the CCSD method, which includes higher order excitation terms like $\hat{T}_2 \cdot \hat{T}_1$ and $\hat{T}_1^3$ called disconnected triples. For practical applications, truncating after $\hat{T}_2$ is usually insufficiently accurate and at least $\hat{T}^3$ should be considered. However, methods like CCSDT have rather high computational costs and thus often CCSD(T) is used, where perturbation theory is applied to include triples more efficiently.[112] While CCSD(T) often yields highly accurate results,[113] it is usually too slow for practical applications due to the scaling of $O(N_{\mathrm{BF}}^7)$. Thus, additional approximations can be applied like using pair natural orbitals (PNOs) to include only the virtual orbitals most relevant for correlating a pair of electrons.[114] Further, the convergence with regard to basis set size can be accelerated with methods like F12b.[115]

Another approach for including electron correlation is related to many-body perturbation theory (MBPT), where the Hamiltonian is divided into an unperturbed ($\hat{H}_{\mathrm{PT}}^{(0)}$) and a perturbed part ($\hat{H}_{\mathrm{PT}}^{\mathrm{p}}$)

$$\hat{H} = \hat{H}_{\mathrm{PT}}^{(0)} + \lambda_{\mathrm{PT}}\hat{H}_{\mathrm{PT}}^{\mathrm{p}}. \tag{2.34}$$

The parameter λ_{PT} determines the strength of the perturbation. Solving the time-independent Schrödinger equation with this Hamiltonian yields the energy $E_{\mathrm{PT}}^{(0)}$ of the unperturbed system and additional corrections that are additive to the energy and depend on the powers of λ_{PT}:

$$E^{\mathrm{PT}} = \lambda_{\mathrm{PT}}^0 E_{\mathrm{PT}}^{(0)} + \lambda_{\mathrm{PT}}^1 E_{\mathrm{PT}}^{(1)} + \lambda_{\mathrm{PT}}^2 E_{\mathrm{PT}}^{(2)} + \ \ldots. \tag{2.35}$$

The form of the energy terms depends on the choice of the unperturbed Hamiltonian. In Møller–Plesset (MP) perturbation theory, the unperturbed Hamiltonian is chosen to be the sum over the Fock operators $\hat{f}(i)$:[116]

$$E_{\mathrm{MP}}^{(0)} = \sum_{k=1}^{N_{\mathrm{elec}}} \epsilon_k. \tag{2.36}$$

Here, the electron-electron repulsion ($\hat{V}_{\mathrm{ee}}$) is doubly counted. Including additionally the first-order term corrects this, yielding the Hartree-Fock energy:

$$E_{\mathrm{MP}}^{(1)} + E_{\mathrm{MP}}^{(0)} = E^{\mathrm{HF}}. \tag{2.37}$$

The second-order correction

$$E_{\mathrm{MP}}^{(2)} = \sum_{k<l}^{\mathrm{occ.}} \sum_{a<b}^{\mathrm{vir.}} \frac{\left(\langle\psi_k\psi_l|\psi_a\psi_b\rangle - \langle\psi_k\psi_l|\psi_b\psi_a\rangle\right)^2}{\epsilon_k + \epsilon_l - \epsilon_a - \epsilon_b}, \tag{2.38}$$

which sums over the occupied orbitals k and l, and the virtual orbitals a and b, covers a part of the correlation energy. Adding it to the Hartree–Fock energy yields the total energy of the MP2 method. MP2 scales formally with $O(N_{\mathrm{BF}}^5)$ and is generally more accurate than Hartree Fock, but often does not reach CCSD(T) accuracy and performs less accurately for systems with small HOMO-LUMO gaps.

2.2.5 Density Functional Theory

Density functional theory (DFT) aims to deduce the molecular energy from an electron density ($\rho(r)$) rather than a wavefunction. In this way, electron correlation is implicitly incorporated, and lower computational costs can, in principle, be reached as the electron density depends only on three spatial and one spin coordinate rather than on the coordinates of multiple electrons. The basis for DFT forms the Hohenberg-Kohn theorem, stating that there is a one-to-one connection between the electron density of a system and its energy. This connection is given by an unknown functional $F[\rho]$ that is universal and thus identical for every system.[117] However, it does not provide information on the form of this functional. For applying DFT, the energy functional may be partitioned analogous to the Hamiltonian into a kinetic energy part $E_{\mathrm{T}}[\rho]$, the attraction between the electron density and the nuclei $E_{\mathrm{ne}}[\rho]$, and the electron-electron interaction $E_{\mathrm{ee}}[\rho]$, leading to the following energy expression:

$$E^{\mathrm{DFT}} = E_{\mathrm{T}}[\rho] + E_{\mathrm{ne}}[\rho] + E_{\mathrm{ee}}[\rho]. \tag{2.39}$$

With reference to Hartree Fock, $E_{\mathrm{ee}}[\rho]$ may further be split into the Coulomb $E_{\mathrm{J}}[\rho]$ and exchange part $E_{\mathrm{K}}[\rho]$:

$$E^{\mathrm{DFT}} = E_{\mathrm{T}}[\rho] + E_{\mathrm{ne}}[\rho] + E_{\mathrm{J}}[\rho] + E_{\mathrm{K}}[\rho]. \tag{2.40}$$

Now, the different parts have to be defined. For $E_{\mathrm{ne}}[\rho]$, a classical expression can be applied that integrates over all points r of the electron density

$$E_{\mathrm{ne}}[\rho] = -\sum_{i}^{N_{\mathrm{nuc}}} \int \frac{Z_i}{|R_i - r|} dr. \tag{2.41}$$

Similarly, the Coulomb repulsion energy ($E_\mathrm{J}[\rho]$) can be deduced from classical expressions:

$$E_\mathrm{J}[\rho] = \frac{1}{2}\int\int \frac{\rho(r)\rho(r')}{|r-r'|}drdr'. \tag{2.42}$$

The kinetic energy of an electron density can, for example, be computed with the Thomas-Fermi model, but this is generally too inaccurate for molecules and atoms.[118] Thus, commonly Kohn-Sham (KS) DFT is applied that introduces auxiliary orbitals via a Slater determinant, which are directly connected to the density. As representing the exact density requires an infinite number of natural orbitals, it is usually approximated with a set of auxiliary functions:

$$\rho_\mathrm{approx} = \sum_{k=1}^{N_\mathrm{elec}} |\psi_k|^2. \tag{2.43}$$

With this approach, the kinetic energy can be computed similarly to Hartree Fock:

$$E_\mathrm{T}^\mathrm{S} = \sum_{k=1}^{N_\mathrm{elec}} \langle\psi_k| -\frac{1}{2}\nabla^2|\psi_k\rangle. \tag{2.44}$$

The small residual error introduced by this procedure is usually included in the exchange-correlation term $E_\mathrm{XC}[\rho]$ that is commonly split into exchange and correlation ($E_\mathrm{C}[\rho]$) energy:

$$E_\mathrm{XC}[\rho] = (E_\mathrm{T}(\rho) - E_\mathrm{T}^\mathrm{S}(\rho)) + (E_\mathrm{ee}[\rho] - E_\mathrm{J}[\rho]) = E_\mathrm{X}[\rho] + E_\mathrm{C}[\rho]. \tag{2.45}$$

With this, the KS-DFT energy results to

$$E^\mathrm{KS-DFT} = E_\mathrm{T}^\mathrm{S}(\rho) + E_\mathrm{ne}[\rho] + E_\mathrm{J}[\rho] + E_\mathrm{XC}[\rho]. \tag{2.46}$$

For determining $E_\mathrm{XC}[\rho]$, different approximations exist, leading to a variety of density functional approximations (DFAs). One possibility is to derive an expression from a uniform electron gas (UEG) for both the exchange and correlation energy. This leads to the local density approximation (LDA), which is reasonable for periodic, metallic systems but results larger errors for molecules and atoms. The generalized gradient approximation (GGA) considers additionally the first derivative of the electron density, thereby allowing a better description of non-uniform electron gases typically occurring in molecules. The inclusion of higher-order derivatives of the electron density, or the kinetic energy density τ that carries similar information, leads to meta-GGA methods. These three types of approximations, LDA, GGA, and meta-GGA, are computationally less expensive than Hartree Fock, while depending on the system, the results can already be better. However, they have some drawbacks that can lead to severe errors. One of these is the self-interaction error (SIE). While in Hartree Fock, the exchange integral fully cancels the self-interaction of an electron with itself included in the Coulomb energy, in DFAs, the exchange integral is only approximated, potentially causing parts of the electron self interaction energy to remain. The result is a tendency toward more delocalized electron densities. A possible way to reduce this error is to mix the approximated exchange energy of a (meta-)GGA with a term that is computed similarly to Hartree Fock (E_X^HF) from the reintroduced orbitals:

$$E_\mathrm{XC} = (1-a_\mathrm{x})E_\mathrm{X}^\mathrm{(meta-)GGA} + a_\mathrm{x}E_\mathrm{X}^\mathrm{HF} + E_\mathrm{C}^\mathrm{(meta-)GGA}. \tag{2.47}$$

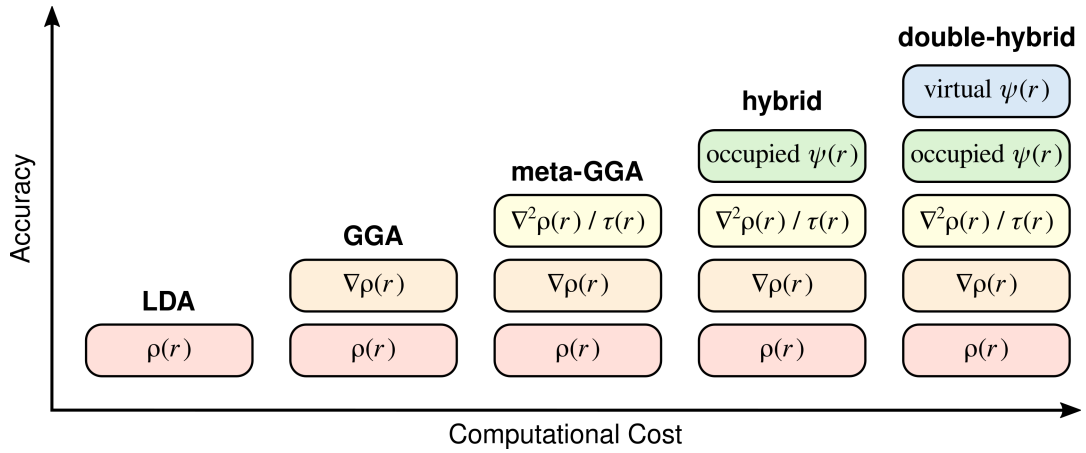

Figure 2.1: Schematic depiction of DFAs categorized according to "Jacob's ladder".[119]

a_x is a mixing factor between 0 and 1. This leads to a class of DFAs called hybrid functionals that comprise a large number of functionals differing by the (meta-)GGA functional it is based on and the amount of Hartree Fock exchange energy that is mixed in. Hybrid functionals usually improve over (meta-)GGAs, especially for cases where the SIE plays a crucial role, e.g., for computing activation energies for transition states. However, due to the Hartree Fock exchange, the formal scaling is $O(N_{BF}^4)$ and thus higher than the (meta-)GGA methods ($O(N_{BF}^3)$). Double hybrid functionals aim to improve hybrids even further. They also change the correlation energy by mixing $E_C^{(meta-)GGA}$ with a correlation energy taken from, e.g., perturbation theory like MP2 (E_C^{MP2}). The overall exchange-correlation energy expression then becomes

$$E_{XC} = (1 - a_x)E_X^{(meta-)GGA} + a_x E_X^{HF} + (1 - a_c)E_C^{(meta-)GGA} + a_c E_C^{MP2}, \tag{2.48}$$

with a_c being a factor between 0 and 1 that determines the amount of MP2 correlation energy mixed in. This causes an increase in formal scaling to $O(N_{BF}^5)$. An overview of the different DFA classes is provided in Figure 2.1 classified according to Jacob's ladder.[119]

Dispersion Corrections

Most DFAs have in common that they miss long-range correlation effects. Thus, instantaneous and correlated fluctuations of the electron charge, important for the London dispersion force, are poorly described.[120] For example, rare gas atoms, which show an attractive force at the equilibrium distance due to dispersion interactions, have a purely repulsive energy curve with many DFAs.[121] As this can introduce a rather significant error, correction schemes have been developed. An example is the Vydrov and van Voorhis nonlocal density-dependent functional kernel (VV10).[122] This kernel aims to describe the dispersion interaction of two electron densities correctly and is added to the E_{XC} of a DFA. The VV10 energy expression contains two parameters that are adjusted to fit the DFA used.

Other commonly used corrections are the D3(BJ) and D4 dispersion models.[123–125] They compute the dispersion energy as a sum over atom pairs. Besides the internuclear distance, also dispersion coefficients are used that are pairwise approximated from precomputed values for the element hydrides. They are combined with Becke-Johnson (BJ) damping and scaled with factors empirically determined

for the individual DFAs. One of the major changes in D4 over D3(BJ) is the introduction of an atomic-partial-charge dependent scaling applied for determining the dispersion coefficients. These partial charges are computed efficiently in each calculation with the classical electronegativity-equilibrium (EEQ) model.[126,127] More details on the DFT-D corrections schemes are given in the Appendix A.1.

Composite DFT Methods

In principle, all DFT methods could be used in large basis sets to achieve their maximum possible accuracy, but this introduces significant computational costs. However, using smaller basis sets introduces errors as mentioned earlier. Methods that aim to balance these two aspects optimally are the composite DFT methods denoted as "3c". They employ smaller basis sets and repair the resulting errors by specially adjusted corrections. Initially started with Hartree Fock (HF-3c),[128] commonly used composite methods are nowadays based on DFT. For example, the r^{2}SCAN-3c (meta-)GGA method employs a tailored triple-ζ basis set combined with refitted D4 and gCP corrections.[129] Further, methods like ωB97X-3c were developed that use a range-separated hybrid functional combined with an adjusted double-ζ basis set optimized to reduce the BSSE.[130]

2.2.6 Semiempirical Quantum Mechanical Methods

Semiempirical quantum mechanical (SQM) methods aim to reduce the computational costs of WFT and DFT methods while maintaining reasonable accuracy. Therefore, different approximations can be applied, such as using minimal basis sets, considering only valence electrons, or applying the zero differential overlap (ZDO) approximation.[131,132] To compensate for the errors introduced by the approximations, parameter-based corrections, optimized using experimental or computational reference data, are applied. SQM methods can generally be derived from WFT, like PM6 and PM7,[133,134] or from DFT leading to tight-binding SQM methods with prominent examples like DFTB[135] and the extended tight binding (xTB)[84] methods GFN1-xTB[48] and GFN2-xTB[49]. In the framework of tight binding, the electron density is split into a reference density ρ_0, usually defined as the sum of spherical, neutral atomic reference densities ($\rho_0 = \sum_A \rho_0^A$) and a density deviation from this reference $\delta\rho$, e.g., occurring upon bonding:

$$\rho = \rho_0 + \delta\rho. \tag{2.49}$$

The total energy consists accordingly of the energy of the reference density (E_0^{TB}), the energy resulting from the density deviation (δE^{TB}) and exchange and correlation terms that are non-separable and depend on the DFT function chosen as starting point. In the case of the extended tight-binding (xTB) methods, an LDA with non-local (NL) correlation energy correction is used as a starting point. Its energy is given as[84]

$$E_{\mathrm{tot}}^{\mathrm{TB}} = E_0^{\mathrm{TB}} + \delta E^{\mathrm{TB}} + E_{\mathrm{XC}}^{\mathrm{LDA}}[\rho] + E_{\mathrm{C}}^{\mathrm{NL}}[\rho, \rho'], \tag{2.50}$$

with $E_{\mathrm{XC}}^{\mathrm{LDA}}[\rho]$ being the exchange-correlation energy and $E_{\mathrm{C}}^{\mathrm{NL}}[\rho, \rho']$ being the energy of the non-local correlation correction that inherently includes dispersion-type interactions. A Taylor expansion around the reference density ($\delta\rho = 0$), with the density deviation typically restricted only to the valence orbitals, yields the energy expression

$$E[\rho] = E^0[\rho_0] + E^1[\rho_0, \delta\rho] + E^2[\rho_0, (\delta\rho)^2] + E^3[\rho_0, (\delta\rho)^3] + \ldots. \tag{2.51}$$

The terms up to the third order include:

$$E^{(0,1,2,3)}[\rho] = \underbrace{\sum_A E_A + E^{(0)}_{\text{rep}} + E^{(0)}_{\text{disp}}}_{=E^{(0)}[\rho_0]} + \underbrace{E^{(1)}_{\text{EHT}} + E^{(1)}_{\text{disp}}}_{\approx E^{(1)}[\rho_0,\delta\rho]} + \underbrace{E^{(2)}_{\text{ES+XC}} + E^{(2)}_{\text{disp}}}_{\approx E^{(2)}[\rho_0,\delta\rho]} + \underbrace{E^{(3)}_{\text{XC}} + E^{(2)}_{\text{disp}}}_{\approx E^{(3)}[\rho_0,\delta\rho]}. \tag{2.52}$$

The term of zeroth order contains the sum over the non-interacting total energies of the atoms (E_{A}), as well as $E^{(0)}_{\text{rep}}$ that is caused by the overlap of the atomic reference densities leading to changes in the Coulomb, exchange, and local correlation energy. $E^{(0)}_{\text{disp}}$ arises from long-range correlation energies only included due to the added non-local correlation term to the LDA-type DFT potential used to derive the energy expression. The first order correction term includes an energy contribution taken from the extended Hückel theory (EHT) $E^{(1)}_{\text{EHT}}$ that is essentially responsible for covalent bonding. However, it does not include changes of the electron density as no Coulomb-type interaction is included, so that the atoms experience the field of the other atoms only by their reference density. In contrast, the first-order dispersion energy $E^{(1)}_{\text{disp}}$ already depends on the difference in density $\Delta\rho$. The second-order energy term introduces another correction for the dispersion energy ($E^{(2)}_{\text{disp}}$) and an electrostatic and exchange-correlation energy term ($E^{(2)}_{\text{ES+XC}}$) that is described by damped Coulomb-type energy expressions. Therefore, the methods now require self-consistent charge (SCC) iterations until optimal partial charges are obtained. The third-order energy term again includes a dispersion correction ($E^{(3)}_{\text{disp}}$), which is not explicitly considered for the GFN methods, and an exchange-correlation term ($E^{(3)}_{\text{XC}}$) that mainly stabilizes relatively highly charged atoms. Both GFN1-xTB and GFN2-xTB employ a minimal basis set. Their parameterizations were focused on geometries, frequencies, and non-covalent interactions (GFN) and they were originally parameterized for elements up to radon. The main difference introduced in GFN2-xTB compared to GFN1-xTB is the inclusion of anisotropic second-order density fluctuation effects that are modeled via cumulative atomic multipole moments (CAMM), making, for example, additional hydrogen or halogen bond corrections used in GFN1-xTB obsolete.[49] Even though the xTB methods proved rather robust and versatile,[85] while being much more efficient than DFT and WFT methods, they still lack accuracy, especially for applications beyond their scope such as the computation of barrier heights.[84] However, they form the basis for some conformer generation tools like *CREST* and can be used for treating various structures that are too large for applying DFT or WFT methods.[44]

2.3 Molecular Mechanics

Molecular mechanics (MM) methods use a different approach to compute molecular energies E. They apply force fields (FFs) that do not explicitly relate the molecular energy to electrons, but rather employ parametric functions of nuclear coordinates with parameters fitted to accurate experimental or computational reference data. With this, significantly lower computational costs compared to WFT, DFT, or SQM methods can be achieved. Depending on the FF used, the energy contributions may vary. A classical form is

$$E_{\text{FF}} = E_{\text{bond}} + E_{\text{bend}} + E_{\text{tors}} + E_{\text{cross}} + E_{\text{vdw}} + E_{\text{ES}}. \tag{2.53}$$

It consists of the sum over energies required to stretch the bonds between two atoms (E_{bond}), the energy for bending the angles (E_{bend}), torsional energy required for rotating around the bonds (E_{tors}), and E_{cross}, the coupling between these three energy terms. Interactions between non-bonded atoms are accounted for by the van-der-Waals E_{vdw} and electrostatic E_{ES} energy. These energy contributions usually require besides parameters and the distances between the nuclei, also partial charges, often approximated with simplified models, for describing the electrostatic energy.[136]

While the computational costs of such a simplified energy description scale only quadratically with the number of atoms, the results are usually less accurate than SQM, DFT, and WFT methods. Further, many FFs are limited to a certain class of compounds, often bio-organic molecules, while there exist only a few general FFs applicable across the periodic table, like the universal force field (UFF)[137] and GFN-FF[50]. The GFN-FF, for example, has the energy expression

$$E_{\text{GFN-FF}} = E_{\text{bond}} + E_{\text{bend}} + E_{\text{tors}} + E_{\text{rep}}^{\text{bond}} + E_{\text{abc}}^{\text{bond}} + E_{\text{IES}} + E_{\text{disp}} + E_{\text{HB}} + E_{\text{XB}} + E_{\text{rep}}^{\text{NCI}}, \quad (2.54)$$

with E_{IES} denoting the isotropic electrostatic energy, $E_{\text{rep}}^{\text{bond}}$ the repulsion forces upon bond compression, and E_{disp} and $E_{\text{rep}}^{\text{NCI}}$ van-der-Waals contributions. Additionally, energy corrections for hydrogen (E_{HB}) and halogen (E_{XB}) bonds are used. Notably, GFN-FF is capable of bond dissociation.

2.3.1 Intermolecular Force Fields

Intermolecular FFs focus on describing NCIs and neglect covalent-bonding terms. An example is the xTB-IFF.[47] This method uses electronic properties of the monomers precomputed by GFN1-xTB or GFN2-xTB to enhance the description of intermolecular interactions. It is especially useful for screening many interaction sites quickly, which is exploited in this work to develop the automated interaction site screening (aISS) algorithm. In the xTB-IFF framework, the energy is computed as

$$E_{\text{int}} = E_{\text{rep}} + E_{\text{ES}} + E_{\text{disp}} + E_{\text{ind}} + E_{\text{CT}}, \quad (2.55)$$

including Pauli repulsion E_{rep}, the electrostatic energy E_{ES}, the dispersion energy E_{disp}, and the induction energy E_{ind} between the fragments. Further, intermolecular charge transfer (E_{CT}) is considered.

The Pauli repulsion energy E_{rep} is a destabilizing energy resulting from the overlap of the molecular orbitals of the fragments. It is computed as a sum over atom pairs and considers atomic distances, effective pair distances accessible via dipole polarizabilities taken from the D4 model, valence electron numbers including the GFN*n*-xTB Mulliken partial charges, and an atomic overlap approximated using atom-centered 1s Slater functions. The electrostatic energy E_{ES} is divided into two contributions, the atom-centered contribution $E_{\text{ES}}^{\text{atom}}$ and a correction for anisotropic electron densities around the atoms $E_{\text{ES}}^{\text{aniso}}$:

$$E_{\text{ES}} = E_{\text{ES}}^{\text{atom}} + E_{\text{ES}}^{\text{aniso}}. \quad (2.56)$$

$E_{\text{ES}}^{\text{atom}}$ is computed by applying the Coulomb law to the model atomic densities resulting from the atom-centered Slater functions. The correction for anisotropic electron densities $E_{\text{ES}}^{\text{aniso}}$ is computed with a damped Coulomb expression of off-centered and atomic charges constructed with the electronic input properties. It is especially important for halogen- and hydrogen bonding, as well as special bonding motifs of π-systems. E_{disp} is computed similar to the D4 model. For the induction energy E_{ind}, a spherical Drude model is applied. Within this approximation, floating charges attached to

the nuclei via a spring constant are equilibrated to counterbalance the respective atomic charge. To accelerate this self-consistent equilibration, it is solved approximately in the xTB-IFF context by evaluating the forces resulting from the electric field of the atoms and from the oscillating charges once, scaling it by a factor of two, and updating the position afterward. The charge transfer energy E_{CT} requires accounting for changes in the initially provided GFNn-xTB electronic properties of the non-interacting fragments. This is done by constructing a combined wavefunction considering the charge transfer and applying the tight-binding Hamiltonian to it. The wavefunction is constructed via a configuration interaction singles (CIS) approach that considers two additional determinants where the energetically highest occupied molecular orbitals (HOMO) of one fragment is replaced by the energetically lowest unoccupied molecular orbitals (LUMO). The extend to which these determinants contribute to the total wavefunction is determined with the HOMO-LUMO gap and a coupling term approximated semiempirically by using the HOMO and LUMO orbital energies and populations of the isolated fragments. By applying the tight-binding Hamiltonian to the combined wavefunction, the energy difference between the isolated wavefunctions yields the charge transfer energy. As charge transfer changes the MO population and thus the atomic partial charges, it has an influence on the induction and electrostatic energies, which are therefore be evaluated after applying the charge-transfer correction. More details on the xTB-IFF method are given in the Appendix A.2.

2.4 Statistical Thermodynamics

As mentioned at the beginning of this chapter, free energies require the energy term $G^{\mathrm{gas}}_{\mathrm{corr.}}(T)$ in addition to the molecular energy E (Equation 2.2). It is computed by applying concepts of statistical thermodynamics, where the free energy can be assumed to be separable into electronic, translational, rotational, and vibrational contributions. All of them are described with discrete energy levels where higher-energy levels are populated more with increasing temperatures. While the electronic contributions are often approximated by only the respective ground state, the other contributions typically have a significant population of multiple states at room temperature. This affects both the enthalpy H_{tot} as well as the entropy S_{tot}, which is accounted for by the thermostatistical correction $G^{\mathrm{gas}}_{\mathrm{corr.}}(T)$:

$$G^{\mathrm{gas}}_{\mathrm{corr.}}(T) = H_{\mathrm{trans}} + H_{\mathrm{rot}} + H_{\mathrm{vib}} - T\left(S_{\mathrm{trans}} + S_{\mathrm{rot}} + S_{\mathrm{vib}}\right). \tag{2.57}$$

The indices *trans*, *rot*, and *vib* refer to the translational, rotational, and vibrational contributions, respectively. For computing the individual contributions, certain approximations are applied, like the ideal gas, and rigid-rotor and harmonic-oscillator (RRHO) approximation. These assumptions lead to the following equations for a non-linear molecule:

$$H_{\mathrm{trans}} = \frac{5}{2}\mathrm{R}_{\mathrm{gas}}T, \tag{2.58a}$$

$$H_{\mathrm{rot}} = \frac{3}{2}\mathrm{R}_{\mathrm{gas}}T, \tag{2.58b}$$

$$H_{\mathrm{vib}} = \mathrm{R}_{\mathrm{gas}} \sum_{i=1}^{3N_{\mathrm{nuc}}-6} \left(\frac{\mathrm{h}\nu_i}{2\mathrm{k}_{\mathrm{B}}} + \frac{\mathrm{h}\nu_i}{\mathrm{k}_{\mathrm{B}}} \frac{1}{e^{\mathrm{h}\nu_i/(\mathrm{k}_{\mathrm{B}}T)} - 1}\right), \tag{2.58c}$$

$$S_{\text{trans}} = \frac{5}{2}\mathrm{R}_{\text{gas}} + \mathrm{R}_{\text{gas}}\ln\left(\frac{V}{\mathrm{N}_{\text{Av}}}\left(\frac{2\pi m_{\text{tot}}\mathrm{k}_{\mathrm{B}}T}{\mathrm{h}^2}\right)^{\frac{3}{2}}\right), \tag{2.58d}$$

$$S_{\text{rot}} = \frac{3}{2}\mathrm{R}_{\text{gas}} + \mathrm{R}_{\text{gas}}\ln\left(\frac{\sqrt{\pi}}{\sigma_{\text{rs}}}\left(\frac{8\pi^2\mathrm{k}_{\mathrm{B}}T}{\mathrm{h}^2}\right)^{\frac{3}{2}}\sqrt{I_1 I_2 I_3}\right), \tag{2.58e}$$

$$S_{\text{vib}} = \mathrm{R}_{\text{gas}}\sum_{i=1}^{3N_{\text{nuc}}-6}\left(\frac{\mathrm{h}\nu_i}{\mathrm{k}_B T}\frac{1}{e^{\mathrm{h}\nu_i/(\mathrm{k}_B T)}-1} - \ln\left(1-e^{\mathrm{h}\nu_i/(\mathrm{k}_B T)}\right)\right). \tag{2.58f}$$

k_B denotes the Boltzmann constant, h the Planck constant, R_{gas} the ideal gas constant, ν_i the vibrational frequencies of the different vibrational levels, N_{Av} the Avogadro constant, m_{tot} the total molecular mass, V the volume of one mol of an ideal gas, σ_{rs} the rotational symmetry number, and I_1, I_2, and I_3 the three moments of inertia. For a linear molecule, the rotational degrees of freedom are reduced, leading to a smaller overall contribution from the rotational terms:

$$H_{\text{rot}} = \mathrm{R}_{\text{gas}}T, \tag{2.59a}$$

$$S_{\text{rot}} = \mathrm{R}_{\text{gas}} + \mathrm{R}_{\text{gas}}\ln\left(\frac{8\pi^2\mathrm{k}_{\mathrm{B}}TI}{\mathrm{h}^2\sigma_{\text{rs}}}\right). \tag{2.59b}$$

As S_{vib} approaches infinity for small vibrational frequencies, often a modified version of the RRHO (mRRHO) is applied, addressing this limitation.[138] Thereby, the contributions of low-lying frequencies are replaced by the corresponding rotational entropy, interpolated by a switching function. Notably, H_{vib} has a non-zero contribution at 0 K (zero-point vibrational energy) to the free energy. Further, these corrections have a huge influence on the association of molecules since the degrees of freedom are usually reduced upon this process.

While the translational and rotational contributions can be computed with information about the molecule's atomic composition and geometry, the vibrational contributions additionally require knowledge about the normal modes. These depend on the bond strengths and thus require information about how the energy changes when stretching and compressing bonds. This information is included in the mass-weighted Hessian matrix (**F**), which is formed with the second derivative of the electronic energy with respect to the atomic positions. The individual matrix elements ij are given by:

$$F_{ij} = \frac{1}{\sqrt{M_i M_j}}H_{ij} = \frac{1}{\sqrt{M_i M_j}}\left(\frac{\partial^2 E}{\partial r_i \partial r_j}\right), \tag{2.60}$$

with M_i and M_j being the atomic masses and r_i and r_j the respective positions of the two atoms i and j. The eigenvalues of the mass-weighted Hessian ϵ_{fi} are used to determine the harmonic vibrational frequencies ν_i:

$$\nu_i = \frac{1}{2\pi}\sqrt{\epsilon_{fi}}. \tag{2.61}$$

For computing the Hessian, every electronic structure method can, in principle, be used as well as FFs. However, when the investigated structure is not a minimum on the PES of the respective method, imaginary modes will result, potentially introducing an error to $G^{\text{gas}}_{\text{corr.}}(T)$.

2.5 Solvation Effects

When modeling molecular processes in solution, the solvent influence on the free energy has to be considered. Computationally, this is typically done with the solvation free energy, δG_{solv}, which is the energy required to transfer a molecule from an ideal gas to solution. It can be added to free energies of systems in the gas phase to get accurate free energies in solution $G_{\text{tot}}^{\text{solu}}$ (Equation 2.3). To compute the solvation free energy, two common approaches exist: explicit and implicit solvent models. Both will be outlined in the following.

2.5.1 Implicit Solvent Models

Idea

Implicit solvent models treat the solvent as a structureless polarizable continuum characterized by properties such as the dielectric constant ϵ_{e} of the solvent.[139] This allows efficient treatment of solvation and routine integration in most simulations. However, implicit solvent models also have certain drawbacks like a tendency to overestimate intramolecular interactions compared to intermolecular ones, potentially causing errors when predicting structures.[67,68] Further, complicated surface curvatures of the solute's structure can introduce larger errors in the solvation free energy.[69]

To compute the solvation free energy with implicit solvent models, different contributions are modeled. First, a cavity has to be created in the continuum, requiring energy (δG_{cavity}). Subsequently, the solute is placed inside, leading to a polarization of the continuum by the solute and thus an electrostatic stabilization (δG_{elec}). Interactions like dispersion, which are not covered by the electrostatic energy, are additionally accounted for. A rather simple approach in this regard is to combine these contributions with δG_{cavity} to form a non-electrostatic energy (δG_{ne}) and to approximate it proportional to the cavity size by summing over the number of atoms N_{atoms}:.[140]

$$\delta G_{\text{ne}} = \sum_{A=1}^{N_{\text{atoms}}} \gamma_A^{\text{S}} S_A^{\text{S}}. \tag{2.62}$$

Whereas γ_A^{S}, an element-specific parameter, is usually fitted to experimental data, the surface area of atom A (S_A^{S}) is often determined from the solvent-accessible surface area (SASA) of the solute. Constructing the SASA can be done, e.g., by moving a probe sphere with a radius appropriate for the modeled solvent over the van-der-Waals surface composed of overlapping spheres of element-specific radii.[141]

The solvation free energy of such an approximation yields

$$\delta G_{\text{solv}} = \delta G_{\text{elec}} + \delta G_{\text{ne}}. \tag{2.63}$$

However, such a simplified approach can introduce large errors, which has led to the development of more sophisticated approaches.[142] Depending on the theoretical framework and model used, additional terms like hydrogen bonding corrections and reference state conversion factors are included.[143–145]

Electrostatic Energy

The electrostatic energy results from the interaction of the polarized continuum with the solute. Computing it requires a way to describe the polarization of the solute and the resulting interaction with the continuum. The solute can be modeled with MM methods and partial charges, as well as QM methods that can be used to converge the electron density of the solute in combination with the polarization of the continuum, as the polarization of the solute and continuum are mutually dependent. To describe the interaction of the solute with the continuum, the nonhomogeneous Poisson (NP) equation can be employed:

$$\nabla \cdot \left(\epsilon_{\mathrm{e}}(r)\nabla\Phi(r)\right) = -4\pi\rho_{\mathrm{c}}(r). \tag{2.64}$$

It connects the electrostatic potential Φ at a position r with the respective charge distribution ρ_{c}. The dielectric permittivity ϵ_{e} is often approximated to equal unity inside the cavity ($\epsilon_{\mathrm{cavity}}$) and that of the solvent (ϵ_{solv}) outside, which is known for common ones.[145] Further, the NP equation can also be modified to include ions in the solvent, resulting in the Poisson–Boltzmann (PB) equation. With the NP and PB equations, the electrostatic potential of the solute in the gas phase ($\epsilon_{\mathrm{gas}} \approx \epsilon_{\mathrm{vac}} = 1$) and in the solvent cavity can be computed. The difference between these potentials, called reaction field (Φ_{reac}), can be used to compute the electrostatic component of the solvation free energy:

$$\delta G_{\mathrm{elec}} = \frac{1}{2}\int \rho_{\mathrm{c}}(r)\Phi_{\mathrm{reac}}(r). \tag{2.65}$$

Solving the NP or the PB equations introduces significant computational costs as they contains slowly converging sums.[146] Thus, different approximations were developed to fit either FFs, SQMs, or DFT and WFT. The ones used in this thesis are outlined in the following.

GBSA and ALPB

Efficient methods like FFs and SQMs require fast implicit solvent models. One of them is the generalized Born (GB) model[147], which assumes a spherical cavity and combines the Born model[148] with the Coulomb interactions of atomic partial charges Q to yield:

$$\delta G_{\mathrm{elec}} = -\frac{1}{2}\left(\frac{1}{\epsilon_{\mathrm{cavity}}} - \frac{1}{\epsilon_{\mathrm{solv}}}\right)\sum_{A=1}^{N_{\mathrm{atoms}}}\sum_{B=1}^{N_{\mathrm{atoms}}}\frac{Q_A Q_B}{f_{AB}^{\mathrm{GB}}}. \tag{2.66}$$

An example of a typical switching function f_{AB}^{GB} was proposed by Still et al.,[147]

$$f_{AB}^{GB,Still} = \left((R_A - R_B)^2 + a_A^{\mathrm{Born}} a_B^{\mathrm{Born}} \cdot \exp\left(-\frac{(E_{\mathrm{A}} - R_B)^2}{4a_A^{\mathrm{Born}} a_B^{\mathrm{Born}}}\right)\right)^{\frac{1}{2}}, \tag{2.67}$$

with R_A and R_B being the positions of atoms A and B, and a_A^{Born} and a_B^{Born} the respective Born radii.[147] Equation 2.66 includes terms where $A = B$, representing a stabilization of charges due to the continuum, as well as cross-terms corresponding to charge-charge interactions screened in the continuum.[149] Combined with the linear approximation for δG_{ne} in Equation 2.62, this forms the basis for generalized Born surface area (GBSA) models, introducing no significant computational overhead for FFs and SQM methods. However, such GB models do not incorporate important physics so that,

e.g., swapping ϵ_{cavity} and ϵ_{solv} results in only a sign change of δG_{elec} even though such symmetry is not seen in nature.[150] Improving on this, analytical solutions to the linearized–Poisson–Boltzmann (ALPB) equation were developed that additionally incorporate dependencies on the shape and size of the molecule. This includes the electrostatic size A_{ES} that provides a relationship of the shape and electrostatic energy of the solute:[151]

$$\delta G_{\text{elec}} = -\frac{1}{2}\left(\frac{1}{\epsilon_{\text{cavity}}} - \frac{1}{\epsilon_{\text{solv}}}\right)\frac{1}{1+\alpha_A\beta_A}\sum_{A=1}^{\text{atoms}}\sum_{B=1}^{\text{atoms}} Q_A Q_B \left(\frac{1}{f_{AB}^{GB,Still}} + \frac{\alpha_A\beta_A}{A_{\text{ES}}}\right). \tag{2.68}$$

α_A is originally suggested to be 0.571412 and $\beta_A = \epsilon_{\text{cavity}}/\epsilon_{\text{solv}}$.[151]

CPCM and COSMO

Computationally more expensive DFT and WFT methods allow the use of more elaborate implicit solvent models, evaluating molecule-shaped cavities to improve the description of solvation. Thereby, the three-dimensional problem of solving the Poisson equation is transformed into an apparent surface charge problem on a two-dimensional solvent surface.[152] In principle, this is done by tessellation of the cavity surface and assigning each fraction a charge. The surface charges $\sigma_{\text{s}}(r_{\text{s}})$ at a point r_{s} on the surface are directly related to the potential inside the cavity

$$\sigma_{\text{s}}(r_{\text{s}}) = \frac{\epsilon_{\text{solv}} - \epsilon_{\text{cavity}}}{4\pi\epsilon_{\text{solv}}}\frac{\partial}{\partial\vec{n}}\left(\Phi_{\text{cavity}}^{\text{elec}} + \Phi_{\sigma}^{\text{elec}}\right), \tag{2.69}$$

where $\vec{n}$ denotes the unit vector. The charges arise due to the electrostatic potential of the solute inside the cavity ($\Phi_{\text{cavity}}^{\text{elec}}$) and the surface charges ($\Phi_{\sigma}^{\text{elec}}$) perpendicular to the surface pointing outwards.[153] The potential resulting from the surface charges at a point r in space,

$$\Phi_{\sigma} = \int \frac{\sigma_{\text{s}}(r_{\text{s}})}{|r - r_{\text{s}}|} dr_{\text{s}}, \tag{2.70}$$

can then be added to the Hamiltonian of the underlying electronic structure method to determine the polarization of the solute and the surface charges self-consistently. Such a computation yields the molecular energy of the solute, including the solvation contribution δG_{elec}. For the conductor-like polarizable continuum model (CPCM)[154] and the conductor-like screening model (COSMO),[155] the solvent is set to have an infinite dielectric constant ($\epsilon_{\text{solv}} = \infty$) to get the ideal screening charges $\sigma_{\text{s}}^{*}(r_{\text{s}})$. These are scaled with a function dependent on ϵ_{solv}:

$$\sigma_{\text{s}}(r_{\text{s}}) = \frac{\epsilon_{\text{solv}} - 1}{\epsilon_{\text{solv}} + k_{\text{PCM}}}\sigma_{\text{s}}^{*}(r_{\text{s}}). \tag{2.71}$$

The scaling factor k_{PCM} depends on the model and is typically between 0 in case of CPCM and 0.5 for COSMO.[153] The more sophisticated COSMO-RS[143,144] and SMD[145] models are based on the results of COSMO and CPCM. For example, in COSMO-RS, corrections for the errors due to the ideal conductor treatment are applied, as well as hydrogen bond and ring corrections. In SMD, an improved atomic-surface-tension-dependent δG_{ne} description is used. These models generally improve over their underlying model, but rely more on training data as they include more parameters.[145,156]

2.5.2 Explicit Solvent Models

Explicit solvent models simulate the solute solvated by individual solvent molecules. To accurately account for the solvent bulk, often periodic-boundary conditions are employed. While this approach closely resembles real systems, it is technically involved, and the large system sizes usually restrict the application of accurate but expensive methods like DFT and WFT.[69] Further, the complex PESs of these systems significantly complicate the conformer generation and structure sampling.[157] A common approach to estimate solvation free energies from explicitly solvated systems is based on the free-energy perturbation using Zwanzig's equation:[158]

$$\Delta G_{S1\rightarrow S2} = G_{S2} - G_{S1} = -\mathrm{k_B}T \cdot \ln\langle exp\left(-\frac{G_{S2} - G_{S1}}{\mathrm{k_B}T}\right)\rangle_{S1}. \tag{2.72}$$

With this, the free energy difference for transitioning a system from state $S1$ to state $S2$ ($\Delta G_{S1\rightarrow S2}$) can be computed with an average over the energies G_{S2} of the ensemble of structures in state $S2$ for which also the respective energies in state $S1$ are computed. In principle, by defining state $S1$ to be the solute in the gas phase and $S2$ to be the solute in the solvent, the energy of transferring one solute molecule from the gas phase to solution results, which can directly be transformed into the solvation free energy.[159] This, however, would be rather inaccurate, as the change in free energy is not necessarily just the difference between the two averaged states, but the function connecting them may be more complex. Thus, for the process of solvation, the path between the two states is split into small steps where the solute and the solvent gradually interact, defined by a parameter that ranges from 0 (no interaction of solute and solvent) to 1 (fully interacting solute and solvent).[158,160] An improvement to this method can be achieved by thermodynamic integration (TI), where not finite differences in energy function are considered, but their differentiated versions.[161,162] Also, the Bennett Acceptance Ratio (BAR) method is commonly employed, where two neighboring states are used to estimate the free energy change with a non-linear equation.[163]

2.5.3 Hybrid Cluster–Continuum Models

While implicit solvent models usually cause only a negligible increase in computational costs and can be routinely integrated in most simulations, they also have several drawbacks and limitations as previously mentioned. In contrast, explicit solvent models could potentially be more accurate, but they are technically involved and often require modeling large system sizes, which limits the scope of applicable methods to efficient, but less accurate methods. Further, covering the dynamics of such large systems accurately requires extensive sampling.[157] Due to the drawbacks of both classes, hybrid cluster–continuum models were developed that aim to combine the strengths of implicit and explicit models.[63–65] They model a few solvent molecules explicitly and cover the bulk solvent by a continuum. With this approach, the expense for conformational sampling is reduced, and also computationally more expensive methods like DFT and WFT can be applied to the system, while solute–solvent interactions can potentially be modeled accurately.

2.6 Conformers

For generating and evaluating conformers, understanding the concept of the PES becomes useful. It maps the molecule's energy with the spatial arrangement of its atoms.[22] An example is illustrated for *n*-pentane in Figure 2.2. Here, the energy minima on the PES correspond to different stable conformers (A, B, C, and D), which are interconnected through rotations around single bonds. The global minimum (A) is the lowest energy conformer, while the other local minima (B, C, and D) are energetically higher.

Identifying the conformers that are present in a real system depends on its energy relative to the global energy minimum (conformational energy) and the system's temperature. The probability p_I of finding a conformer I in the system can be computed via the Boltzmann weight over the total number of conformers N_{conf}:

$$p_I = \frac{e^{-\frac{G_I}{k_B T}}}{\sum_{i=1}^{N_{\text{conf}}} e^{-\frac{G_i}{k_B T}}}. \tag{2.73}$$

Consequently, higher-energy conformers are more likely to be found the higher the temperature of the system is. Hence, it is of practical relevance to find all conformers that are energetically close to the energy minimum and thus might have a significant Boltzmann weight. These conformers contribute to the total free energy of the structure ensemble:[35]

$$G_{\text{ensemble}} = \overline{G} + G_{\text{conf}}, \tag{2.74}$$

where $\overline{G}$ is the Boltzmann-weighted average of the individual conformer free energies G

$$\overline{G} = \sum_{I}^{N_{\text{conf}}} p_I G_I, \tag{2.75}$$

and G_{conf} a contribution arising from the conformational entropy caused by the population of different conformational minima.[35,164–166]

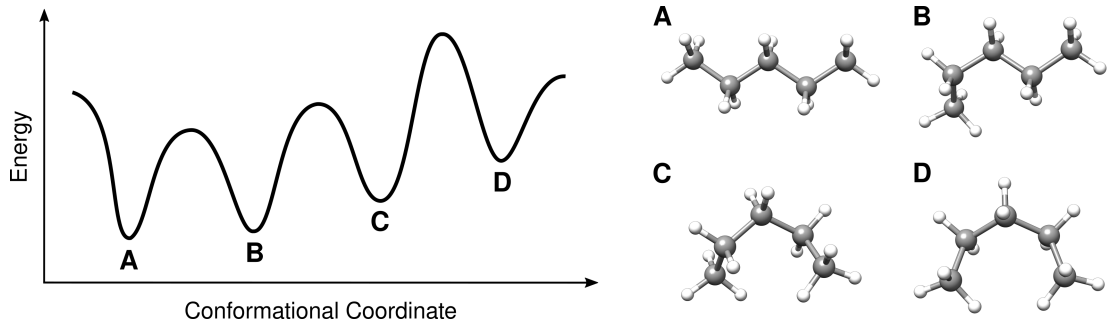

Figure 2.2: Schematic representation of the 2D potential energy surface of the staggered *n*-pentane conformers A-D.

2.6.1 Exploring the Potential Energy Surface

For generating conformers, methods for moving on the PES of a system can be used. They often neglect thermostatistical contributions due to technical reasons. How the energy derivative can be computed with various methods like WFT, DFT, SQM, and MM, and used for moving on the PES, is outlined in the following.

Geometry optimization

Geometry optimizations converge a given structure to the nearest local minimum on the PES. Existing algorithms for this rely, in principle, on the same concept: They iteratively change atomic positions until the force $\vec{F}$ acting on the atoms is nearly zero, or until the Hessian contains no imaginary modes anymore.[167] For example, the steepest descent method moves the atoms step-wise downhill on the PES by following the force, which is computed as the opposite of the gradient $\vec{g}$

$$\vec{F} = -\vec{g} = -\left(\frac{\partial E}{\partial r}\right). \tag{2.76}$$

However, this procedure tends to converge rather slowly, leading to significant computational overhead. More efficient methods are based on the Newton-Raphson approach, where the Hessian is considered for accelerating the optimization. As computing the Hessian is rather computationally costly, it is often approximated in Quasi-Newton methods like the Broyden-Fletcher-Goldfarb-Shanno (BFGS)[168] and Limited-memory BFGS (L-BFGS)[169] algorithms. The step length of each geometry optimization cycle is usually adjusted dynamically for optimal convergence with algorithms like the trust radius method (TRM).[168] They aim to prevent overshooting resulting from too large step sizes and a large number of optimization cycles caused by small step sizes.[170]

Molecular Dynamics Simulations

Molecular Dynamics (MD) simulations model the dynamics of a system by propagating it over time. Thereby, depending on the temperature applied during the simulations, energy barriers can be overcome that, e.g., transform one conformer into another. They are based on evaluating Newton's second equation in the differential form

$$\vec{F} = m_\mathrm{p}\frac{d^2 r}{dt^2}. \tag{2.77}$$

In an MD simulation, the force $\vec{F}$ acting on a particle with the mass m_p is used to predict the position r after the time step t. At this position, a new force is evaluated and the system's propagation in time is continued step-wise.

When running MD simulations, choosing the time step for evaluating the next force is crucial, as too small time steps lead to increased computational costs while too large steps might lead to atom collision and thus unstable MD simulations. To address this, algorithms exist that use a fixed time step for the whole system, but also methods like adaptive verlet,[171] or Multiple Time Step (MTS)[172] were developed that adjust the time steps, e.g., individually for all atoms depending on their speed. Further, algorithms like SHAKE can be used to constrain certain bond lengths that would vibrate rather fast

(e.g., bonds involving hydrogen atoms) allowing larger time steps without instability problems.[173]

Controlling the temperature during MD simulations is done by a thermostat that adjusts the velocity and thus the kinetic energy of the atoms. Examples are the Berendsen thermostat, which scales the velocity directly with a time-dependent damping factor,[174] or the Nosé-Hoover thermostat, which introduces a fictitious thermal reservoir maintaining correct statistics for NVT ensembles (constant number of particles, volume, and temperature).[175,176]

Metadynamics Simulations

While MD simulations can be used to overcome energy barriers, they are rather inefficient for exploring the potential energy surface, as energetically favored structures will dominate the simulations, making the exploration of PES parts further away from them complicated. Furthermore, overcoming rather high barriers is hardly or not even possible. Addressing these limitations, meta-dynamic (MTDs) simulations were developed.[177] The principal idea is the addition of a Gaussian-shaped, history-dependent potential to the energy of the systems. The potential $E_{\text{bias}}^{\text{MTD}}$ depends on two parameters, k_{MTD} and α_{MTD}, and a collective variable Δ^{MTD} like the root mean square deviation (RMSD) between a reference structure i_{r} and the actual structure in the simulation:

$$E_{\text{bias}}^{\text{MTD}} = \sum_{i_{\text{r}}}^{n_{\text{MTD}}} k_{\text{MTD}} e^{-\alpha_{\text{MTD}} \Delta_{i_{\text{r}}}^{\text{MTD}}}. \tag{2.78}$$

This bias potential increases with the number of reference structures n_{MTD}. By defining k_{MTD} to be positive, a repulsive potential results, with a strength dependent k_{MTD}. α_{MTD} defines the width of the biasing potential. This potential increases the closer a structure is to an already visited one, and thus the simulation is pushed away from already simulated areas of the PES, significantly accelerating its exploration.

Conformer Generation

For generating conformers, different approaches exists. Besides employing knowledge-based information about optimal arrangements,[178] physically motivated methods that rely on FF and SQM methods for PES exploration became increasingly important. One of these is *CREST*, which combines multiple MTD and MD simulations with geometry optimizations and efficient filtering methods to identify unique conformers as well as rotamers.[44,179] Thereby, multiple MTD simulations are run in parallel, whereby snapshots are taken that are optimized. To sample locally around the energy minimum, additional MD simulations in conjunction with geometry optimizations are performed. Techniques like Z-matrix crossing can also be used to find additional conformers. By default, the procedure is restarted up to five times each time a new energy minimum is found. To cover different molecular systems, also modified versions of this algorithm exist, e.g., the NCI-MTD runtype that is optimized for systems composed of multiple molecules and applies a repulsive wall potential to prevent dissociation of non-covalently bound molecules.

CHAPTER 3

Automated and Efficient Generation of General Molecular Aggregate Structures

Christoph Plett,* Stefan Grimme*

Received: October 01, 2022
Published online: November 17, 2022

Reprinted (adapted) in Appendix B with permission‡ from:
C. Plett and S. Grimme, Angew. Chem. Int. Ed. **62** (2023) e202214477.

Own contributions

- Optimization of the original xTB-IFF algorithm
- Integration of the GFN methods
- Development of additional features
- Implementation in the *xtb* program suite
- Compilation of the test set and showcases
- Performing of all calculations
- Evaluation of the results
- Writing of the manuscript

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

The question of how multiple molecules arrange relative to each other and interact non-covalently is often essential for understanding molecular properties and processes. Hence, there is a great interest in computational tools capable of finding realistic intermolecular orientations. While conformer generation tools like *CREST*[44] can be used for this task, they are usually designed to be generally applicable rather than specialized on intermolecular interactions and thus might become inefficient for large systems composed of multiple molecules. Consequently, docking tools specifically designed to generate structures of interacting molecules by docking monomers have become of great importance. While they can be much more efficient than general conformer generation tools, most existing docking tools focus mainly on biomolecules and are often not applicable to other types of molecular systems.[52] To address this gap, the automated Interaction Site Screening (aISS) algorithm was developed and implemented in the open-source *xtb* software package.[180] It enables efficient and automatic docking of two molecules with almost arbitrary elemental composition to generate dimers, oligomers, and aggregates subsequently. It combines the intermolecular force field xTB-IFF[47] for screening potential interaction sites with GFN*n*-xTB[48,49] or GFN-FF[50] for geometry optimizations.

The first step of the aISS algorithm after processing the input geometries is the computation of electronic properties of the monomers with GFN1-xTB[48] or GFN2-xTB[49]. These are required by xTB-IFF to evaluate interaction energies in the following steps. Next, the algorithm performs a grid-based prescreening of potential interaction sites around one molecule. This involves using neutral and artificially charged rare-gas atoms to probe favorable positions in terms of dispersion and electrostatic interactions. Based on the thereby gained information, general orientations of the two monomers are generated, as well as structures with chemically relevant interaction motifs such as pocket filling and π-π stacking interactions. Afterward, the most favorable structures in terms of xTB-IFF interaction energy are refined with a genetic algorithm that is based on random crossovers and mutations, allowing structure exploration beyond static grid positions. To ensure computational feasibility, intramolecular conformations are fixed during the screening. To account for the influence of the NCIs on the monomer geometries, some of the best structures found so far (per default 15) are fully optimized with GFN1-xTB, GFN2-xTB, or GFN-FF, optionally including the implicit solvation models ALPB or GBSA.[181] An additional runtype is available, in which all energetically favorable structures from the genetic algorithm are optimized to yield an ensemble as complete as possible.

To quantify the performance of the aISS algorithm, a diverse set of seven dimer and trimer systems was compiled including heavy main-group elements and transition metals, as well as charged and open-shell systems. aISS was tested employing the default and ensemble runtype against *CREST*, all of them using GFN2-xTB. Subsequent r^{2}SCAN-3c[129] DFT refinements of the structures and energies was done for more accurate results. Overall, aISS yielded structures of similar quality to those from *CREST*. In some cases, like a Ti-complex dimer,[182] aISS yielded structures less favorable in terms of GFN2-xTB interaction energy, but the subsequent DFT refinements lead to energetically similar structures. Contrary, aISS also generated structures of lower energy than *CREST*, especially for electronically challenging systems where the GFN methods are prone to errors. For example, monomers of a cationic hypervalent iodine complex[183] had to be constrained during the *CREST* runs to avoid bond breaking during the high-energy MTD simulations, which limited the respective structure search. Additionally, for larger and electronically more challenging systems like a Hg-complex trimer,[184] *CREST*'s MTD sampling failed to identify the global minimum within reasonable simulation times, resulting in an energetically significantly higher structure than those resulting from aISS.

Beyond its use for for such challenging cases, the test set outlined another significant advantage of

aISS: its efficiency. For example, a *CREST* run for a standard organic molecule with 200 atoms in total[185] required over 64 h on 14 standard CPU cores, while the aISS algorithm identified the same global minimum in just 15 min. This high efficiency allows not only the routine treatment of small- to mid-sized molecules, but also expands the capabilities of computational structure generation toward large and diverse molecular systems. As a showcase, a rhodium-organic cuboctahedra[186] was docked into the self-assembled Goldberg polyhedron $Pd_{48}L_{96}(BF_4)_{96}$[187] resulting in a 4296-atom system with a diverse elemental composition including Se, B, Rh, and Pd.

Further, the aISS algorithm can be used for the exploration of reactive sites and low-barrier reactions. For example, a micelle known to cleave under aqueous acidic conditions[188] was studied by docking an oxonium ion with aISS employing GFN2-xTB and the ALPB(water) solvation model. Despite the many possible basic functional groups, aISS identified the correct protonation site and the experimentally observed proton transfer occurred during its geometry optimizations.

A special feature of aISS is the directed docking mode, which allows the user to define interaction sites that should preferably be occupied during docking. This is extremely useful when studying reactive sites that might energetically not be favorable and for restricting the search space based on chemical intuition. As a demonstration, a Buchwald-Hartwig amination[189] was investigated, where unbiased docking yielded an intermediate that does not undergo further reactions. With the directed docking, the reactive intermediate that can be used for further reaction modeling was found. Similarly, a faujasite-based zeolite was investigated in which different positions of sodium cations were shown to have a significant influence on the reactivity, e.g., of a following Diels-Alder.[190]

In summary, the aISS algorithm significantly expands the capabilities of automated structure generation for interacting molecules by offering high efficiency, broad applicability across elements up to radon, and unique features like directed docking. It enables routine assessment of NCIs in computational workflows and allows to study systems beyond the scope of common docking tools.

CHAPTER 4

Automated Molecular Cluster Growing for Explicit Solvation by Efficient Force Field and Tight Binding Methods

Sebastian Spicher,[*,†] Christoph Plett,[*,†] Philipp Pracht,[*] Andreas Hansen,[*] Stefan Grimme[*]

Received: March 10, 2022
Published online: April 28, 2022

Reprinted (adapted) in Appendix C with permission[‡] from:
S. Spicher, C. Plett, P. Pracht, A. Hansen, and S. Grimme, J. Chem. Theory Comput. **18** (2022) 3174.

Own contributions

- Improvements of the QCG algorithm
- Integration of *CREST* workflows for conformer search
- Implementation in the *CREST* program suite
- Performing of all calculations except for the microsolvation examples
- Evaluation of the results
- Writing of the manuscript

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

An important example of non-covalently interacting molecules is solvation. Modeling solvent molecules poses specific challenges, such as ensuring a physically meaningful distribution of solvent molecules that goes beyond the purely energetic criteria discussed in Chapter 3. Therefore, the Quantum Cluster Growth (QCG) algorithm was developed and implemented in the open-source *CREST* program.[44,66,179] This hybrid cluster–continuum model automates and streamlines the inclusion of explicit solvent molecules in quantum chemical calculations by automatically generating and evaluating solute–solvent clusters. A key strength of QCG is its robustness across a wide variety of solutes and solvents, enabled by the underlying GFN method family [48–50] supporting elemental compositions up to radon.

The QCG algorithm consists of several modular steps. First, a single cluster is grown by sequential solvent addition using xTB-IFF-based docking[47] and geometry optimizations employing the GFN methods. The size of the cluster is either specified by the user or automatically determined by energy convergence criteria. During the growth, two repulsive wall potentials are applied to ensure realistic solute–solvent clusters. An inner wall potential keeps the solute centered while providing enough conformational flexibility for the solute to adapt to the growing solvent shell. The outer potential acts on the entire cluster, preventing solvent molecules from clustering without the solute. These potentials dynamically adapt based on solute and solvent properties such as volume, shape, and the increasing number of solvent molecules. After initial cluster growth, the second step is an ensemble generation to refine the initially obtained structure and to identify further relevant conformers. Thereby, either a single MD or MTD simulation combined with geometry optimizations of snapshots can be used or more sophisticated *CREST* workflows, all of which employ the two wall potentials. In a final step, solvation free energies can be computed. Therefore, additional clusters of pure solvent molecules are generated and δG_{solv} is obtained by subtracting the Boltzmann-weighted free energy of the solvent clusters and the gas-phase solute from that of the solute–solvent clusters.

The performance of QCG was evaluated using an example of phenylalanine solvated by 60 water molecules. As competitors to the QCG growth, structures generated with the TIP3P water model[191] and a space-filling algorithm were compared, both optimized with GFN2-xTB. The cluster generated by QCG was energetically most favorable and provided an even distribution of water molecules around the solute. The structure was about 30 kcal mol^{-1} lower in formation energy than the one generated employing TIP3P, not only with GFN2-xTB, but also on the more accurate r^{2}SCAN-3c[129] DFT energy surface. The space-filling algorithm generated the least meaningful structure, demonstrating the necessity of energetic considerations for cluster growth. The effect of an additional ensemble generation was demonstrated exemplarily for ethanol in acetonitrile. Even for smaller clusters of up to eight solvent molecules, the ensemble generation improved the structures energetically by up to 4 kcal mol^{-1} and the energy improvement increased for larger cluster sizes.

Evaluating the solvation free energies for typical organic solutes and solvents, the QCG results showed overall larger errors compared to the implicit solvent model COSMO-RS.[192] Nevertheless, the results were promising for such a direct approach and serve as a good starting point for future improvements.

A possible application of the QCG algorithm is generating the structures of microsolvation approaches, which was demonstrated using an example of benzoic acid and aminobenzothiazole solvated by three water molecules. QCG identified the relevant structures reported in previous studies within minutes on a standard desktop computer, making it significantly more efficient than other microsolvation strategies.[193] Further, the use of QCG for structure elucidation was shown for the antibiotic bacillaene.[194] MD simulations with GFN2-xTB and the implicit ALPB(water) solvent

model[181] showed almost no bond breaking of the two intramolecular hydrogen bonds present in the gas phase, reflecting that implicit solvent models tend to overestimate intramolecular hydrogen bond strengths.[67,68] During a similar simulation with explicit water molecules added by the QCG algorithm, insertion of water molecules into the intramolecular hydrogen bonds was observed and thus completely different structure motifs resulted. Lastly, QCG was used for simulating infrared (IR) spectra, where explicit solvation is essential for accurately reproducing spectral features. The implicit solvent model COSMO[155] failed to reproduce the experimentally measured spectra of dimethyl sulfoxide (DMSO), acetonitrile, and chloroform due to additional or missing peaks. In contrast, QCG-generated structures resulted in spectra that closely matched experimental measurements.

In conclusion, QCG is a powerful tool for automated and efficient inclusion of explicit solvent molecules covering a wide range of solute–solvent combinations. It expands the capabilities of computational methods toward challenging molecular systems, where efficient implicit solvent models reach their limits.

CHAPTER 5

Conformational energies of biomolecules in solution: Extending the MPCONF196 benchmark with explicit water molecules

Christoph Plett,* Stefan Grimme,* Andreas Hansen*

Received: August 23, 2023
Published online: November 19, 2023

Reprinted (adapted) in Appendix D with permission‡ from:
C. Plett, S. Grimme, and A. Hansen, J. Comput. Chem. **45** (2024) 419.

Own contributions

- Generation and compilation of the structures of the benchmark set
- Performing of all calculations with the tested methods
- Evaluation of the computational results
- Writing of the manuscript

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

Structure generation tools like QCG often rely on efficient SQM or FF methods to maintain computationally feasible. However, this can limit the accuracy of the resulting structure ensembles, necessitating refinement with more accurate methods. Suitable methods for this are commonly identified by consulting benchmark sets, but only a few exist for explicitly solvated structures. To address this gap, the solvMPCONF196 was introduced, comprising conformers of biologically relevant structures solvated by water molecules.

The benchmark is based on the 196 conformers of 13 peptides and organic macrocycles composed in the MPCONF196 set,[76] which were extended by solvation with up to nine water molecules. Thereby, the functional groups expected to interact most strongly with water were solvated using QCG. The solvent shell was further optimized with the r^{2}SCAN-3c[129] DFT method, refining the intermolecular interactions. During these optimizations, the solute was kept fixed to allow a direct comparison to the original MPCONF196 structures and a clear analysis of additional contributions and error sources introduced by solvation. Finally, highly accurate PNO-LCCSD(T)-F12[195–197] reference values, with a remaining error of less than 0.5 kcal mol^{-1} ,[198] were generated to assess various DFT, WFT, SQM, and FF methods.

Generally, it was found that explicit solvation significantly increases the complexity of computing accurate conformational energies. In addition to the energy differences between the solute conformers, solute–solvent and solvent–solvent interactions, often dominated by hydrogen bonding, had a major influence on the conformational energies. Notably, a broad range of interaction motifs was observed, partly due to the diverse functional groups of the solutes, which can act as both hydrogen-bond donors and acceptors. These interactions significantly altered the energy landscape compared to the isolated gas-phase structures of the MPCONF196. Structures that were unstable in the gas phase due to exposed polar groups could become stable when saturated with water molecules, leading to a significant influence of solvation on the energetic ordering and an increased range of conformational energies. Compared to this, exemplary relaxations of solutes solvated by water molecules had a minor influence on the energies, although significant geometry changes were observed.

Evaluation of common DFT functionals in a large quadruple-ζ basis set confirmed the usual trend of Jacob's ladder:[119] double-hybrid functionals were most accurate but also the most expensive, followed by hybrid, and (meta-)GGA methods. The best functional tested was the double hybrid revDSD-PBEP86-D4[125,199] with a mean absolute deviation (MAD) from the reference of below 0.4 kcal mol^{-1} , a maximum error around 1.5 kcal mol^{-1} , and a Spearman rank correlation coefficient of 0.995, indicating a near-perfect ranking accuracy. However, this method was computationally about five times more expensive than the tested hybrid methods, among which the range-separated hybrid functionals performed best (MADs of about 0.5 kcal mol^{-1} to 0.8 kcal mol^{-1}). Other hybrid methods like B3LYP-D4,[200,201] PW6B95-D4,[202] and PBE0-D4[203] performed slightly worse, but also MADs of up to 1.3 kcal mol^{-1} were observed for M06-2X[204]. The tested (meta-)GGA methods showed a wider range of errors with MADs ranging from about 0.9 kcal mol^{-1} (B97M-V[205,206]) to 1.5 kcal mol^{-1} (BP86-D4[207–209]). However, compared to the hybrid methods, they reduce computation time by a factor of about seven.

Investigations on the basis set size revealed that going from quadruple-ζ to triple-ζ basis sets caused only a slight loss of accuracy. In contrast, employing basis sets of double-ζ size led to significant errors attributed, among others, to the additional intermolecular BSSE. This trend was also observed for composite DFT methods, where HF-3c[128] and PBEh-3c[210], both using a small single-ζ or double-ζ basis set, showed large MADs of up to 3.4 kcal mol^{-1} . Conversely, r^{2}SCAN-3c with a modified triple-ζ basis set was much more accurate with an MAD of 1.2 kcal mol^{-1} , surpassing most

of the (meta-)GGA methods with quadruple-ζ basis sets while being up to seven times faster.

Among the tested SQM methods, GFN2-xTB[49] performed best with an MAD of 3.2 kcal mol^{-1} and a Spearman coefficient of 0.934. While these errors were larger compared to DFT methods, GFN2-xTB offered a tremendous speed-up of more than an order of magnitude. While being even less computational expensive then SQM methods, the tested FFs performed overall worse, with MMFF94[211–215] yielding the lowest MAD of 4.4 kcal mol^{-1} , followed by GFN-FF[50] with an MAD of 8.0 kcal mol^{-1} .

In summary, the solvMPCONF196 benchmark set represents a significant contribution to the rather unexplored field of benchmarking methods for describing explicit solvation. It revealed challenges arising from the additional solvent molecules when evaluating conformers and provides guidelines for method selection to refine and analyze the respective ensembles.

CHAPTER 6

ONIOM meets *xtb*: efficient, accurate, and robust multi-layer simulations across the periodic table

Christoph Plett,[*] Abylay Katbashev,[*] Sebastian Ehlert,[†] Stefan Grimme,[*] Markus Bursch[‡]

Received: May 12, 2023
Published online: June 22, 2023

Reprinted (adapted) in Appendix E with permission[§] from:
C. Plett, A. Katbashev, S. Ehlert, S. Grimme, and M. Bursch, Phys. Chem. Chem. Phys. **25** (2023) 17860.

Own contributions

- Conception of the study
- Preparation and calculation of the explicit solvation example
- Evaluation of the results
- Writing of the manuscript

[*]Mulliken Center for Theoretical Chemistry, University of Bonn, Beringstr. 4, D-53115 Bonn, Germany.
[†]Microsoft Research AI4Science, Evert van de Beekstraat 254, 1118 CZ Schiphol, The Netherlands.
[‡]Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, 45470 Mülheim an der Ruhr, Germany.

As discussed in Chapter 5, accurate structure ensembles typically require the use of computationally demanding methods. However, practical applications may involve systems too large to be modeled with conventional DFT or WFT methods. To bridge this gap, multilevel approaches like the ONIOM scheme[83] can be used to combine efficient SQM and FF methods with DFT and WFT. For ONIOM, the system is divided into different parts, whereby more accurate but computationally expensive methods are applied to the region of primary interest. The rest of the system is treated with more efficient methods, whereby the GFN methods are particularly well-suited due to their robustness and applicability across a broad range of elements. To facilitate the use of GFN methods in multilevel approaches, a subtractive ONIOM scheme was implemented in the *xtb* program suite. An interface to common QM software further allows combining the GFN[48–50] with DFT and WFT methods. Further, energies and gradients are implemented, enabling, e.g., geometry optimizations and conformational energy refinements within the ONIOM framework.

The *xtb* ONIOM implementation involves partitioning the molecular system into layers that are treated at different levels of theory and subsequently combined in a subtractive scheme. Thereby, cut bonds are handled by placing hydrogen atoms automatically at the vacant positions according to tabulated bond distances.[216] While the additional hydrogen atoms do not affect the subtractive computation of the total energy from the fragments, they introduce artificial forces affecting the gradient. To correct this, a Jacobian matrix is used to map the force to individual atoms and to identify the artificial component.

The use of the ONIOM scheme combined with the GFN methods was demonstrated in the context of explicit solvation for an exemplary system of the bis(cyclopentadienyl)silylniobium complex $[Cp_2Nb(H)_2(SiCl^iPr_2)]$[217] solvated by 20 DMSO molecules. Here, geometry optimizations with GFN methods alone resulted in severe structural errors due to a haptotropic shift of the cyclopentadienyl rings. DFT optimizations yielded reasonable structures, but required about 49 h computation time for this large example of 245 atoms. Using the ONIOM scheme for treating the solvent shell with GFN2-xTB and the solute with DFT yielded a structure nearly identical to the full DFT optimization (RMSD of 0.01 Å) in just one hour.

ONIOM in conjunction with the GFN methods offers a broad applicability beyond solvation, shown for the example of a zirconium-based UiO-66 MOF.[218] DFT optimizations of a cut-out model reproduced experimentally measured geometries accurately, but required 46 h to 191 h, depending on the chosen method. By applying ONIOM with GFN2-xTB or GFN-FF in combination with the respective DFT methods, similarly accurate structures were obtained with significantly reduced computational costs of 10-20 min. Systematic improvement in accuracy was observed when the high-level region was extended. Also, combining GFN2-xTB for the metal-organic nodes and GFN-FF for the organic linkers yielded reasonable structures in just 2 minutes.

In addition to geometry optimizations, ONIOM can be used to accelerate energy calculations, shown for a reaction mechanism discovery of the Rh-functionalized metal–organic cuboctahedron catalyzing a cyanosilylation.[186,219] For such a large system of 404 atoms in total, (meta-)GGA methods took about 30 min for a single energy evaluation. Combining the same DFT methods with GFN2-xTB via ONIOM reduced this time to 40 sec, with small energy deviations of just 1–3 kcal mol^{-1} . With this level of efficiency for such large systems, even hybrid functionals could be applied, potentially improving accuracy beyond what is achievable using (meta-)GGAs alone. Further highlighting the potential of the ONIOM implementation, the fusion of a porphyrin-based macromolecular spoked-wheel complex[220] in toluene was investigated theoretically. With its 870 atoms, this system exceeds the practical limits of DFT methods. By combining GFN2-xTB or GFN1-xTB with DFT methods via the ONIOM

scheme, it was possible to model the reaction and obtain reaction energies of about $-44.9\ \mathrm{kcal\,mol^{-1}}$. In contrast, applying the SQM methods alone resulted in values that were about $10\ \mathrm{kcal\,mol^{-1}}$ off. Even GFN-FF, which showed severely incorrect positive reaction energies when applied to the whole system, yielded qualitatively correct results ($-35.3\ \mathrm{kcal\,mol^{-1}}$) when combined with DFT through the ONIOM scheme.

In summary, the ONIOM implementation of *xtb* allows matching the strengths of the GFN and DFT methods, improving the cost–accuracy ratio beyond what the individual methods can achieve. This enables the treatment of large structures that require a highly accurate quantum mechanical description for a region of central importance but are too large for applying DFT or WFT methods to the entire system.

CHAPTER 7

Toward Reliable Conformational Energies of Amino Acids and Dipeptides—The DipCONFS Benchmark and DipCONFL Datasets

Christoph Plett,* Stefan Grimme,* Andreas Hansen*

Received: June 20, 2024
Published online: September 11, 2024

Reprinted (adapted) in Appendix F with permission‡ from:
C. Plett, S. Grimme, and A. Hansen, J. Chem. Theory Comput. **20** (2024) 8329.

Own contributions

- Compilation of the structures for both sets
- Performing of all calculations with the evaluated methods
- Evaluation of the results
- Writing of the manuscript

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

In addition to multilayer schemes such as ONIOM, machine-learning-based approaches are a promising alternative for treating large systems with reasonable accuracy. However, these models require extensive and accurate reference data for training.[91,93] To identify suitable reference methods for generating these data and to extend the existing ones, the DipCONFS and DipCONFL sets were developed. They cover conformers of the 17 natural amino acids neutral in aqueous solution and their 289 possible dipeptides. Conformers for both sets were selected from the PeptideCs database, which contains over 400 million different structures.[221] To avoid bias in the DipCONF sets, all conformers except the originally energetically lowest one were selected randomly. For accurate geometries in aqueous solution, all structures were refined using the r^{2}SCAN-3c method[129] with the implicit CPCM(water) solvent model[154]. This led to structures where the most important interactions governing conformational stability are intramolecular hydrogen bonds within the backbones of the peptides. Depending on the side chains, additional motifs such as sulfur-based hydrogen bonding and π–π stacking in aromatic systems were also found to be important.

The smaller DipCONFS benchmark set contains 918 conformers with highly precise PNO-LCCSD(T)-F12[195–197] conformational reference energies, with remaining errors of less than 1 kcal mol^{-1}. A comprehensive assessment common DFT functionals with a large quadruple-ζ basis set revealed that the revDSD-PBEP86-D4[125,199] double hybrid performed best among the tested methods with an MAD of only about 0.2 kcal mol^{-1} and a maximum error of 0.6 kcal mol^{-1}. However, the high computational cost of this method limits its application for generating large-scale datasets. Other tested double hybrid and hybrid methods also performed reasonably well, whereby no large deviations were observed between the different functionals. For the tested (meta-)GGAs, the accuracy on the DipCONFS was significantly more method-dependent. The best performer of this class, r^{2}SCAN-D4,[222] achieved a similar accuracy as the tested hybrid functionals, whereas M06-L[223] yielded outliers of up to several kcal mol^{-1}. Aligning with these results, the hybrid method ωB97M-D3(BJ)/def2-TZVPPD,[106,123,124,205] used to generate other datasets like SPICE,[224] also proved suitable for generating training data. In contrast, BP86-D3(BJ)/DGauss-DZVP[208,209,225], employed for generating the original PeptideCs database, was found to be not sufficiently accurate for generating reliable reference data. Despite the tuned D3 parameters,[226] outliers of up to 3.5 kcal mol^{-1} were observed. Among the tested methods, r^{2}SCAN-3c was found to have a favorable balance between accuracy and efficiency. This composite method achieved an accuracy similar to most of the hybrid methods in the large basis set while being about 28 times faster than the best tested hybrid functional and four times faster than r^{2}SCAN-D4 in the large basis set. Among the tested SQM and FF methods, none was found to have sufficient accuracy for generating accurate datasets, as the best performer from these classes, PM7,[134] had an MAD of 1.6 kcal mol^{-1}, which was worse than the MLIP AIMNet2[227] (MAD of 1.3 kcal mol^{-1}).

Based on the results of the DipCONFS benchmark set, r^{2}SCAN-3c was selected for generating gas-phase properties for the 29,128 conformers of the DipCONFL dataset. It showed a well-balanced conformational energy distribution of approximately Poisson shape with a peak at about 8–9 kcal mol^{-1} and conformational energies of up to 37 kcal mol^{-1}. Notably, structural differences did not consistently correlate with the conformational energies. For instance, a minor conformational change, such as the rotation of a hydroxyl group, could disrupt an intramolecular hydrogen bond leading to an energy change of 5–6 kcal mol^{-1}.[228]

Evaluation of various SQM, FF, and MLIP methods showed generally larger errors on DipCONFL compared to DipCONFS. Consequently, large-scale benchmarks such as the DipCONFL are particularly valuable for both the training and assessment of general-purpose computational methods. Notably, the

MMFF94[211–215] as best performer among the tested FFs, still showed overall large errors of up to 22.4 kcal mol^{-1} . Further, the best performer among the tested SQM methods, PM7, yielded deviations of up to 11 and was outperformed by the MLIP AIMNet2 in terms of MAD and maximum error.

In summary, the DipCONFS benchmark set with it's highly accurate conformational energies of amino acids and dipeptides was used to identify reliable computational methods for large-scale data generation. With these insights, the DipCONFL dataset was compiled, significantly expanding the available training data and thereby supporting the development of robuts MLIPs.

CHAPTER 8

Conclusion and Outlook

In summary, understanding the three-dimensional molecular structure remains an important aspect of reliable computational studies. This often includes identifying relevant conformers, stable spatial arrangements of atoms that can be interconverted via rotation around single bonds and that can have significantly different properties.[21,23] In this regard, reliable workflows must generate all possible low-energy structures and accurately evaluate their conformational energies.[35] This is especially challenging for large systems of non-covalently interacting molecules due to their complex conformational space and the additional contributions to the conformational energy.[43,46] The work presented in this thesis aimed to extend the capabilities of computational structure generation toward interacting molecules, to assess existing methods and support the development of future methods for conformer evaluation, and to facilitate the accurate simulation of extended system sizes by employing multilayer schemes.

In Chapters 3 and 4, automated and efficient tools for generating structures of interacting molecules were presented. The automated Interaction Site Screening (aISS) algorithm allows the docking of nearly arbitrary molecules. It leverages an intermolecular force field for efficient structure screening of rigid monomers combined with the GFN2-xTB, GFN1-xTB, or GFN-FF method for geometry optimization. The performance of aISS was tested on a set of standard organic molecules, transition metal complexes, and electronically challenging heavy main-group compounds where the aISS algorithm yielded structures of similar quality to those generated by *CREST*. As the docking tool is computationally much more efficient than such a general conformer generation tool, aISS makes the treatment of even systems composed of thousands of atoms possible. With its broad applicability, the aISS algorithm thereby extends the capabilities of common docking tools that are typically restricted to biologically relevant molecules. Additionally, the directed docking feature allows structure generation focused on specific interaction sites. With this, aISS can replace the manual modeling of reaction intermediates, thereby improving quality and efficiency. The application of aISS was already demonstrated beyond the contents of this thesis in diverse areas, such as molecular sensors,[229,230] biologically relevant applications,[231,232] elucidating molecular mechanisms,[233] and as part of other workflows.[234]

The second tool presented in this work, the Quantum Cluster Growth (QCG) algorithm, extends the capabilities of docking tools like aISS toward larger clusters for modeling solvation. This hybrid cluster–continuum approach grows solute–solvent clusters through sequential docking of solvent molecules by simultaneously modeling the bulk solvent with implicit solvent models. Individually

adjusted repulsive potentials ensure physically meaningful structures during the growth and an equal distribution of solvent molecules around the solute. In addition to the cluster growth, QCG allows ensemble generation and computation of solvation free energies from the generated structures. When assessing the quality of QCG-generated structures, it outperformed algorithms based on common explicit solvation models like TIP3P and algorithms that rely solely on geometric criteria. The use of QCG was demonstrated in several cases where implicit solvent models were insufficient. For example, in IR spectra simulations, implicit solvent models predicted additional or missing peaks, while the spectra calculated with QCG-generated structures showed much better agreement with exmperimental findings. Further, implicit solvent models overestimated intramolecular hydrogen bonds during an exemplary MD simulation of the drug bacillaene, while the addition of explicit solvent molecules led to significantly different structures. For test cases of typical microsolvation examples, QCG proved to be an efficient tool for generating all relevant solute–solvent structures. Beyond these examples, QCG was applied in various studies like simulations of spectroscopic properties,[235–237] structure analysis in solution,[238,239] reaction mechanism elucidation,[240,241] and modeling of ions and ionic liquids.[242,243] Furthermore, while it was initially based on the docking procedure originally proposed with the xTB-IFF, it was later modified to use the aISS algorithm, allowing e.g., the use of directed docking to solvate certain functional groups of the solute.

While QCG relies on SQM and FF methods to efficiently explore the conformational space, reliable ensembles usually require refinement with more accurate DFT or WFT methods. To guide the selection of suitable higher-level methods for such a refinement, Chapter 5 introduced the solvMPCONF196 benchmark set. It complements the very limited number of benchmark sets on explicit solvation by solvating the 196 conformers of 13 biologically relevant molecules included in the MPCONF196 benchmark with up to nine water molecules. The additional solvent molecules had a significant influence on the conformational energies and increased the complexity of their computation due to the additional solute–solvent and solvent–solvent interactions. In particular, these NCIs, mainly dominated by intermolecular hydrogen bonds, were found to have a strong impact on the conformer ordering and the structure of the solute. The provided PNO-LCCSD(T) reference data with less than 0.5 kcal mol^{-1} remaining error was used to evaluate common WFT, DFT, SQM, and FF methods, revealing that the choice of the method has a substantial impact on the ensemble quality. Generally, the tested methods followed the trend of the Jacob's ladder of DFAs proposed by Perdew[119] with double-hybrid functionals like revDSD-PBEP86-D4 being the most accurate, followed by hybrids and (meta-)GGA functionals. However, the high computational costs of double-hybrid methods make the application of less expensive hybrids or (meta-)GGA methods desirable. Within these classes, the accuracy of the conformational energies were strongly dependent on the functional choice. Moreover, it was shown that at least a triple-ζ basis set should be used for reliable results, as smaller basis set sizes led to significant errors. Notably, the r^{2}SCAN-3c composite method proved valuable for ensemble refinement due to an outstanding cost-accuracy ratio. Among the tested SQM and FF methods, the best performer, GFN2-xTB, showed reasonable results given the tremendous speed-up of orders of magnitude compared to DFT methods, but did not yield a highly accurate conformer ranking.

The results of the solvMPCONF196 benchmark set indicate a gap for, e.g., large solutes with many solvents. While the application of even the most efficient DFT methods becomes computationally less feasible for such extended system sizes, GFN2-xTB might not yield sufficiently accurate results. Chapter 6 introduced a solution to bridge this gap with the *xtb* implementation of a subtractive ONIOM scheme.[83] This allows the combination of efficient GFN methods, capable of treating a variety of systems efficiently and robustly, with the accuracy of DFT and WFT methods for energies and

geometry optimizations. The ONIOM implementation was used to achieve an accurate description of an electronically challenging transition metal complex solvated by 20 DMSO molecules. By combining DFT methods for treating the solute with the GFN methods for describing the explicit solvent molecules, accurate geometries could be generated with only a fraction of the computational time of treating the whole system with DFT. Additionally, the application of ONIOM with the GFN methods was demonstrated apart from solvation for challenging MOF systems. For example, the optimization of the zirconium-based UiO-66 MOF[218] was accelerated up to ten times with ONIOM compared to treating the whole system with DFT with only negligible loss in accuracy. Further, the ONIOM implementation was used to elucidate the mechanism of a cyanosilylation reaction at a Rh-functionalized metal–organic cuboctahedron. Thereby, multiple orders of magnitude in computation time were saved while yielding almost identical energies. This approach was also applied to elucidate the formation of a Zn-based nanoring with 870 atoms, where standard DFT methods would barely be applicable.

Besides efficient schemes like ONIOM, the rapidly evolving field of machine-learned interatomic potentials (MLIPs) shows great potential for efficient conformer evaluation. However, compared to SQM and FF methods, they require large amounts of training data, limiting their possible range of application and maximum accuracy. Contributing to this field, Chapter 7 presented a benchmark and dataset combination extending the available training data and providing insights into generating reliable datasets. The conformers for both sets covering 17 amino acids and their 289 possible dipeptides were selected almost randomly from the PeptideCs[221] database, which contains approximately 400 million structures. This approach ensured the inclusion of both low- and high-energy conformers, providing a representative sampling of the conformational space and a well-balanced distribution of conformational energies. The DipCONFS benchmark set comprises 918 conformers and highly accurate PNO-LCCSD(T) conformational energies with a remaining error of less than 1 kcal mol^{-1} . With this set, various DFT and WFT methods were evaluated with regard to their suitability as a reference method for MLIP training data generation. It was found that upon using a large basis set, all tested double-hybrid and hybrid density functionals, as well as most (meta-)GGA functionals, were in principle suitable reference methods for computing conformational energies. However, when generating large amounts of data, computational efficiency becomes crucial. In this regard, r^2SCAN-3c performed exceptionally well as it was much more efficient than any other tested method applied with a large basis, while maintaining high accuracy (MAD of 0.3 kcal mol^{-1}). This rendered r^2SCAN-3c an ideal method for extending the DipCONFS benchmark to the DipCONFL dataset. With 29,128 conformers and accurate DFT data including conformational energies, gradients, and partial charges, the DipCONFL set provides valuable data for MLIP training. The evaluation of various SQM and FF methods on both sets showed larger errors for DipCONFL than DipCONFS. Notably, even the most accurate SQM and FF methods tested were outperformed by the MLIP AIMNet2, highlighting the great potential of MLIPs.

In summary, the scientific advances presented in this thesis comprise the development of generally applicable and automated tools for discovering spatial arrangements of interacting molecules, alongside novel insights into how such structures can be refined to achieve high accuracy. Further, practical solutions to combining different methods for structure refinement were introduced that expand the accessible cost-accuracy ratio of one method alone, and valuable data for developing future methods for conformer evaluation were provided. Together, these contributions offer powerful strategies for robust and routine structure exploration, enabling accurate conformational modeling and hence advancing the capabilities of computational chemistry.

In future studies, the application of the aISS and QCG tools for generating larger molecular clusters beyond solvation could be further explored. This is of great interest, e.g., for modeling solids[244] and aggregation processes in the atmosphere.[245] Although the stepwise use of the aISS algorithm is, in principle, suitable for generating large clusters, applications like QCG and workflows for monolayer generation[234] have shown that constraining the cluster growth might be necessary to achieve physically meaningful models. Besides employing repulsive potentials similar to QCG, the introduction of symmetry-based or volumetric criteria might be beneficial. Additionally, using conformer ensembles instead of single structures as input for the aISS algorithm would cover cases where the energetically favored monomer conformation does not relate to the best configuration in the dimer. However, the computational costs involved in docking all combinations of two conformer ensembles can become prohibitive, even with efficient tools like aISS. Thus, smart selection of reliable candidates for docking should be implemented, e.g., based on energetic criteria, scans for accessible interaction sites, or symmetrical constraints. Embedding the resulting clusters in an appropriate electrostatic continuum might be sufficient for simulating properties like the stability of solids, one of the essential steps for predicting their solubilities.[246]

Regarding the physically motivated approach of computing solvation free energies introduced by the QCG algorithm, various future developments of more robust structure generation methods could directly improve the values. Additionally, emerging methods like g-xTB[247] for energy evaluation with better cost-accuracy ratios would enhance solvation free energy predictions. With these improved methods, also the automated identification of cases where implicit solvent models lack accuracy might become possible, e.g., by evaluating association free energies of the solute and solvent molecules.[248]

In terms of identifying and assessing beneficial methods for treating the respective solute–solvent clusters, solvMPCONF196 provided already a solid foundation for organic molecules solvated by water. However, real applications often involve more diverse solutes and also solvents beyond water ranging from standard organic ones up to increasingly important ionic liquids.[249] Therefore, additional benchmark sets have to be developed that expand the covered solute and solvent combinations, allowing for robust refinements of solute–solvent clusters of such systems.

In closing, this thesis represents a substantial contribution to the computational modeling of molecular structures, particularly with regard to non-covalent interactions. The insights gained and the methods introduced have already facilitated various research studies. Moreover, they provide a robust foundation for future advancements, expanding the possibilities of reliable, efficient, and routinely applicable approaches for structure exploration further.

APPENDIX A

Supporting Information to Chapter 2: Theoretical Background

A.1 DFT-D Dispersion Corrections

The D3 model with Becke-Johnson damping and the D4 model have the general form of the two-body dispersion energy:[123–125]

$$E_{\text{disp}}^{(2)} = -\sum_{AB}\left[s_6\frac{C_6^{AB}}{R_{AB}^6 + f(R_{AB}^0)^6} + s_8\frac{C_8^{AB}}{R_{AB}^8 + f(R_{AB}^0)^8}\right], \tag{A.1}$$

with s_6 scaling the dipole-dipole and s_8 scaling the dipole-quadrupole contribution. For (meta-)GGAs and hybrids, s_6 is usually chosen to be 1.0. C_6^{AB} and C_8^{AB} denote the dispersion coefficients, and R_{AB} the internuclear distance of the atom pairs. $f(R_{AB}^0)$ defines the Becke-Johnson damping as

$$f(R_{AB}^0) = a_1^{\text{BJ}} R_{AB}^0 + a_2^{\text{BJ}}. \tag{A.2}$$

The parameters a_1^{BJ} and a_2^{BJ} are, like s_6 and s_8, individually determined for the different functionals. R_{AB}^0 is the cut-off radius, defined as

$$R_{AB}^0 = \sqrt{\frac{C_8^{AB}}{C_6^{AB}}}. \tag{A.3}$$

The damping reduces the dispersion energy for smaller interatomic distances to avoid near-singularities and the double-counting of electronic effects already included in the DFA. The C_6^{AB} dispersion coefficients are approximately determined in each calculation by weighting precomputed values for the element hydrides obtained from the dynamic polarizabilities via the Casimir-Polder relation.[250] The C_8^{AB} coefficients result from the C_6^{AB} according to

$$C_8^{AB} = 3C_6^{AB}\sqrt{Q_A Q_B}, \tag{A.4}$$

with

$$Q_{A/B} = s_{42}\sqrt{Z_A}\frac{\langle r^4\rangle^{A/B}}{\langle r^2\rangle^{A/B}}. \tag{A.5}$$

$\langle r^4 \rangle^{A/B}$ and $\langle r^2 \rangle^{A/B}$ are multipole-type expectation values derived from atomic densities and s_{42} is a factor chosen so that reasonable C_8^{AA} values for He, Ne, and Ar result. In addition to the two-body dispersion energy $E_{\text{disp}}^{(2)}$, the energy due to three-body effects $E_{\text{disp}}^{(3)}$ can also be added for generally improved results. It is computed via the Axilrod–Teller–Muto (ATM) term[251,252]

$$E_{\text{disp}}^{(3)} = \sum_{ABC} \frac{C_9^{ABC}(3cos(\theta_{AB})cos(\theta_{AC})cos(\theta_{BC}) + 1)}{(r_{AB}r_{AC}r_{BC})^2}, \tag{A.6}$$

with θ_{AB}, θ_{AC}, and θ_{BC} denote internal angles of a triangle formed by the atoms A, B, and C that are separated by the distances r_{AB}, r_{AC}, and r_{BC}. The triple-dipole dispersion coefficient C_9^{ABC} can be approximated by

$$C_9^{ABC} \approx -\sqrt{C_6^{AB}C_6^{AC}C_6^{BC}}. \tag{A.7}$$

With the three-body term, the dispersion energy is approximated as

$$E_{\text{disp}} \approx E_{\text{disp}}^{(2)} + E_{\text{disp}}^{(3)}. \tag{A.8}$$

Depending on the size and geometry of the system, higher-order many-body effects can become similarly important as three-body terms.[253] Within the D4 model, they can be approximated via coupling the atomic dipole polarizabilities based on quantum harmonic oscillators (QHOs).[254,255]

A.2 xTB-IFF

The following provides more details on how the different contributions of the xTB-IFF intermolecular interaction energy are computed (Equation 2.55.[47]

The Pauli repulsion energy E_{rep} is calculated as a sum over atom pairs A and B of two different molecules

$$E_{\text{rep}} = \sum_{AB} N_A N_B \left(k_{r1} \frac{S_{AB}}{R_{AB}} + k_{r2} e^{-(k_p R_{AB} \alpha_{AB})} \right). \tag{A.9}$$

k_{r1}, k_{r2}, and k_p are fit parameters, N_A and N_B the valence electron number, R_{AB} is the distance between the atoms, S_{AB} the atomic overlap, and α_{AB} an effective pair distance. The valence electron number is computed with the valence nuclear charge $Z_{A/B}^{\text{val}}$ and Mulliken atomic charges $q_{A/B}$, available from the GFNn-xTB electronic properties of the monomers:

$$N_{A/B} = Z_{A/B}^{\text{val}} + q_{A/B}. \tag{A.10}$$

The effective pair distance α_{AB} is computed with the dipole polarizabilities of atom A (α_A) and B (α_B), taken from the D4 model,[254] and a parameter k_α:

$$\alpha_{AB} = k_\alpha \frac{1}{2} \left(\alpha_A^{1/3} + \alpha_B^{1/3} \right) \tag{A.11}$$

The atomic overlap S_{AB} is computed via atom-centered 1s Slater functions ϕ_{1s}:

$$S_{AB} = \int \phi_{1s}^A(r) \phi_{1s}^B(r) dr, \tag{A.12}$$

whose shapes are defined by an exponent ζ_{1s} that is individually computed for each atom A/B:

$$\zeta_{1s}^{A/B} = k_{A/B}^{R0} \alpha_{A/B}^{-1/3} (1 + CN_{A/B})^{k_{CN}}. \tag{A.13}$$

$k_{A/B}^{R0}$ denotes an element-specific fit parameter and the coordination number $CN_{A/B}$ is taken from the D3 model.[123] k_{CN} is set to be $-\frac{1}{2}$ for hydrogen atoms yielding contracted densities and $-\frac{1}{3}$ for other elements.

The electrostatic energy E_{ES} is divided into two contributions, the atom-centered contribution $E_{\mathrm{ES}}^{\mathrm{atom}}$ and a correction for anisotropic electron densities around the atoms $E_{\mathrm{ES}}^{\mathrm{aniso}}$:

$$E_{\mathrm{ES}} = E_{\mathrm{ES}}^{\mathrm{atom}} + E_{\mathrm{ES}}^{\mathrm{aniso}} \tag{A.14}$$

The atom-centered contribution is computed by applying the Coulomb law to the model atomic densities resulting from $\phi_{1s}^{A}(r)$ and $\phi_{1s}^{B}(r)$:

$$E_{\mathrm{ES}}^{\mathrm{atom}} = \sum_{AB} \frac{Z_A Z_B}{R_{AB}} + \int\int \frac{\phi_{1s}^{A}(r_i)\phi_{1s}^{B}(r_j)}{|r_i - r_j|} dr_i dr_j - \int \frac{Z_A \phi_{1s}^{B}(r_j)}{|R_A - r_j|} dr_j - \int \frac{\phi_{1s}^{A}(r_i) Z_B}{|r_i - R_B|} dr_i. \tag{A.15}$$

The correction for anisotropic electron densities $E_{\mathrm{ES}}^{\mathrm{aniso}}$ is computed with a damped Coulomb expression summing over pairs of off-centered (Q^{off}) and atomic charges q:

$$E_{\mathrm{ES}}^{\mathrm{aniso}} = \sum_{i,j} \frac{Q_i^{\mathrm{off}} Q_j^{\mathrm{off}}}{|r_i - r_j| + \frac{\mathrm{k_{dmp1}}\alpha_{AB}}{|r_i - r_j|}} + \sum_{i,B} \frac{Q_i^{\mathrm{off}} q_B}{|r_i - R_B| + \frac{\mathrm{k_{dmp1}}\alpha_{AB}}{|r_i - R_B|}} + \sum_{j,A} \frac{Q_j^{\mathrm{off}} q_A}{|r_j - R_A| + \frac{\mathrm{k_{dmp1}}\alpha_{AB}}{|r_j - R_A|}}. \tag{A.16}$$

$\mathrm{k_{dmp1}}$ is a damping constant, i and j refer to the off-centered charges at position r_i and r_j, and A and B to the atoms at position R_A and R_B. The off-centered charges Q^{off} at position r_i are constructed based on Foster-Boys LMOs[256] stemming from the GFNn-xTB electronic properties of the fragments. For retaining charge conservation and to complete the dipols, also positive counter parts are considered that are placed according to the type of orbital used for constructing the off-centered charges: For π orbitals and lone pairs, they are placed at the position of the corresponding nucleus, for σ orbitals, they are placed on the opposite side of the nucleus.

The dispersion contribution E_{disp} is computed similar to the D4 model (Equation A.1), only changing the fit parameters s_8, a_1^{BJ}, and a_2^{BJ}. The ATM term (Equation A.6) can additionally be included.

The induction energy E_{ind} is computed based on the spherical Drude model.[257] Within this approximation, floating charges q^f are equilibrated that counter-balance the respective atomic charge. They are attached to the nuclei with spring constants $\mathrm{k_{D1}}$, e.g. for a atom A:

$$q_A^{\mathrm{D}} = -\mathrm{k_{D2}} \sqrt{\mathrm{k_{D1}} \eta_A \alpha_A}. \tag{A.17}$$

$\mathrm{k_{D2}}$ denotes the Drude charge proportionally constant and η_A the chemical hardness of atom A. The

induction energy within the Drude oscillator model is computed according to:

$$E_{\text{ind}} = \sum_{i,B} \frac{q_i^{\text{D}} q_B}{|r_i - R_B| + \text{k}_{\text{dmp2}}} + \sum_{j,A} \frac{q_j^{\text{D}} q_A}{|r_j - R_A| + \text{k}_{\text{dmp2}}} + \sum_{i,j} \frac{q_i^{\text{D}} q_j^{\text{D}}}{|r_i - r_j| + \text{k}_{\text{dmp2}}} + \frac{1}{2}\text{k}_{\text{D1}} \left(\sum_{i \in A} |r_i - R_A|^2 + \sum_{j \in B} |r_j - R_B|^2 \right) - E_0^{\text{D}}, \quad \text{(A.18)}$$

with i and j referring to the Drude charges where i is connected to atom A of fragment 1 and j to atom B of fragment 2. E_0^{D} refers to the energy when the negative Drude charge is located at the respective atomic position and k_{dmp2} is a damping constant.

The charge-transfer energy E_{CT} is computed according to

$$E_{\text{CT}} = \langle \Psi_0^{AB} | \hat{H}_{A+B} | \Psi_0^{AB} \rangle - \langle \Psi_0^{A+B} | \hat{H}_{A+B} | \Psi_0^{A+B} \rangle, \quad \text{(A.19)}$$

where Ψ_0^{AB} denotes the wavefunction of the combined systems after charge transfer and Ψ_0^{A+B} the combined wavefunctions of the non-interacting fragments. To both, the tight-binding Hamiltonian of the independent fragments $\hat{H}_{A+B}$ is applied and the resulting energy difference equals the energy change due to charge transfer. The combined wavefunction considering charge transfer is set up as

$$\Psi_0^{AB} = c_0 \Psi_0^{A+B} + c_{i_A b_B} \Psi_{i_A b_B}^{A+B} + c_{j_B a_A} \Psi_{j_B a_A}^{A+B}, \quad \text{(A.20)}$$

where $\Psi_{i_A b_B}^{A+B}$ defines the combined wavefunction of fragment A and B where the HOMO of A is replaced by the LUMO of B (thus allowing the charge transfer of A to B) and $\Psi_{j_B a_A}^{A+B}$ has the HOMO of B replaced by the LUMO of A. The coefficients of how much these states contribute to the coupled wavefunction (c_0, $c_{i_A b_B}$, and $c_{j_B a_A}$) are determined by solving an eigenvalue problem:

$$\begin{pmatrix} 0 & f_{ij}^{\text{CT}} & f_{ji}^{\text{CT}} \\ f_{ij}^{\text{CT}} & \Delta E_{ij}^{\text{CT}} & 0 \\ f_{ji}^{\text{CT}} & 0 & \Delta E_{ji}^{\text{CT}} \end{pmatrix} \begin{pmatrix} c_0 \\ c_{i_A b_B} \\ c_{j_B a_a} \end{pmatrix} = 0. \quad \text{(A.21)}$$

Here, $\Delta E_{ij}^{\text{CT}}$ and $\Delta E_{ji}^{\text{CT}}$ denote the respective HOMO-LUMO gap, corrected by an empirically determined factor, as GFNn-xTB methods usually underestimate these gaps.[258] f_{ij}^{CT} and f_{ji}^{CT} determine how strongly the charge-transfer states couple to the ground state. They are approximated semiempirically by using the HOMO (ϵ^{HOMO}) and LUMO (ϵ^{LUMO}) orbital energies, the atomic densities of the isolated fragments (p^{HOMO} and p^{LUMO}) from a Mulliken population analysis, and the additional fit parameters k_{CT}, β_{CT1}, and β_{CT2}:

$$f_{ji}^{\text{CT}} = -k_{\text{CT}} e^{-\beta_{\text{CT1}}(\epsilon_j^{\text{LUMO}} - \epsilon_i^{\text{HOMO}})} e^{-\beta_{\text{CT2}}\sqrt{r_{AB}/\alpha_{AB}}} \sum_{i \in A} \sum_{j \in B} p_A^{\text{HOMO}} \times p_B^{\text{LUMO}}. \quad \text{(A.22)}$$

APPENDIX B

Automated and Efficient Generation of General Molecular Aggregate Structures

Christoph Plett,* Stefan Grimme*

Reprinted (adapted) with permission‡ from:
C. Plett and S. Grimme, Angew. Chem. Int. Ed. **62** (2023) e202214477.

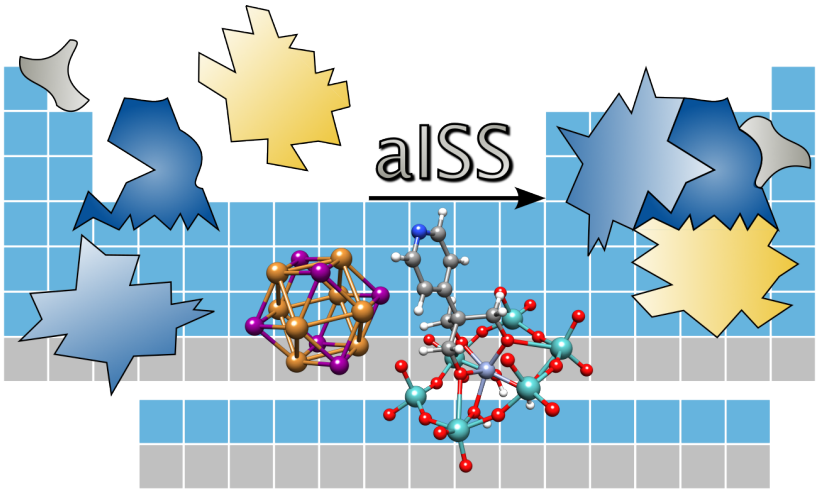

Figure B.1: Associated Table of Contents graphic for publication in Angewandte Chemie - International Edition.

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

Abstract

Modeling intermolecular interactions of complex non-covalent structures is important in many areas of chemistry. To facilitate the generation of reasonable dimer, oligomer, and general aggregate geometries, we introduce an automated computational interaction site screening (aISS) workflow. This easy-to-use tool combines a genetic algorithm employing the intermolecular force-field xTB-IFF for initial search steps with the general force-field GFN-FF and the semi-empirical GFN2-xTB method for geometry optimizations. Compared with the alternative *CREST* program, aISS yields similar results but with computer time savings of 1-3 orders of magnitude. This allows for the treatment of systems with thousands of atoms composed of elements up to radon, e.g., metal-organic complexes, or even polyhedra and zeolite cut-outs which were not accessible before. Moreover, aISS can identify reactive sites and provides options like site-directed (user-guided) screening.

B.1 Introduction

The question of how molecules interact non-covalently and arrange geometrically in an optimal way is of fundamental relevance for many chemical systems and plays an important role in various physical, biological, and chemical processes.[42] It is crucial for the association and function of supermolecular dimers and oligomers[259], self-assembly of molecules,[260] or protein-protein and protein-ligand complexes[39,261] to name only a few examples. Nowadays, computational studies often complement experimental work to elucidate structure and interaction (free) energies in great detail[262] and finally even predict complex reaction mechanisms.[263] Usually, well-established wave-function theory (WFT) or Kohn-Sham (KS) density functional theory (DFT) are used as computational methods for simulating basic molecular properties, but they quickly reach their limits for the structure generation of large oligomer and aggregate systems containing a few hundred atoms due to high computational costs. This encouraged the development of various semi-empirical and force field methods that extend the treatable system size drastically.[264,265] Recent examples are the semi-empirical GFN*n*-xTB[84] molecular orbital, tight-binding methods, as well as the generally applicable force field GFN-FF[50], which proved to provide reasonable Geometries, Frequencies, and Non-covalent interactions (GFN).[266–268] If only non-covalent interactions (NCIs)[269,270] are of interest, specialized approaches can be used that describe exclusively intermolecular interactions[271] and otherwise treat the molecules as rigid fragments. One of these methods is the intermolecular force field xTB-IFF,[47] which uses specifically pre-computed electronic properties from xTB for the two interacting molecular fragments to improve the theoretical description without introducing too much computational overhead. Even though all of these methods can be used to optimize and analyze intermolecular structures, they cannot generate them efficiently, for which special algorithms are required. Typically, existing software solutions for docking molecules focus on protein-ligand binding,[272] which is either based on simple scoring functions for binding affinities or on a very simplified calculation of the interaction energies between the fragments.[273,274] Commonly used algorithms can either do rigid docking(, e.g., *ZDOCK*,[275] and *RDOCK*)[276], flexible-rigid docking(, e.g., *Flex X*,[277,278] *AutoDock*,[279] and *Autodock Vina*)[280], or fully flexible docking like *Gold*,[281] *Glide*,[282] and *LeDOCK*.[283] These highly specialized algorithms have a limited range of applications (typically restricted to biomolecules) and cannot be applied to general chemistry. Until now, a universal and easy-to-use method that is capable of finding interaction sites and generating reasonable dimer, oligomer, and aggregate structures for molecules with sizes

up to several thousand atoms and arbitrary elemental composition has been missing. To close this gap, we here introduce a robust and efficient algorithm for this purpose, named automated interaction site screening (aISS). It enables the investigation of intermolecular geometries for a wide variety of systems like transition-metal catalysts, zeolites, or MOFs that were inaccessible before due to their size or chemical composition. The freely available aISS algorithm is designed to be easily applicable also by computational non-experts and provides useful additional features like site-directed (user-guided) screening. After a short description of the algorithm, aISS is tested for chemically interesting example systems, and the interaction energies of the resulting structures are re-evaluated with high-level DFT methods.

B.1.1 aISS Method

The principal idea of aISS is to find the energetically lowest structure (global energy minimum) of the largest interaction between two given fragments, which often has a dominant impact in the real system. A few energetically higher structures (per default 15) are also obtained by the algorithm and optionally, a more complete ensemble of thermally populated structures can be generated. As the possible bonding motifs and geometrical structures can be rather diverse and complex, multiple steps applying different approximations are performed during an aISS run (Figure B.2). The automated algorithm is invoked with one simple command-line keyword. Only two sets of input coordinates for the molecular fragments to be combined have to be provided, but information about charges, unpaired electrons, and different geometrical constraints can be set as well. The procedure then starts with GFN*n*-xTB computations of the electronic properties of fragments A and B, which are required for the xTB-IFF interaction energy calculations. This has to be done only once as the information can be used for any generated intermolecular geometry of fragments A and B. Next, a grid-based xTB-IFF energy pre-screening is conducted around fragment A with a neutral and artificially +/−0.1 charged rare-gas atom (Kr), which allows a fast exploration of possible interaction sites. The following three steps are independently used for generating different intermolecular geometries from the rigid fragments. Thereby, fragment B is moved around fragment A, and the structures are evaluated in terms of the xTB-IFF energy. Fragment A will always be the first molecular geometry file given in the program call. It is recommended to provide the larger fragment first, as the screening process will generally be more efficient if the smaller fragment is moved around the larger one. Two of the structure-generating steps aim for typical bonding motifs: a search for pockets in fragment A and a screening for π–π-stacking interactions along different directions in three dimensions (3D). The third structure-generating step is a search for general orientations of fragment B on an angular grid around fragment A including the best positions of the Kr atom pre-screening. The resulting structures of the pocket search are fully optimized including all intramolecular degrees of freedom with GFN*n*-xTB or GFN-FF, while the energetically lowest structures (per default 100) of the stack and angular search are combined and refined by a two-step genetic algorithm.[284] This ensures the inclusion of positions and orientations that are not yet covered by the grid-based searches. During this genetic optimization, first, a random

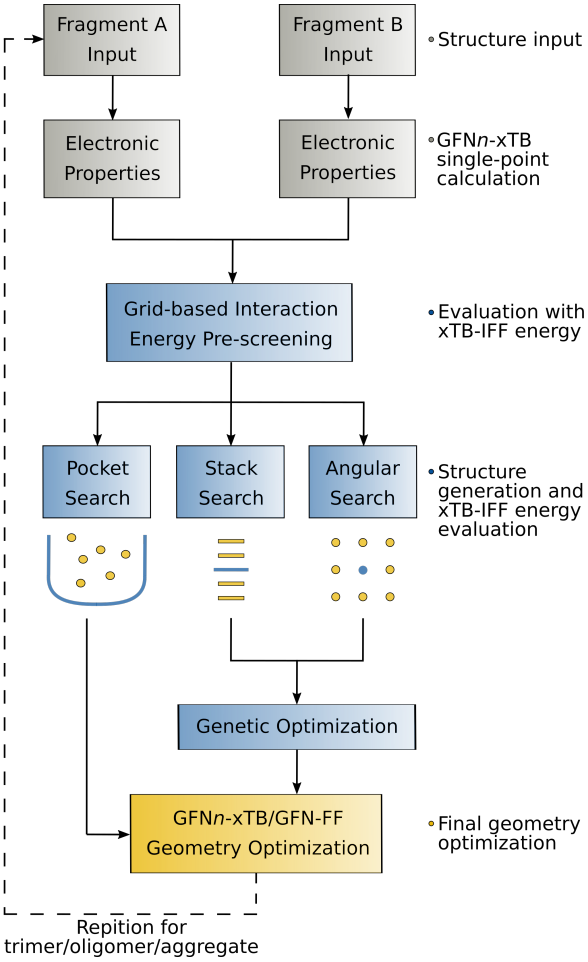

Figure B.2: Schematic depiction of the aISS algorithm.

crossover of each pair of positions of fragment B around fragment A is done according to

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\begin{pmatrix} x_{ij} \\ y_{ij} \\ z_{ij} \\ \alpha_{ij} \\ \beta_{ij} \\ \gamma_{ij} \end{pmatrix} = \sum_{i=1}^{N}\sum_{j=1}^{N}\left[\begin{pmatrix} x_i \cdot f_1 \\ y_i \cdot f_2 \\ z_i \cdot f_3 \\ \alpha_i \cdot f_4 \\ \beta_i \cdot f_5 \\ \gamma_i \cdot f_6 \end{pmatrix} + \begin{pmatrix} x_j \cdot (1-f_1) \\ y_j \cdot (1-f_2) \\ z_j \cdot (1-f_3) \\ \alpha_j \cdot (1-f_4) \\ \beta_j \cdot (1-f_5) \\ \gamma_j \cdot (1-f_6) \end{pmatrix}\right], \tag{B.1}$$

where x, y, and z are Cartesian coordinates that describe the center of mass (CMA) of fragment B with respect to the CMA of fragment A, α, β, and γ are the three Euler rotational angles of fragment B, and $f_1 - f_6$ are random numbers between zero and one. The second step of the genetic optimization is a random mutation of 50 % of the structures in position and angle. The xTB-IFF energy of each newly generated structure is used for ranking, and the genetic step is repeated per default ten times (for further details, see the Supporting Information). Finally, either a few energetically lowest geometries

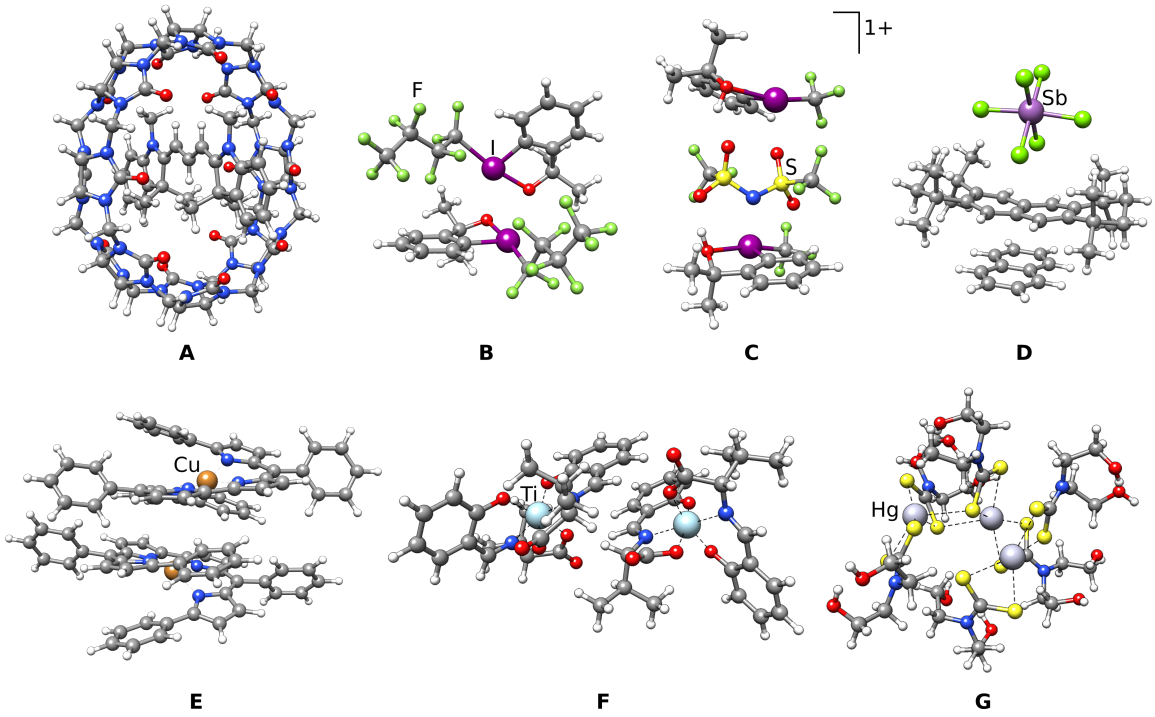

Figure B.3: The lowest energy structures for an example set of complexes resulting from the aISS algorithm.

(per default 15) or all structures with attractive xTB-IFF interaction energy are optimized, depending on whether just a single structure or an ensemble is requested. For these geometry optimizations, either the default GFN2-xTB (aISS//GFN2-xTB), or alternatively GFN1-xTB (aISS//GFN1-xTB) or GFN-FF (aISS//GFN-FF) can be chosen. Optionally, they can be conducted including a standard continuum solvation model.[181] This ensures intramolecular relaxation of the rigidly added fragments and refinement of the entire complex. If not only a dimer, but a trimer or oligomer is desired, the dimer geometry can be used as an input for a subsequent run to either add a third molecule or to combine two dimers. This can be repeated iteratively with different fragments until a desired size and composition of the aggregate is reached. Because the GFN2-xTB electronic property calculations are conducted for fragments of increasing size, polarization as well as other many-body effects are accounted for already at this intermolecular force-field level, thereby increasing accuracy and physical reliability. To ensure the free availability and easy application of the aISS algorithm, we implemented it in the open-source *xtb*[84,180] program and provided a detailed documentation.[285]

B.2 Results and Discussion

B.2.1 General Applicability and Evaluation

To demonstrate the general applicability and efficiency of the algorithm and to validate the resulting structures, we investigated a variety of currently researched dimers and trimers (Figure B.3).[182–185,286,287] These small to medium-sized systems contain up to 300 atoms including heavy main-group elements and transition metals, which are usually challenging or not even included for

Table B.1: GFN2-xTB and r^{2}SCAN-3c (DFT) interaction energies in kcal mol^{-1} of the molecules shown in Figure B.3. The structures were generated with the aISS algorithm in ensemble (aISS$^{\mathrm{E}}$) and single-structure (aISS$^{\mathrm{S}}$) run-type. For comparison, interaction energies for structures generated with the NCI-iMTD workflow of the *CREST* program are shown together with computational timings on 14 cores of an Intel(R) Xeon(R) CPU E5-2660 v4 @ 2.00GHz.

	A	**B**	**C**	**D**	**E**	**F**	**G**
$E_{\mathrm{int}}^{\mathrm{GFN2-xTB}}$ (*CREST*)	−62.0	−29.3	−153.0	−84.0	−26.1	−39.1	−95.0
$E_{\mathrm{int}}^{\mathrm{GFN2-xTB}}$ (aISS$^{\mathrm{S}}$)	−62.0	−26.6	−153.7	−83.4	−22.9	−29.2	−101.5
	(100 %)	(91 %)	(101 %)	(99 %)	(88 %)	(75 %)	(107 %)
$E_{\mathrm{int}}^{\mathrm{GFN2-xTB}}$ (aISS$^{\mathrm{E}}$)	−62.0	−28.3	−155.4	−83.5	−25.5	−36.0	−109.5
	(100 %)	(97 %)	(102 %)	(99 %)	(98 %)	(92 %)	(115 %)
$E_{\mathrm{int}}^{\mathrm{r^2SCAN-3c}}$ (*CREST*)	−61.1	−12.1	−114.2	−119.2	−23.3	−23.9	−50.9
$E_{\mathrm{int}}^{\mathrm{r^2SCAN-3c}}$ (aISS$^{\mathrm{S}}$)	−60.6	−11.7	−115.3	−119.2	−21.5	−23.7	−57.2
	(99 %)	(97 %)	(101 %)	(100 %)	(92 %)	(99 %)	(112 %)
$E_{\mathrm{int}}^{\mathrm{r^2SCAN-3c}}$ (aISS$^{\mathrm{E}}$)	−61.1	−12.1	−119.8	−121.9	−22.4	−24.1	−54.4
	(100 %)	(100 %)	(105 %)	(102 %)	(96 %)	(101 %)	(106 %)
Comp. time (*CREST*)	64 h 23 min	5 h 3 min	2 h 9 min	12 h 30 min	45 h 2 min	17 h 15 min	32 h 4 min
Comp. time (aISS$^{\mathrm{S}}$)	15 min	3 min	8 min	4 min	7 min	5 min	11 min
Comp. time (aISS$^{\mathrm{E}}$)	13 h 57 min	2 h 35 min	4 h 32 min	5 h 19 min	9 h 36 min	9 h 22 min	27 h 52 min

common force fields that do not employ electronic information. Thus, the default GFN2-xTB[49] was chosen for the final geometry optimizations in the aISS algorithm (aISS//GFN2-xTB). For evaluating the performance, a comparison with the more elaborated but well-tested and established NCI-iMTD workflow of the Conformer-Rotamer Ensemble Sampling Tool (*CREST*)[44,288] is given, which is based on metadynamics[177] and molecular dynamics simulations to find different conformers. GFN2-xTB was also employed for this NCI-iMTD workflow of *CREST*. For a quantitative comparison, the interaction energies (E_{int}) for the energetically lowest GFN2-xTB structures are compared. Because both algorithms search for the global minimum energy structure, the most attractive (negative) interaction energy corresponds to the best result. For GFN2-xTB (and alternatively for GFN-FF or DFT) this value was calculated as the difference in total energy of the fully relaxed monomers ($E(A)$ and $E(B)$) and dimer ($E(AB)$), i.e., $E_{\mathrm{int}} = E(AB) - E(A) - E(B)$. Additionally, r^{2}SCAN-3c[129] interaction energies were calculated after geometry re-optimization of the GFN2-xTB structures to validate the results with a very accurate approach for NCIs and conformational energies. Therefore, the Commandline Energetic Sorting of Conformer Rotamer Ensembles (*CENSO*)[35] program was employed to filter and to re-optimize the ensembles resulting from either the NCI-iMTD workflow of *CREST* or the ensemble run-type of the aISS algorithm, respectively. For the single-structure aISS run-type, only the best GFN2-xTB structure was re-optimized. The resulting interaction energies are shown in Table B.1 as absolute values and as percentages of E_{int} relative to the NCI-iMTD results for comparison together with the timings on a usual desktop computer. For the cationic trimethine cyanine within the cucurbit[8]uril (Figure B.3, A), [185] the interaction energies of the aISS structures are almost identical to those from the NCI-iMTD algorithm. This indicates a similar structure quality and holds also for the electronically difficult cationic, open-shell octamethylated-naphthalene complex between naphthalene and the anionic antimony hexafluoride (Figure B.3, D)[286] and the paramagnetic copper complex (Figure B.3, E).[287] Noteworthy is the tremendous difference in computational time.

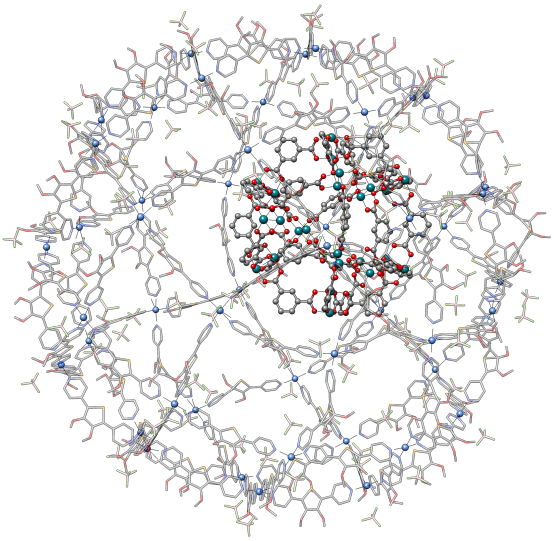

Figure B.4: The best structure of the rhodium-organic cuboctahedra inside the $Pd_{48}L_{96}(BF_4)_{96}$ Goldberg polyhedron found with the aISS//GFN-FF algorithm. Hydrogen atoms are omitted for clarity. Pd is depicted in light blue, Se in orange, B in pinkish, and Rh in light sea green.

While the *CREST* runs take days to weeks (up to 64 h 23 min), the aISS runs complete within a few minutes. The examples of the perfluoroorganyl iodine dimer (Figure B.3, B),[183] and the titanium complex in Figure B.3, F,[182] show a somewhat less attractive GFN2-xTB interaction energy for the aISS structures, but upon employing r^{2}SCAN-3c for E_{int} they become almost identical again. This shows that the aISS single-structure run-type is able to find reasonable structures, even for cases where the GFN2-xTB potential energy surface deviates from the usually more realistic DFT energy surface. The example of two perfluoroorganyl iodine cations and one bistriflimide anion (Figure B.3, C),[183] reveals another advantage of the aISS over MD-based approaches. During the *CREST* run, the electronically complex monomers had to be constrained to avoid covalent bond breaking in the molecules during the biased (high-energy) metadynamics simulations, while for the aISS algorithm no further constraints had to be set. The CREST run led therefore to a less relaxed structure compared to the aISS result as seen in a slightly more attractive E_{int} for the latter. The imposed restriction is also notable in the CREST simulation time which is only in this case smaller compared to the aISS ensemble run-type. If more complex systems with multiple molecules like the mercury-complex trimer (Figure B.3, G)[184] are modeled, the conformational space is often too large to find the global minimum energy geometry within reasonable MD simulation times.[289] In such cases, the more systematic aISS algorithm yields better results in only a tiny fraction of computational time. In comparison, the r^{2}SCAN-3c geometry optimizations took between 1 h 18 min (Figure B.3, D) and 8 h 31 min (Figure B.3, E). Hence, if DFT-optimized geometries are required, this can become the computational bottleneck, especially for the CREST and aISS ensemble approaches, where a large number of structures are optimized.

The high efficiency of aISS, especially in combination with GFN-FF geometry optimizations (aISS//GFN-FF) instead of GFN2-xTB, together with its universal applicability, enables the treatment of systems that were not possible before. As an example, a rhodium-organic cuboctahedron[186] was added into the largest self-assembled Goldberg polyhedron $Pd_{48}L_{96}(BF_4)_{96}$.[187] The resulting

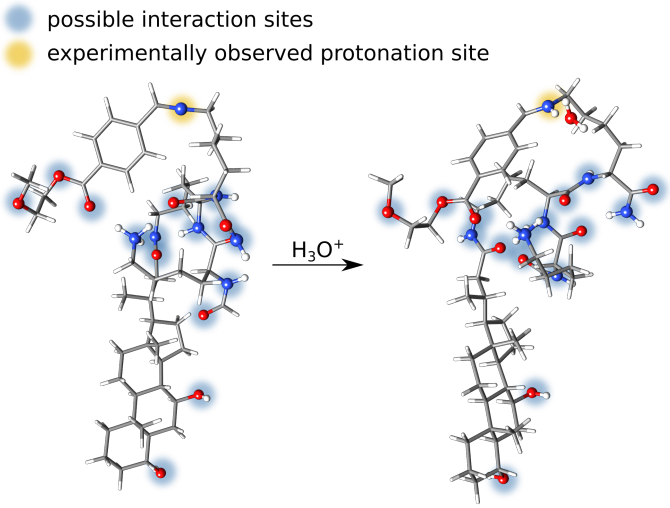

Figure B.5: Addition of oxonium to the micelle. Possible interaction sites are marked in blue. The experimentally observed protomer found correctly by aISS is marked in yellow.

structure (Figure B.4) contains 4296 atoms in total, which makes other computational approaches like *CREST* unfeasible due to their much higher cost. Moreover, the system is composed of borate anions, selenium, palladium, and rhodium, which excludes the application of other current docking software focusing on biomolecules. The total computation time on 14 cores of a usual desktop computer amounts to just 29 h, where the searching of interaction sites and the generation of the structures took 1 h while the final GFN-FF geometry optimizations took about 24 h.

To further highlight the general applicability, we treated the single-chain proteins Barnase and Barstar[290] with the aISS//GFN-FF algorithm employing the ALPB water solvation model.[181] Different from common docking software for biomolecules, no manual atomic charge assignment or other preparation had to be done as the required properties are automatically generated by the GFN2-xTB single-point calculations of the fragments. The energetically lowest, GFN-FF-optimized Barnase-Barstar dimer resulting from the aISS algorithm (Figure S4 in the Supporting Information) shows a reasonable interaction geometry, similar to the one observed in the X-ray structure.[291] Its GFN2-xTB interaction energy in water, calculated as the difference between the electronic energies including solvation of the dimer and the isolated monomers in the dimer geometry, amounts to a realistic value of $-143.1\ \mathrm{kcal\,mol^{-1}}$.

B.2.2 Reactive Sites

With the aISS//GFN2-xTB workflow, also real chemical reactions are accessible if the energy barriers from the initially found NCI complexes are small enough to be overcome during the final GFN2-xTB geometry optimizations, which simplifies the identification of reactive sites. For demonstration, the protonation of a micelle cut-out was investigated (Figure B.5).[188] These kinds of biomolecular structures gained a lot of attention due to their possible use as an anti-tumor drug targeting system. From the experiment, it is known that the benzoic imine moiety (Figure B.5) is the pH-sensitive linker and gets cleaved reversibly in acidic solutions.[188] To explore this behavior, aISS was used with the implicit ALPB water solvation model[181] to add an oxonium ion to the micelle cut-out to simulate

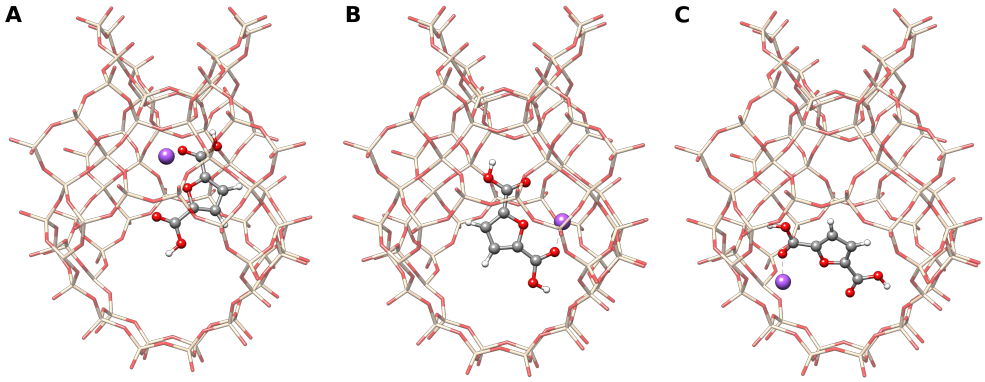

Figure B.6: Favored interaction site (A) and two directed interaction sites (B, C) of a sodium cation and a tetrahydrofuran-2,5-dicarboxylic acid at a faujasite-based zeolite according to the aISS//GFN2-xTB algorithm. Hydrogen atoms are omitted for clarity. Sodium cations are depicted in purple, silicon in beige.

acidic conditions. Even though the substrate exhibits many different basic interaction sites like alcohol, ether, amino, and aldehyde functionalities, the result shows that not only was the correct protonation position found, but also the experimentally observed proton transfer occurred during the subsequent GFN2-xTB geometry optimization. Thus, realistic modeling of small-barrier reactions can easily be done with the aISS.

B.2.3 Directed Interaction Site Screening

Another unique feature of the aISS algorithm is the directed addition of molecules to certain functional groups and sites defined by the user. Hence, it becomes possible to study interactions at possibly active sites of a molecule that must not be energetically most favored or to restrict the search space according to chemical knowledge or interest. This can be done in two ways: by either adding a distance-dependent repulsive potential to the atoms outside a user-defined region or by adding an attractive potential to the desired interaction site. Both bias potentials are added only for the xTB-IFF energy screening to prevent unphysical structures after the subsequent geometry optimization. The repulsive potential is useful for strongly interacting molecules with large interaction sites, e.g., large biomolecules like proteins and DNA. Here, an attractive potential would possibly result in a too short distance between the two fragments as the strong attraction is further enhanced. The attractive potential is preferred for weakly interacting or sterically crowded molecules like stereoselective catalysts. Applying the repulsive potential to systems with such sterically crowded interaction sites might lead to a complete repulsion of the added fragment. As an example, different interaction sites of a faujasite zeolite (Figure B.6) are investigated for catalytic activity. These kinds of catalysts offer different confinements for exchangeable sodium cations that can be used for Diels-Alder reactions. Thereby, the sodium placement on multiple accessible sites does play an important role in the course of different Diels-Alder reactions.[190] These sites can be modeled easily by performing a directed placement of the sodium cations first and then adding the substrate. Fixing the cut-out geometry of the zeolite in the aISS//GFN2-xTB run ensures that the crystal structure geometry is kept. Some of the resulting complexes are shown in Figure B.6. The different positions of the sodium cation directly influence the orientation of the substrate, which will result in spatially and electronically different surroundings.

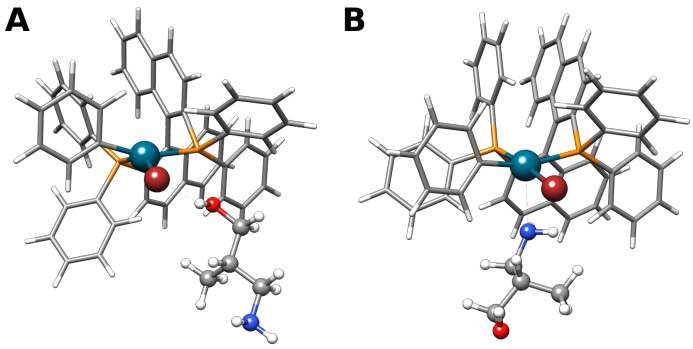

Figure B.7: Favored interaction site (A) and to the amine directed interaction site (B) according to the aISS//GFN2-xTB algorithm. Pd is depicted in dark turquoise, Br in dark red, and P in orange.

This causes different GFN2-xTB interaction energies: $-206.7\,\mathrm{kcal\,mol^{-1}}$ for the non-directed docking (Figure B.6, A), while the other positions yield energies of $-201.1\,\mathrm{kcal\,mol^{-1}}$ (Figure B.6, B) and $-192.9\,\mathrm{kcal\,mol^{-1}}$ (Figure B.6, C). Another example is from the field of transition-metal catalysis: the Buchwald Hartwig amination of bromobenzene and (S)-3-amino-2-methylpropan-1-ol with a Pd(BINAP) catalyst.[189] After the oxidative addition of bromobenzene to the catalyst, the amine approaches the palladium center for the following reductive elimination. Modeling this intermediate without the use of the directed interaction site screening leads to the formation of a halogen bond between bromine and the alcohol moiety (Figure B.7, A). To focus on the amine group, the attractive bias potential in the aISS algorithm can be used. This leads to the low-energy structure with the amine directly bound to the palladium of the catalyst (Figure B.7, B), as required for the further course of the reaction. Hence, the directed interaction site screening can not only be used to investigate different adsorption sites, but also to generate realistic geometries for the clarification of reaction mechanisms.

B.3 Conclusion

The generation of general dimer, oligomer, and aggregate geometries is a challenging problem for computational chemists. With aISS, a general and robust automated interaction site screening workflow is presented that efficiently generates physically reasonable intermolecular geometries of molecules containing thousands of atoms with elements up to radon. It is easily applicable, requires only fragment input coordinates, and can iteratively be employed to investigate even complex reaction mixtures. For medium-sized test systems, the generated structures are comparable in quality (and sometimes even better) than those from elaborated MD-based searching methods. The huge computational time savings enable the investigation of systems with several thousand atoms that could not be treated before. Moreover, reactive sites can be identified, and low-barrier reactions like protonations can be modeled with aISS//GFN2-xTB. One of the unique features is the directed interaction site screening of molecules at certain regions or functional groups, allowing the treatment of user-defined reactions or binding sites. A computer program implementing aISS can be downloaded free of charge,[180] and detailed instructions on how to use the program can be found online.[285]

Supporting Information

The Supporting Information is available free of charge at
https://onlinelibrary.wiley.com/doi/full/10.1002/anie.202214477.

- Computational details, details on the aISS algorithm, and additional examples (PDF)
- All relevant structures in xyz format (ZIP)

Conflict of Interest

There are no conflicts to declare.

Acknowledgements

This work was financially supported by Merck KGaA. The authors thank Dr. M. Bursch, Dr. S. Ehlert, T. Gasevic, Dr. A. Hansen, and Dr. S. Spicher for helpful discussions.

APPENDIX C

Automated Molecular Cluster Growing for Explicit Solvation by Efficient Force Field and Tight Binding Methods

Sebastian Spicher,[*,†] Christoph Plett,[*,†] Philipp Pracht,[*] Andreas Hansen,[*] Stefan Grimme[*]

Reprinted (adapted) with permission[‡] from:
S. Spicher, C. Plett, P. Pracht, A. Hansen, and S. Grimme, J. Chem. Theory Comput. **18** (2022) 3174.

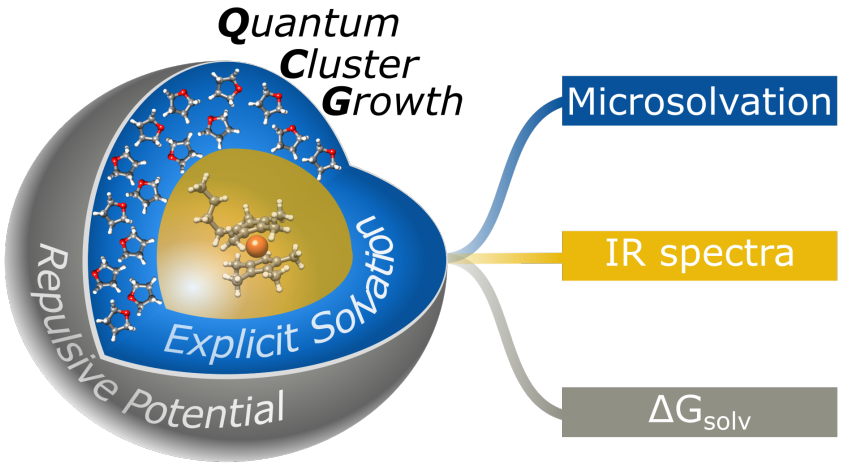

Figure C.1: Associated Table of Contents graphic for publication in Journal of Chemical Theory and Computation.

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

Abstract

An automated and broadly applicable workflow for the description of solvation effects in an explicit manner is introduced. This method, termed *Quantum Cluster Growth* (QCG), is based on the semiempirical GFN2-xTB/GFN-FF methods, enabling efficient geometry optimizations and MD simulations. Fast structure generation is provided by the intermolecular force field xTB-IFF. Additionally, the approach uses an efficient implicit solvation model for the electrostatic embedding of the growing clusters. The novel QCG procedure presents a robust cluster generation tool for subsequent application of higher-level (e.g., DFT) methods, to study solvation effects on molecular geometries explicitly or to average spectroscopic properties over cluster ensembles. Furthermore, the computation of the solvation free energy with a supermolecular approach can be carried out with QCG. The underlying growing process is physically motivated by computing the leading-order solute–solvent interactions first and can account for conformational and chemical changes due to solvation for low-energy barrier processes. The conformational space is explored with the NCI–MTD algorithm as implemented in the CREST program, using a combination of metadynamics and MD simulations. QCG with GFN2-xTB yields realistic solution geometries as well as reasonable solvation free energies for various systems without introducing many empirical parameters. Computed IR spectra of some solutes with QCG show a better match to the experimental data compared to well-established implicit solvation models.

C.1 Introduction

Current chemical research, e.g., on stereoselective catalysis in organic synthesis,[292] electrochemical capacitors for energy storage,[293] or protein folding in living organisms,[294,295] is all connected by the fundamental question of how solvent molecules interact with solutes and surfaces in the condensed phase. To answer this question, an adequate description of solvation effects is inevitable. Nowadays, theoretical chemistry is capable of providing highly accurate quantum mechanical (QM) calculations in the gas phase, whereas most experiments are carried out in solution. To compare experimental findings and theoretical simulations, a reliable solvation model has to be included in the calculation. Quantitative theoretical predictions of thermodynamic properties for molecules in solution require an accurate description of the interaction between solvent molecules themselves and their very specific interaction with the solute.[296] Therefore, the calculation of mass densities, enthalpies of vaporization, heat capacities, surface tensions, dielectric constants, solvation free energies, and other properties of molecules in solution remains a challenging task for computational chemistry and is part of current theoretical research[297–299].

Methods for evaluating solvation effects can be roughly classified into two categories.[300] Explicit solvation models[97,301,302] describe the individual solvent molecules, whereas implicit models[153,303] treat the solvent as a continuous medium, mainly characterized by its dielectric constant ε [303]. Combinations of explicit/implicit solvation models are conceivable, in which, e.g., the first solvation shell is built up from explicitly placed solvent molecules, while the remaining shells are treated implicitly by continuum embedding. Furthermore, each of these methods may be conducted at the classical molecular mechanical (MM) or quantum mechanical (QM) level of theory, or the combination of both in so-called QM/MM schemes. [304,305] Recently, great advances were made in the context of polarizable molecular mechanics models that are commonly used to compute many molecular properties and that can reproduce solvation effects.[306–308] Continuum solvation models consider

the solvent as a continuous isotropic medium. The solvent is replaced by an electric "reaction field" that represents a statistical average of all solvent degrees of freedom at thermal equilibrium.[153] The solute is placed in a suitably shaped hole in the medium, thus creating a cavity. In standard polarizable continuum models (PCM),[309] the polar electrostatics can be decoupled from the nonpolar interactions, and the solvation free energy can be written as

$$\delta G_{\text{solv}} = \delta G_{\text{cavity}} + \delta G_{\text{disp}} + \delta G_{\text{elec}}, \tag{C.1}$$

where the three terms are the free energy of cavitation, dispersion, and electrostatic energy.[310] The latter contribution in PCM models is mostly approximated through solving the Poisson–Boltzmann (PB) equation[311,312] or further simplified by using the generalized Born (GB)[147,313–315] model. Nowadays, the most commonly applied version of the PCM model is a reformulation of dielectric PCM (DPCM) in terms of the integral equation formalism termed IEFPCM[316]. The conductor-like screening model/conductor-like screening model for real solvents (COSMO[155]/COSMO-RS[143,144]) is a variation on Poisson–Boltzmann PCM. In COSMO, the dielectric permittivity is set to infinity ($\varepsilon = \infty$), which defines the solvent as a conductor. COSMO-RS results from the combination of the COSMO approach with the statistical thermodynamics of interacting surfaces.[317] The COSMO approach is also employed in a further variation of PCM, the conductor-like polarizable continuum model (CPCM)[154]. Other continuum solvation models, e.g. SMD, also include the QM charge density of a solute molecule interacting with a continuum description of the solvent.[145] Implicit solvation models are computationally efficient and were successfully applied in many computational studies (see, e.g., Refs.318,319,320,321). COSMO and COSMO-RS are the default models for computational chemistry work at the DFT level in our group. Nevertheless, the main disadvantage of implicit models remains the inadequate description of very polar or charged species, mainly because strong directional and local interactions between solute and solvent molecules, *i.e.*, noncovalent interactions, ionic and hydrogen bonds, are not treated properly. By neglecting explicit solvent molecules, important interactions are missing or only described poorly.[68,322–324] Explicit solvation models use molecular dynamics (MD)[325,326] or Monte Carlo (MC) statistical mechanics[327,328] simulations to generate molecular ensembles and corresponding energies. Differences in free energies are obtained by applying free energy perturbation (FEP) theory,[158] thermodynamic integration (TI),[161] or Bennett's acceptor ratio (BAR) method.[163] Free energy methods that use data from MD or MC simulations require a large number of steps to converge, and thus, suffer from the issue of insufficient phase space sampling to estimate the ratio of partition functions. For a short overview of explicit solvation treatments and related approaches, see Refs. 159,329,330,331. At this point, quantum cluster equilibrium (QCE) methods should also be mentioned, comprising the essential idea of applying statistical mechanics to quantum chemically calculated clusters to obtain thermodynamic properties of the liquid (condensed) and the vapor phase.[332–334]. In the SAMPL5 challenge of calculating host–guest binding free energies, it was found that methods involving explicit solvent molecules, in general, perform better than implicit solvation models.[335] Therefore, recent developments turned their sights on hybrid cluster–continuum model approaches, where explicit water molecules were added to the continuum model to describe alkane complexation in self-assembled capsules or calculate the solvation free energies of small molecules in aqueous solution.[65,299,336] In a recent study by Bensberg *et al.*, the electrostatic PCM energy was partially replaced by an explicit solvent molecules treatment in the context of sub-system density functional theory (DFT), which was already successfully applied to systems of different sizes containing water and cyclohexane as solvents. [63]

Here, the idea of combined cluster–continuum models is extended significantly in terms of molecular size and versatility. We propose a new QM and force-field (FF) based hybrid solvation model, in which large molecular clusters of any solute are generated fully automatically by successively adding explicit solvent molecules. With cluster sizes up to a few hundreds of atoms and the possibility to include all elements up to radon ($Z \leq 86$), a large part of the chemical compound space can be covered and any solute–solvent combination is in principle accessible, with no restriction in charge or spin state. This newly developed procedure is denoted as *Quantum Cluster Growth* (QCG) and can be applied to study the effect of explicit solvation on various properties at the QM level of theory. For computational efficiency reasons, semiempirical quantum mechanical (SQM) methods in combination with even faster FFs are employed[84] in the generation process. The resulting cluster ensembles may serve as input for subsequent high-level DFT or wave function theory calculations. In the context of microsolvation, the herein proposed cluster growing algorithm can automate the detection of important interaction sites and therefore, replace laborious approaches[248,337–339] or can be an alternative to other workflows.[193,340] An example for a recently developed, automated explicit solvation workflow is the AutoSolvate toolkit for generating clusters of organic solutes in several solvents.[341] Further, the QCG algorithm is extended to compute solvation free energies by a novel supermolecular ansatz. In contrast to many existing proposals, this includes the explicit calculation of cluster entropies, giving access to δG_{solv} as well as δH_{solv} values.

After a description of the theoretical background, technical details of the QCG algorithm are given and the cluster generation process is examined statistically. In a first application example, the quality of the generated clusters is assessed in comparison to other (micro)solvation tools. The effects of explicit solvation on molecular geometries are evaluated in the framework of MD simulations and IR spectra calculation. Solvation free energies are computed for a test set of small organic molecules in comparison to established implicit solvation models. As an outlook for future applications, free association energies of supermolecular complexes in solution are calculated.

C.2 Theoretical Background

QCG represents a fully automated approach to describe a molecule in solution in an explicit manner at a QM level of theory. Therefore, molecular clusters of the solute (A) with a given number n of solvents (S) are generated by adding one solvent molecule at a time to an energetically favorable position

$$\begin{aligned} A_{(\mathrm{gas})} + S_{(\mathrm{gas})} &\longrightarrow [AS]_{(\mathrm{gas})} \\ [AS]_{(\mathrm{gas})} + S_{(\mathrm{gas})} &\longrightarrow \left[AS_2\right]_{(\mathrm{gas})} \\ &\vdots \\ \left[AS_{n-1}\right]_{(\mathrm{gas})} + S_{(\mathrm{gas})} &\longrightarrow \left[AS_n\right]_{(\mathrm{gas})} . \end{aligned} \tag{C.2}$$

The square brackets in Equation C.2 indicate a noncovalent interaction (NCI) complex. The formation of new covalent chemical bonds between solvent and solute is possible as long as the underlying QM or FF method can describe it and deserves no special attention. First, the fully automated cluster ensemble generation procedure is outlined in Section C.2.1. In Section C.2.2, an extension to the QCG algorithm is proposed that enables the computation of solvation free energies.

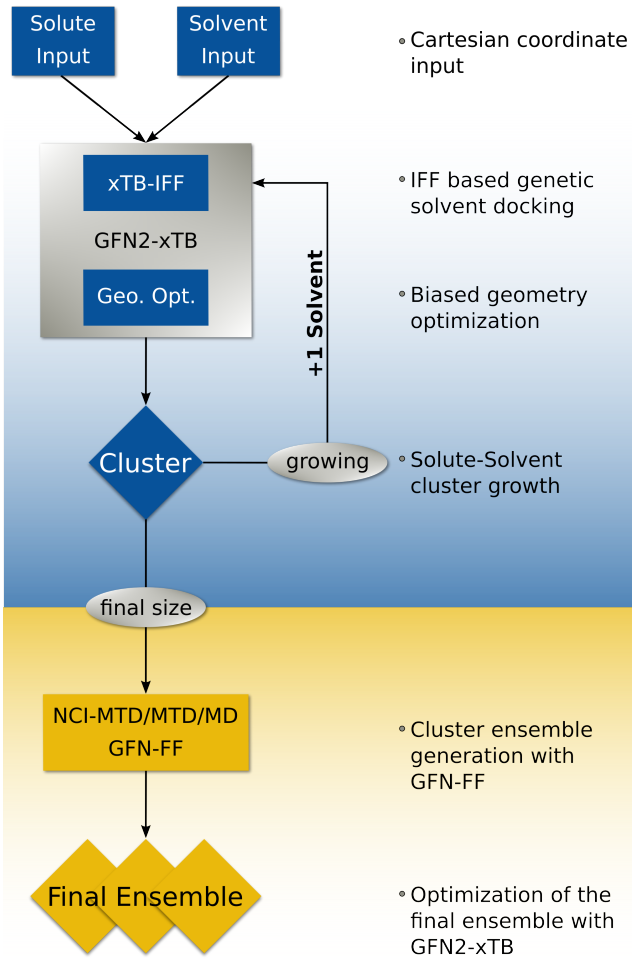

Figure C.2: Schematic illustration of the QCG algorithm.

C.2.1 Cluster Ensemble Generation

The automated cluster generation part of the QCG algorithm may be subdivided into two steps. Both are illustrated in the QCG workflow in Figure C.2.

The first part describes the cluster growth process, consisting of a repeating cycle in which each turn increases the cluster size by one solvent molecule. As input, only the solute and solvent geometries are required. Optimal complexation (docking) positions for added solvents are determined with a genetic (global) optimization algorithm employing the intermolecular force field xTB-IFF[47]. The necessary QM information is generated on-the-fly with GFN2-xTB[49] and consists of the Mulliken atomic charges, charge centers of localized molecular orbitals, frontier orbitals, and orbital energies. The interaction energy surface between the growing cluster and the added solvent molecule is screened, and the most favorable positions at xTB-IFF level are determined and re-optimized at the GFN2-xTB level of theory. Repulsive wall potentials (*vide supra*) are applied throughout the growing process to "shape" the solute–solvent cluster properly. Complete and consistent coverage of the solute with a minimum number of solvent molecules is the target. The great advantage of QCG in comparison to

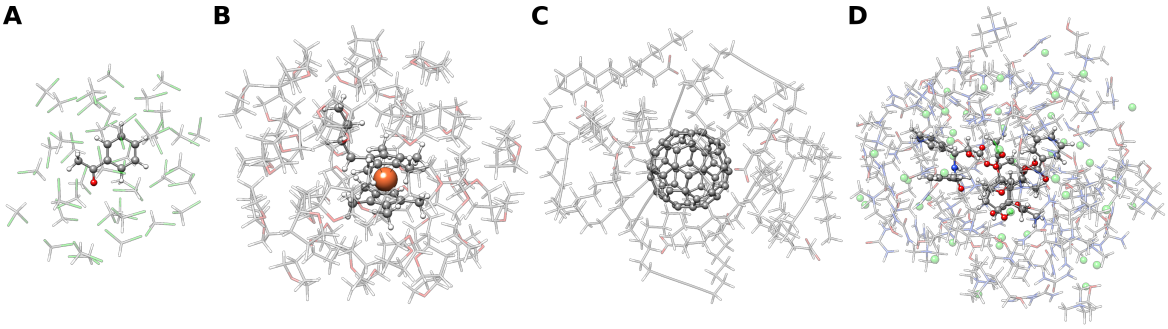

Figure C.3: Examples of QCG generated solute–solvent clusters: (A) acetophenone solvated by 40 explicit molecules of dichloromethane. (B) Butylferrocene (n-Butylcyclopentadienyl(cyclopentadienyl)iron(II)) surrounded by 55 molecules of THF. (C) Fullerene C_{60} solvated by 10 PCDA (10,12-pentacosadiynoic acid) molecules. (D) Taxol within a eutectic solvent consisting of choline chloride and urea.

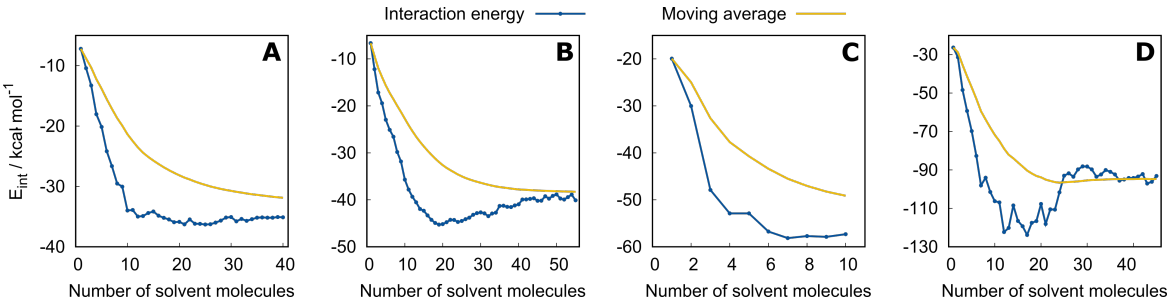

Figure C.4: Solute–solvent interaction energy (blue) and moving average (yellow) as a function of cluster size for the examples in Figure C.3.

other explicit solvation tools is the general applicability, as most of the required parameterization is inherent and already existing in the underlying SQM and FF methods. To highlight this characteristic, molecular clusters of various (exotic) solute–solvent combinations were generated with QCG and are depicted in Figure C.3.

In QCG, the interaction energy of the solute with the surrounding solvent molecules is computed in each growing step. It could be taken as a convergence criterion for the growing and is fulfilled when the change in energy by the addition of another solvent molecule becomes smaller than a certain threshold. We observed an irregular convergence behavior due to fluctuations in the solvent shell, which was also observed in other cluster–continuum approaches.[65,340] Hence, the simple moving average of the interaction energy is chosen instead as an alternative convergence criterion, where the threshold is set to 10^{-4} E_h. The convergence behavior of the interaction energy E_{int} (and its moving average) with regard to the cluster size is shown in Figure C.4 for the previously introduced examples (*cf.* Figure C.3). It is found that for systems (A) (acetophenone in CH_2Cl_2) and (C) (C_{60}) in PCDA) the interaction energy converges when roughly the first solvation shell is filled. Interestingly, for (B) (butylferrocene in THF) and (D) (taxol in an eutectic solvent) filling the second solvation shell leads to rearrangements in the first shell, leading to an increase in the solute–solvent interaction energy. As the QCG algorithm optimizes the total cluster energy rather than the solute–solvent interaction, an increase in E_{int} can occur. Although the moving average would be applicable as a convergence threshold in general, the number of solvent molecules can vary in different cluster growth runs. Hence, in the

following, the target number of added solvent molecules (user input) is chosen as an exit criterion in the algorithm due to practical reasons and to obtain better reproducibility.

The QCG algorithm is non-deterministic, and a single cluster geometry is of limited value. With increasing cluster size, the number of local minima on the PES grows drastically, giving rise to millions of possible conformations already for medium-sized systems. Thus, statistical averages over many parallel generated clusters must be computed to determine equilibrium properties and the entropic contribution of each cluster. Therefore, the second step of the cluster generation process in QCG is the sampling of phase space by a combination of MD and metadynamic (MTD) simulations. We employ the NCI-MTD algorithm as implemented in the CREST program to generate an ensemble of energetically low cluster structures.[44,177] However, the computational demands of this algorithm for common computational resources can only be met at the FF level of theory. Therefore, the recently developed GFN-FF[50] method is used as the underlying level of theory. This combination has already been successfully applied in similar situations to determine protein conformations and to bind gases in metal-organic cages.[267,342] As an alternative to NCI-MTD, similar algorithms are implemented in QCG, where just one MTD or MD simulation is performed instead of multiple ones. The cluster ensemble generated at the FF level is then re-optimized at the GFN2-xTB level of theory. Thereupon, final single-point energy calculations are performed in the absence of any constraints (wall potentials), always at the same level to obtain the averaged ensemble energy $\overline{E}$. This can be done, with or without the implicit ALPB or GBSA solvation model[84,181] to minimize artificial surface effects.

Wall Potentials

An unbiased cluster growing process may lead to incomplete coverage of the solute and to an inconsistent description of the solvated system as it usually occurs in the bulk solvent. To enhance the coverage of the solute and to prevent irregularities in the growing process, a dynamic, repulsive outer wall potential is applied to shape the outermost solvent shell. The form of the repulsive potential is chosen as an ellipsoid. This choice allows for adapting to the solute's geometry, varying from spherical to axial, and resulting in a more uniform distribution of the solvent molecules around the solute.

Thus, the outer wall potential prevents the accumulation of solvent molecules at a specific binding site. In addition, a second inner wall potential is applied to keep the solute fixed in the center of the growing cluster. This prevents, e.g., the move of hydrophobic molecules in polar solvents like water to the cluster surface. Both potentials are applied within the xTB-IFF docking steps, the GFN2-xTB geometry optimizations, and the conformer search (NCI-MTD). The arrangement of the potentials is illustrated for an exemplary system in Figure C.5. The reasonable choice of the potential is a challenging task. On the one hand, the potentials must ensure that any solute molecule is covered by any solvent, which is difficult for solute–solvent combinations with very different polarities and structures. On the other hand, the potentials should allow conformational reorganization during the growing process and hence should not be too restrictive. The QCG algorithm calculates the three principal axes of an ellipsoid according to different geometrical and solute/solvent-specific criteria. The closer a molecule is to the surface of the ellipsoid cavity, the more repulsive is the applied wall potential. The energy contribution E_{pot} given by the ellipsoid potential is defined as a steep polynomial function

$$E_{\mathrm{pot}} = \sum_{i}^{N} \left(\frac{\mathbf{R}_i - \mathbf{O}}{\mathbf{R}_{\mathrm{pot}}} \right)^{10} . \tag{C.3}$$

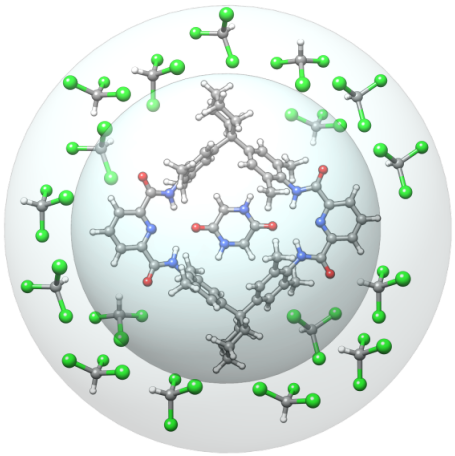

Figure C.5: Inner and outer wall potential applied during the QCG procedure to shape the molecular clusters.

In Equation C.3, 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 $\mathbf{R}_{\text{pot}}$ are the principal semi-axes of the ellipsoid potential parallel to $\mathbf{R}_i - \mathbf{O}$. The inner potential is only applied to the solute to fix its position during the cluster generation process. Therefore, it is rigid and does not change within the growing process. The outer potential is dynamic and increases as the number of solvent molecules within the cluster rises. For the solute and solvent molecule, diagonalization of the inertia tensor results in the principal moments and unit ellipsoid axes $a \geq b \geq c$ and yields important information about the molecular geometry. In addition, the excluded volume of overlapping spheres V is computed via analytic equations for solute and solvent molecules using the `arvo` package.[343] To further describe the geometry of a molecule, a structural factor F_α is introduced to account for further geometrical properties

$$F_\alpha = \sqrt{1 + \frac{a - c}{\frac{1}{3}(a + b + c)}}. \tag{C.4}$$

F_α is a measure of how strongly a molecule's shape differs from an ideal sphere. For spherical molecules, F_α simply reduces to unity, and for arbitrarily shaped molecules $F_\alpha \geq 1$ holds true. The volume and radius of a cavity hosting the solute and n solvent molecules need to be properly sized. For the outer cavity, all requirements are fulfilled in Equation C.5

$$R_{out} = \left[\frac{3}{4\pi}\left(\frac{F_\alpha}{2} \cdot n\, V_{solvent} + V_{solute}\right)\right]^{\frac{1}{3}} + \beta \cdot R_{max} + \gamma_1, \tag{C.5}$$

where the cavity radius is determined by calculating the third root of the added molecular volumes. The contribution of the solvent molecules to the overall volume is scaled by the geometrical correction factor F_α. R_{max} is the maximal internal distance within the solvent molecule scaled by the empirically determined factor β (usually = 0.5). γ_1 is an added constant to damp the long-range effects of the polynomial wall potential. Adding more solvent molecules to the cluster increases the radius of the cavity, leading to a dynamic outer wall potential. The static inner cavity is hosting the solute, and its radius is calculated according to Equation C.6. The radius is independent of the number, where R_{max}

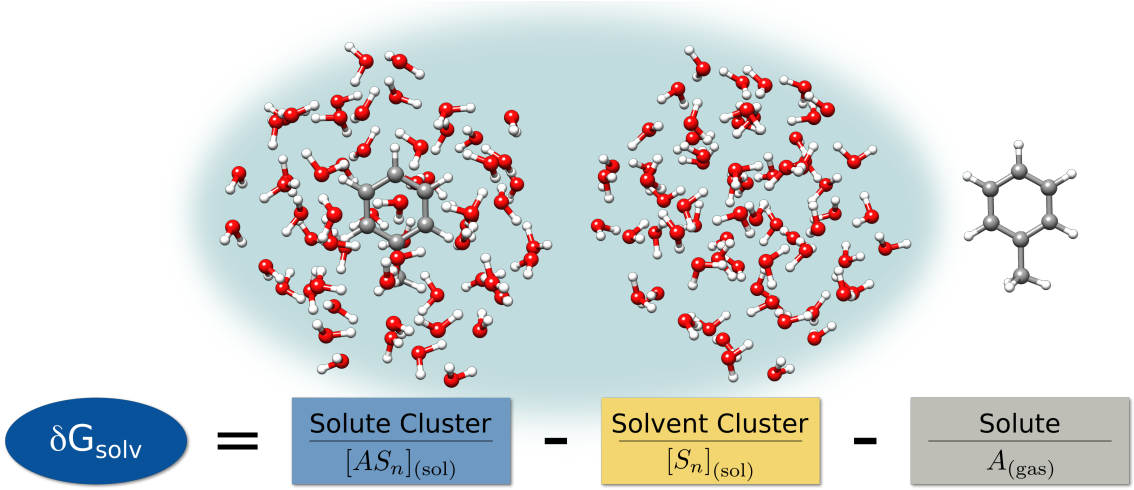

Figure C.6: Example system of toluene in 60 molecules of water for the calculation of the free solvation free energy δG_{solv} according to a supermolecular approach.

is the maximal internal distance within the solute

$$R_{in} = \left(\frac{3\,F_{\alpha}}{4\pi} \cdot V_{solute} \right)^{\frac{1}{3}} + \beta \cdot R_{max} + \gamma_2. \tag{C.6}$$

To obtain the desired principal semi-axes $\mathbf{R}_{pot}$ (*cf.* Equation C.3) of an ellipsoid, the radii of the inner and outer cavity are projected onto the unit axes of the solute. This results in an ellipsoid capable of hosting the cluster during the generation process.

C.2.2 Solvation Free Energies

One of the fundamental quantities to describe the interaction of a solute with a surrounding solvent is the free energy of solvation δG_{solv} that describes the change in free energy upon transferring a molecule from the (ideal) gas state to a solvent at a certain temperature and pressure.[159,338,344,345] This quantity can be calculated with QCG in a supermolecular approach as the difference in total free energy of the isolated solute A (gas phase), and the "filled" $[AS_n]$ and "empty" $[S_n]$ clusters visualized in Figure C.6. This can be applied to any solute and requires the total free energies of the solute in the gas phase $G(A)_{(\text{gas})}$, the pure solvent cluster $G([S_n])_{(\text{sol})}$ and the solute cluster $G([AS_n])_{(\text{sol})}$ (Equation C.7), each consisting of the respective total electronic energy and the sum of corrections from energy to free energy in the modified rigid-rotor-harmonic-oscillator approximation (mRRHO) including zero-point-vibrational energy (ZPVE).[138]

$$\delta G_{\text{solv}}(A) = G([AS_n])_{(\text{sol})} - G([S_n])_{(\text{sol})} - G(A)_{(\text{gas})}. \tag{C.7}$$

To keep the algorithm computationally efficient also for large molecular clusters, GFN2-xTB is employed to calculate the electronic energies and GFN-FF for the thermostatistical mRRHO contributions. In QCG, the solute is surrounded by a limited number of solvent molecules, which resembles only a cutout from the infinitely diluted solute in the condensed phase of the solvent. Hence,

for finite cluster sizes, an additional embedding in a GBSA continuum model is employed.[181] This approach is common practice in cluster–continuum models and leads to faster convergence of δG_{solv} as a function of the cluster size.[65,340,346] For $n \rightarrow \infty$ the effect of the continuum model is vanishing. In the next step, it is important to distinguish between the free energy of an individual equilibrium structure and the ensemble value of the previously generated and optimized cluster structures (see Section C.2.1). Assuming that all degrees of freedom (DOFs) are separable, the free energy of the (cluster) structure ensemble (SE) is obtained as[35],

$$G_{\mathrm{SE}} = \overline{G} + G_{\mathrm{conf}}, \tag{C.8}$$

where G_{conf} is the conformational free energy part of N distinguishable conformers in the SE calculated from the Gibbs–Shannon entropy according to Ref. 347

$$G_{\mathrm{conf}} = -T\, S'_{\mathrm{conf}} = RT \sum_{i}^{N} p_i \, ln\left(p_i\right) \ . \tag{C.9}$$

Hence, the conformational free energy G_{conf} is included if a complete ensemble of low-energy clusters is found. Further, the average $\overline{G}$ in Equation C.8 is given by

$$\overline{G} = \sum_{i}^{N} p_i G_i, \tag{C.10}$$

with the Boltzmann weight p_i

$$p_i = \frac{e^{-G_i/k_{\mathrm{B}}T}}{\sum_j^N e^{-G_j/k_{\mathrm{B}}T}}, \tag{C.11}$$

and the molecular free energy $G = E + G_{\mathrm{mRRHO}}$ of the ensemble member i. Note that the molecular entropy of each species (solute cluster, solvent cluster, solvent) is explicitly calculated. Taking into account the SE of the solute clusters, the solvent clusters, and the solute molecules, Equation C.7 must be rewritten in terms of structure ensembles

$$\delta G_{\mathrm{solv}}(A) = G_{\mathrm{SE}}([AS_n])_{(\mathrm{sol})} - G_{\mathrm{SE}}([S_n])_{(\mathrm{sol})} - G_{\mathrm{SE}}(A)_{(\mathrm{gas})}. \tag{C.12}$$

Hence, for a complete cluster ensemble, *i.e.*, $n \rightarrow \infty$, the solvation energy should approach the true solvation free energy for transferring 1 mol/L solute from the gas phase into solution with the same concentration. The required volume work $p\Delta V$ is added to Equation C.12. Considering the SE for the solute molecule in the gas phase is necessary for flexible molecules, and is mentioned here for the sake of completeness. The application examples (*vide infra*) are restricted to rather rigid molecules and hence the ensemble average for $[A]_{(\mathrm{gas})}$ can be neglected. If all entropic contributions are discarded and only the zero-point-vibrational and $H(T)$ terms are taken from the mRRHO calculation, the solvation enthalpy $\delta H_{\mathrm{solv}}(A)$ is obtained directly, without additional computational effort.

$$\delta H_{\mathrm{solv}}(A) = H_{\mathrm{SE}}([AS_n])_{(\mathrm{sol})} - H_{\mathrm{SE}}([S_n])_{(\mathrm{sol})} - H_{\mathrm{SE}}(A)_{(\mathrm{gas})}. \tag{C.13}$$

Hence, the extended QCG approach also enables the calculation of solvation enthalpies and entropies separately, which is, to the best of the authors' knowledge, a unique feature among existing hybrid-

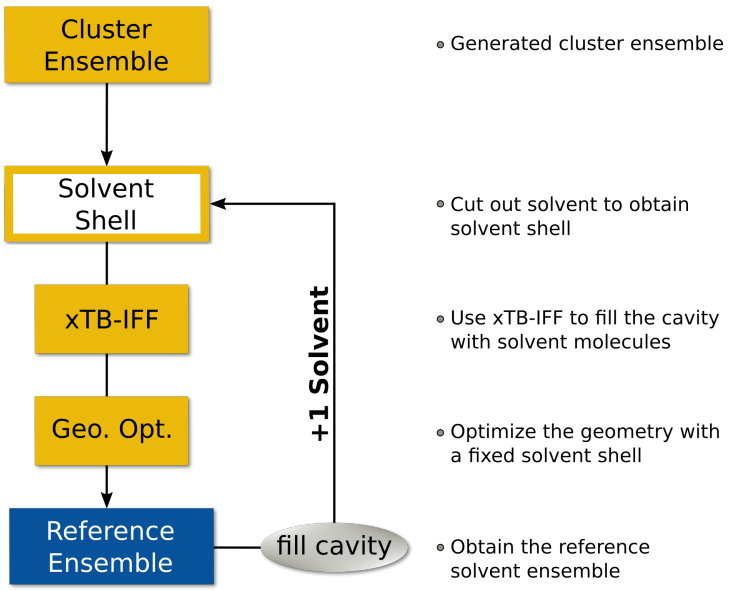

Figure C.7: Workflow of the cut-fix-fill (CFF) algorithm. The solute is cut out from the cluster and replaced by solvent molecules. The energy of the desired reference cluster is interpolated from the filled cluster.

continuum models. The subtracting scheme given in Equation C.12 automatically includes the energy required to create a cavity inside the solvent G_{cavity}, including entropy contributions and the loss of solvent–solvent van der Waals interactions. Hence, in contrast to implicit solvation models, the estimation of G_{cavity} is not needed in the QCG theory. Moreover, the electronic energies of the ensembles are computed by single-point computations without the wall potentials to ensure that the biasing potential does not enter the solvation free energy directly.

Generating Reference Clusters

Reference ("empty") clusters of pure solvent molecules are generated by employing the newly developed *cut-fix-fill* (CFF) algorithm. A schematic representation is given in Figure C.7. The computational effort invested in the cluster ensemble generation is recycled in the process of generating the reference cluster ensemble by replacing the solute molecule in the generated clusters with a number of solvent molecules. In a first step, the solute is cut out of the cluster and the remaining solvent shell is kept fixed in the following steps by setting the forces on the respective solvent molecules to zero. Secondly, analogous to the growing procedure, xTB-IFF docking and GFN2-xTB optimization steps are applied to fill the created cavity in a step-wise fashion. The cluster is considered to be filled when no further solvent molecule can be placed inside the cavity, which is supervised by energetic (positive interaction energy) and geometric (volume of inserted solvent exceeds solute volume) criteria. The CFF algorithm unavoidably increases the number of molecules. Thus, the energy of the reference cluster with the initial count of solvents is calculated from the energy of the filled cluster by multiplication with a factor $n/(n+x)$, where n is the initial number of solvents and x the number of solvents placed inside the cavity. The reason for using the CFF approach compared to an additional MD-based ensemble generation is that a close structural similarity of the solvent shell in both the solute and the reference

cluster is ensured, leading to beneficial error compensation effects. This minimizes statistical noise and is important since slight changes in the solvation shell may cause significant energy changes, especially for highly polar and protic solvents.

C.3 Technical Details

The QCG algorithm is implemented in the open-source CREST[348] program and a detailed manual with examples can be found online.[285] Here, we will limit the explanation only to the most important features. The QCG algorithm is invoked by `crest <solute> -qcg <solvent>`. The user has to provide only the solute and solvent input coordinates. As discussed in the previous sections, the QCG algorithm consists of: i) the cluster growth, ii) the ensemble generation, and iii) the reference cluster construction for the computation of solvation free energies, including the respective calculation of thermostatistical contributions. The cluster growth is invoked with the addition of `-grow` to the command line input and can be carried out with all GFN methods. By default, GFN2-xTB is employed as the xTB-IFF docking is usually the time-limiting factor. The added number of solvent molecules can either be set manually (`-nsolv`), or automatically determined by a moving average threshold of the solute–solvent interaction energy (*cf.* section C.2.1). In the case of water as solvent, the outer wall potential is scaled per default by 0.7. This factor is increased by 5 % each time the interaction energy is positive or if the default of 1.0 is reached. It can be adjusted by the user for any solvent.

To conduct the ensemble generation of the grown cluster, the `-ensemble` flag is employed (instead of `-grow`). As a starting point, a cluster after the growth algorithm is used that is either generated during the same or a previous run. Therefore, a single MD/MTD simulation or the NCI-MTD run type is available. By default, the latter is performed at the GFN-FF level of theory with MTD and MD simulations of 10 ps length, respectively. To obtain the qualitatively best cluster ensemble for reasonable computational costs, final geometry optimizations are conducted by GFN2-xTB. The MD/MTD length can also be varied by the command line input. Choosing the single MD/MTD simulation instead of the NCI-MTD run type, a GFN2-xTB MD/MTD simulation of 10 ps length is performed by default at 298 K. The computation of the solvation free energy is invoked by `-gsolv` and requires a solute–solvent cluster ensemble. δG_{solv} values are computed according to Equation C.12, where the free energies of the solute and reference clusters are obtained as the Boltzmann weighted average of the ensemble (G_{SE}, *cf.* Equation C.10). By default, every solute cluster populated by more than 10 % is taken into account. Starting from these structures, the reference cluster ensemble is constructed by the CCF algorithm (*cf.* section C.2.2). Every geometry optimization during the CFF algorithm is performed with GFN2-xTB, similar to the ensemble generation step. For the final single-point calculation, the solute- and reference clusters are additionally embedded in a continuum solvation model. Either the GBSA or ALPB solvation model can be chosen.[84,181] The total free energy of a cluster is calculated as the sum of the single-point energy and thermostatistical contribution G_{mRRHO}. Therefore, QCG computes the harmonic vibrational frequencies of all solute- and reference clusters populated more than 10 % to ensure computational efficiency by considering the most relevant structures. For this step, GFN-FF is the default, but all GFN methods can be applied. Entering the liquid phase limits the translational and rotational degrees of freedom of the solute molecule. To mimic this effect, the corresponding entropic contributions to G_{mRRHO} are reduced by 25 % (*cf.*

Equation C.14)

$$G_{\mathrm{mRRHO}} = \left(H_{\mathrm{trans}} + H_{\mathrm{rot}} + H_{\mathrm{vib}}\right) - T\left[0.75\left(S_{\mathrm{trans}} + S_{\mathrm{rot}}\right) + S_{\mathrm{vib}}\right], \tag{C.14}$$

whereas the vibrational contribution remains unchanged. The scaling factor of 0.75 was empirically determined and can be adjusted for each solvent individually. Nevertheless, this imposes no restriction to the general applicability of the QCG approach. Lastly, the conformational free energy part G_{conf} of the cluster ensembles and the volume work to transfer a solute molecule from an ideal gas to an ideal solution at molar concentration (1 mol L^{-1}) are included in QCG.

C.4 Computational Details

The QCG algorithm was applied as implemented in the CREST program.[44,348] Unless stated otherwise, the default settings were employed throughout this work. Final single-point energies were calculated at the GFN2-xTB level of theory with the implicit GBSA solvation model for the solute and reference clusters.[84,181] All GFN*n*-xTB and GFN-FF calculations were performed with the xtb 6.4.0 program package.[349] To decrease the statistical error, each solvation free energy computation with QCG was performed ten times and averaged. The number of solvent molecules was determined to complete at least the first solvation shell. For comparison, solvation contributions to the free energy were calculated with COSMO-RS[143], also including the volume work required for changing from an ideal gas at 1 bar to 1 mol L^{-1} to solution. For the COSMO-RS free energy, the BP_TZVP_C30_1601 parameterization was used. Two single-point calculations with BP86[208,209]/def-TZVP[350] (one in gas-phase and one in an ideal conductor) were performed, and the output of these calculations was then processed by the *COSMOtherm* program package[144,351]. The efficient B3LYP-3c[352] and r^2SCAN-3c[129] DFT composite methods were employed as implemented in the *TURBOMOLE* program package (version 7.5.1)[353–355] together with the COSMO model. Harmonic vibrational frequencies were calculated analytically using the `aoforce` implementation in *TURBOMOLE* and scaled by a factor of 0.97 in the case of B3LYP-3c.[352] Visualization of molecules was performed with the *UCSF Chimera* (version 1.15) [356] program, and *gnuplot* (version 5.0)[357] was employed for plotting.

C.5 Results and Discussion

C.5.1 Reproducibility

The QCG algorithm consists of multiple steps (*c.f* section C.2.1), with each of them containing non-deterministic components (*i.e.*, docking or MD/MTD). Regarding the complexity of the phase space for a large, explicitly solvated cluster, prohibitively long simulation times would be required to always converge to the same solution. In practice, the finite simulation time introduces a statistical error, which is investigated here. First, we determine the standard deviation (SD) resulting from the docking procedure in the growth step, the subsequent finite simulation times for the solute ensemble, and the reference ensemble generation. Exemplarily, a system of acetonitrile in ethanol (Figure C.8) is elaborated. Different sized clusters (1–40 solvent molecules) were generated ten times each using the same settings and the SD and maximal spread (MinMax) respective to the averaged electronic energies were analyzed as well as their average over all number of solvents ($\overline{\mathrm{SD}}$ and $\overline{\mathrm{MinMax}}$). It is found that

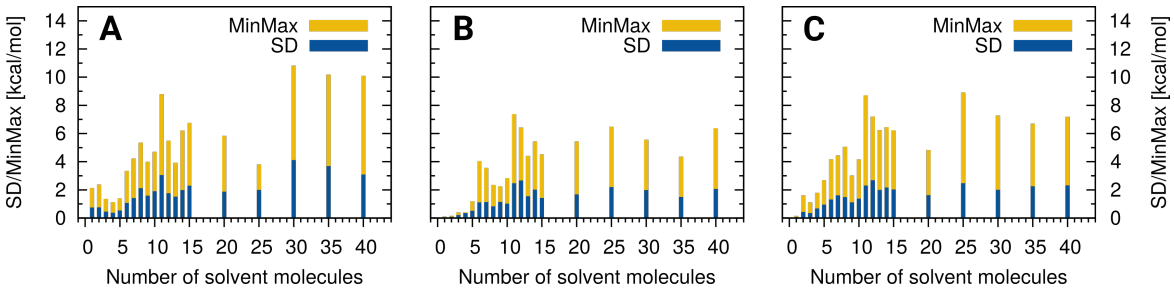

Figure C.8: Standard deviation (SD) and spread between the energetically lowest and highest cluster (MinMax) given for: (A) the electronic energies after the cluster growth, (B) conformer sampling (ensemble generation), and (C) reference ensemble generation of ethanol in acetonitrile averaged over ten runs each.

already the cluster growth algorithm consisting of xTB-IFF docking and GFN2-xTB optimizations (Figure C.8(A)) yields (slightly) different structures for the repeated computations with the same setup and introduces a scattering of the cluster energies. Averaging over all cluster sizes, the $\overline{\text{SD}}$ and $\overline{\text{MinMax}}$ amount to 1.8 kcal mol^{-1} and 5.1 kcal mol^{-1} in respect to the average energy. Different structures within the growth process result from the docking algorithm in xTB-IFF. Typically, the first few solvent molecules are always added at the same position, explaining the small SD and MinMax for small clusters. However, larger clusters exhibit energetically very similar docking positions and hence, different binding sites are occupied.

QCG employs the NCI-MTD algorithm to explore the low-energy conformational space after the growing process. In Figure C.8(B), the ensemble energy is shown as the Boltzmann-weighted sum of all clusters within a generated ensemble. Again, the statistical errors are evaluated over ten equivalent runs. In general, the conformer sampling compensates partially for the scattering introduced by the growth process, as similar energetically low clusters are found by repeated conformational sampling. The average $\overline{\text{SD}}$ reduces to 1.3 kcal mol^{-1} and the $\overline{\text{MinMax}}$ value to 3.6 kcal mol^{-1} upon the ensemble generation. For a small number of added solvents, the SD and MinMax values are close to zero. As the conformational space becomes larger with increasing cluster size, the SD and MinMax values also increase.

A further source of error is introduced by the CFF algorithm (Figure C.8(C)). Differently shaped cavities within the frozen cluster shells can be filled with a varying number of solvent molecules. This inconsistency introduces an average $\overline{\text{SD}}$ and $\overline{\text{MinMax}}$ of 1.5 kcal mol^{-1} and 4.9 kcal mol^{-1}, respectively. Overall, the error of the CFF algorithm is larger than the error after ensemble search and comparable to the error of the growing process because the added solvent molecules introduce an additional conformational error. Additionally, problems arise if spatially different conformations of the solute molecule are possible. For example, a bent conformation of n-octanol can be placed in cavities, which is otherwise too small for the extended conformation.

Increased simulation times in the NCI-MTD run were investigated. Therefore, again 10 calculations were repeated per number of solvent molecules (ranging from 1 to 15) for different MTD lengths. The resulting SDs and MinMax values for each number of solvent molecules were averaged for each MTD length. These averages ($\overline{\text{SD}}$ and $\overline{\text{MinMax}}$) are shown in Table C.1. The comparison of different simulation times shows that longer MTDs systematically reduce the scattering, e.g., increasing the MTD simulation time from 1.2 ps to 100 ps reduces the $\overline{\text{SD}}$ by 40 %. One also notices that the NCI-MTD algorithm employing 1.2 ps of simulation time has the same effect on the standard deviation

Table C.1: $\overline{\text{SD}}$ and $\overline{\text{MinMax}}$ values of the ensemble energy (in kcal mol^{-1}) for ethanol in acetonitrile with different MTD lengths during the NCI-MTD step. SD and MinMax are averaged over 15 different cluster sizes with up to 15 solvent molecules.

	1.2 ps	10 ps	50 ps	100 ps	single MD (10 ps)
$\overline{\text{SD}}$	1.05	0.97	0.83	0.63	1.08
$\overline{\text{MinMax}}$	3.19	3.02	2.54	1.87	3.25

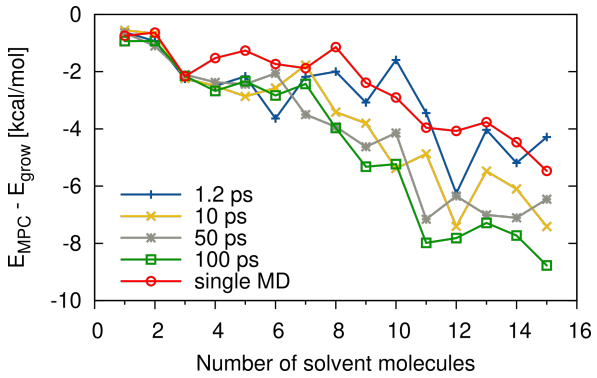

Figure C.9: Difference in energy between the most populated cluster (MPC) found by the NCI-MTD algorithm and the input cluster from the growth step for differently sized clusters of ethanol in acetonitrile. Each value is averaged over 10 individual runs.

as a normal MD simulation of 10 ps. Increased simulation times further yield energetically lower clusters. Throughout this work, we chose the energy gain upon cluster formation as a measure of its quality. Figure C.9 shows the energy difference between the clusters after the growth process and the most populated clusters (MPC) after the conformational sampling. Except for some fluctuation, longer MTD times yield, in general, larger energy gains ($E_{\text{MPC}} - E_{\text{grow}}$). Thus, the MTDs in the NCI-MTD run should be chosen as long as possible to yield better ensembles and to reduce the statistical error. However, for larger cluster sizes, long MTD simulations become prohibitively expensive.

C.5.2 Cluster Quality

In this section, we want to assess the quality of the generated clusters with QCG, *i.e.*, the physically meaningfulness of the found binding motifs, including solute-solute and solvent–solvent interactions. For quantification, the energy of cluster formation E_{form} is computed according to

$$E_{\text{form}} = E_{\text{cluster}} - E_{\text{solute}} - n \cdot E_{\text{solvent}}, \tag{C.15}$$

where the energy of the solute E_{solute} and n times the energy of a solvent molecule E_{solvent} are subtracted from the energy of the cluster E_{cluster}. The lower (more negative) the energy of formation, the higher is the quality of the assessed cluster.

The performance of QCG in terms of cluster generation (`-grow`) is compared to different algorithms. This intention turned out to be rather difficult as, to the best of the authors' knowledge, hardly any other algorithm exists that automatically generates solute–solvent clusters. One program package is

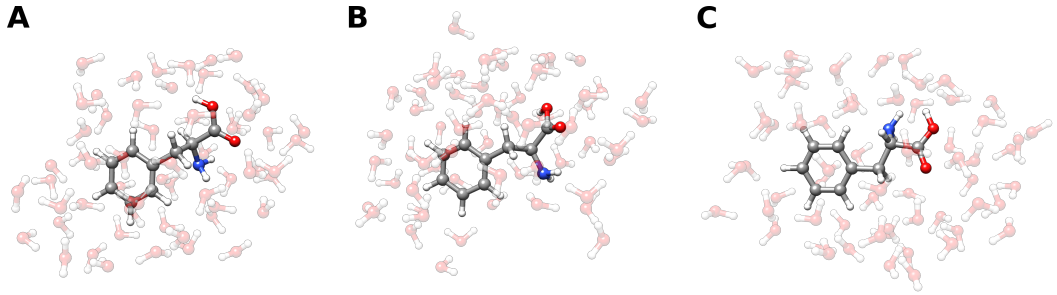

Figure C.10: Phenylalanine surrounded by 60 water molecules generated by QCG (A), the TIP3P water model (B), and space-filling algorithm (C) with subsequent GFN2-xTB geometry optimization.

Table C.2: Energy of formation (in kcal mol^{-1}) for the cluster out of phenylalanine and 60 water molecules computed with GFN2-xTB and r^{2}SCAN-3c single-point DFT computations. The QCG algorithm is compared to *AMBERtools* and the space-filling algorithm, both with subsequent GFN2-xTB geometry optimizations.

	QCG (*-grow*)	*AMBERtools* (TIP3P)	Space-Filling
GFN2-xTB	−712.4	−684.1	−665.3
r^{2}SCAN-3c//GFN2-xTB	−656.7	−632.1	−622.8

AMBERtools.[358] Therein, different solvent models can be chosen, but they are restricted to a few solvents. One of these is water, e.g., employing the TIP3P model.[191] During the cluster generation with *AMBERtools*, a sphere with a user-defined radius is cut out of a pre-equilibrated box of 216 water molecules. Subsequently, the solute is placed inside this cavity, and solvent molecules that collide are removed. We also implemented a second competitor that relies only on geometrical criteria. Therein, solvent molecules are randomly placed around the solute within a sphere of a given radius until the entire volume is filled. To ensure a reliable comparison, the structures resulting from *AMBERtools* and the space-filling algorithm were post-optimized with GFN2-xTB as in the QCG algorithm.

As an example, phenylalanine was solvated with 60 water molecules. Figure C.10 depicts the structures resulting from the QCG algorithm and *AMBERtools* after geometry optimization with GFN2-xTB. Table C.2 shows the corresponding formation energies for the three methods. The space-filling algorithm shows the highest formation energies and hence performs worst. Tentatively, this can be assigned to the neglect of intermolecular interactions during cluster generation. Even though subsequent GFN2-xTB geometry optimizations improve the structure, the starting point is still too far off from a global minimum to be repaired in just a single geometry optimization. The *AMBERtools* procedure yields a lower (better) formation energy than the space-filling model. The priorly equilibrated water box ensures a more physical solvent structure and for bulk water, the TIP3P water model delivers a good description. For a molecular cluster, this description may not be optimal, as explicit solute–solvent interactions during the cluster generation are neglected. This may lead to an unreasonable solute surrounding, which can only partially be compensated by the GFN2-xTB geometry optimization. The QCG algorithm optimizes the solute–solvent NCIs during each step of the iterative cluster growth and hence, yields the best cluster formation energies. At the r^{2}SCAN-3c//GFN2-xTB level, the energy gain is 24.6 kcal mol^{-1} (28.3 kcal mol^{-1} for GFN2-xTB) larger than with the TIP3P model, and 33,9 kcal mol^{-1} (47.1 kcal mol^{-1}) larger than with the space-filling approach. The

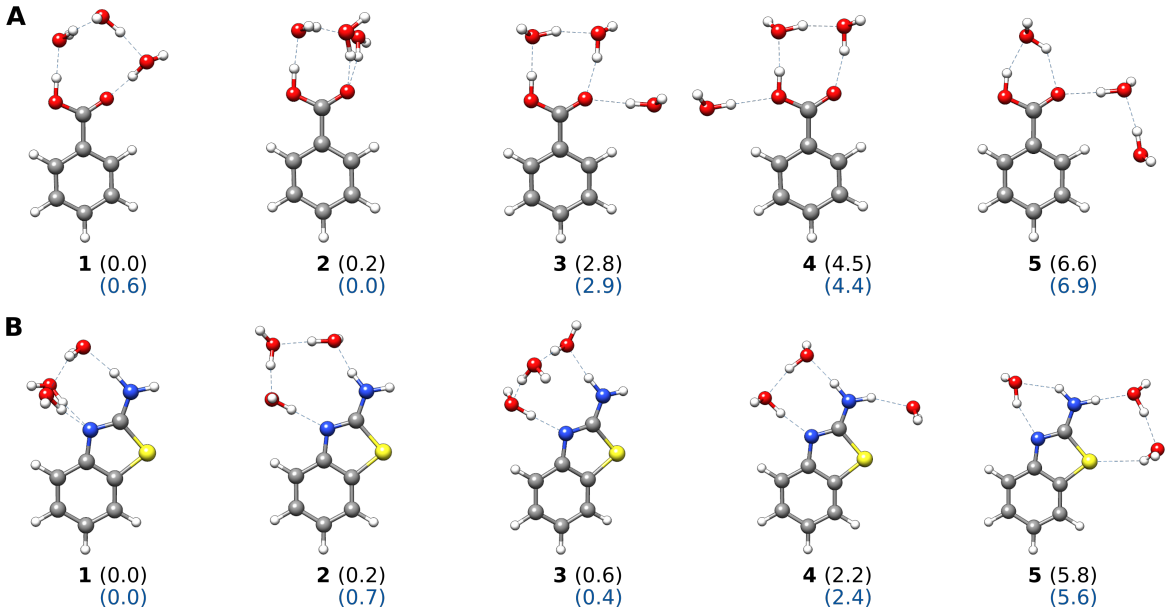

Figure C.11: Microsolvated structure ensemble of benzoic acid (A) and aminobenzothiazole (B) with three explicit water molecules, respectively. Shown are five different conformations each. Relative conformational energies (in kcal mol^{-1}) were calculated by r^{2}SCAN-3c and GFN2-xTB (in blue) and are given in parentheses.

generation of an initial energetically low cluster is mandatory for the subsequent ensemble generation (conformer sampling) in QCG. Due to the enormous complexity of the phase space and the limitation to finite simulation lengths, the outcome of the NCI-MTD algorithm strongly depends on the quality of the input cluster. Hence, the elaborated growing routine is indispensable for QCG. The computational cost of the cluster growth (16 min on four cores of an Intel® Xeon® CPU E3-1270 v5 @ 3.60 GHz) is much lower than that of the subsequent ensemble generation (2 h 4 min with the same CPU). In passing, we note the good agreement between the total cluster formation energies of the very reliable r^{2}SCAN-3c DFT composite and semi-empirical GFN2-xTB methods (deviation of about 10%).

C.5.3 Microsolvation

For small cluster sizes, the QCG algorithm shows small statistical errors (*cf.* section C.5.1). Hence, it seems to be a promising and easy-to-use tool for applications in the context of microsolvation. For demonstration, ensembles of benzoic acid and 2-amino-benzothiazole were generated with QCG (`-ensemble`) at the GFN2-xTB level of theory by adding three explicit water molecules to the solutes. Distinct conformations of the found clusters were energetically sorted by r^{2}SCAN-3c single-point calculations. Selected structures are shown in Figure C.11. Relative r^{2}SCAN-3c and GFN2-xTB (in blue) energies are given, respectively. In the case of benzoic acid (Figure C.11(A)), the most favorable structures at the r^{2}SCAN-3c level are cyclic conformations, including three water molecules and the carboxylic acid functional group. The energetic order of the conformers for (A1) and (A2) changes at the GFN2-xTB level. Structures (A3) and (A4) form a cyclic arrangement between two water molecules and the carboxylic acid group. The energy difference between these structures of 1.7 kcal mol^{-1} (1.5 kcal mol^{-1}) occurs due to the different position of the third water molecule,

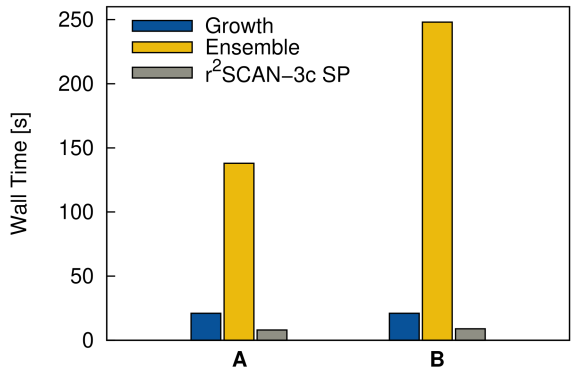

Figure C.12: Computational timings of benzoic acid (A) and aminobenzothiazole (B) for the cluster growth, the ensemble generation, and the r^{2}SCAN-3c single-point calculations with three explicit water molecules. The calculations were done on four cores of an Intel® Xeon® CPU E3-1270 v5 @ 3.60 GHz.

forming either an intermolecular hydrogen bond to the C=O or the O-H group of the benzoic acid molecule. In the least favorite conformation (A5), only two water molecules are interacting with the COOH group. For the structures (A3)–(A5), the energetic order of GFN2-xTB and r^2SCAN-3c agree. For 2-amino-benzothiazol (Figure C.11(B)), the lowest energies are observed for the structures with intramolecular hydrogen bonds (HBs) between the water molecules and the amine fragments. Again, cyclic arrangements of the amine group with three water molecules are favored over those with two or one water molecule, respectively. The energetically higher clusters also contain intramolecular HBs to the sulfide group, which are less strong. For small energy differences between the conformers ($\leq$ 1 kcal mol^{-1}), a partially different ordering is observed for GFN2-xTB compared to r^{2}SCAN-3c. Thus, in the context of microsolvation, we recommend a re-ranking by DFT as already suggested in Ref. 35 for non-rigid molecules, since noncovalently bound clusters may also be regarded as highly flexible.

To further validate the QCG ensembles, they were compared to results from a different microsolvation approach proposed in Ref. 193. Here, the ensembles for the same systems were obtained from MD simulations with subsequent grid inhomogeneous solvation theory (GIST) analysis.[359] The resulting structures after B3LYP[201,360]-D3[123]/def2-TZVP[361] optimization are similar to the QCG ones and have the same interaction motifs. For example, the GIST analysis found a structure similar to (A3), and also a cluster with the same water docking positions as (B4) was observed. Even though the energy ranking depends on the chosen density functional, the inclusion of an implicit solvation model, as well as on the respective methods employed for the thermostatistical contributions, both approaches (GIST analysis and QCG) yield similar orderings. In terms of computation time, QCG, with subsequent r^{2}SCAN-3c single-point calculations, takes only a few minutes on a regular desktop computer (Figure C.12). The QCG approach combined with r^{2}SCAN-3c single-point calculation (3 min 7 sec and 4 min 39 sec) outperforms other approaches solely based on MD simulations[193,362] in terms of computation time. This computational efficiency, coupled with high accuracy and robustness, makes QCG a promising candidate for future applications in calculating free energies of reactions based on microsolvated structures.

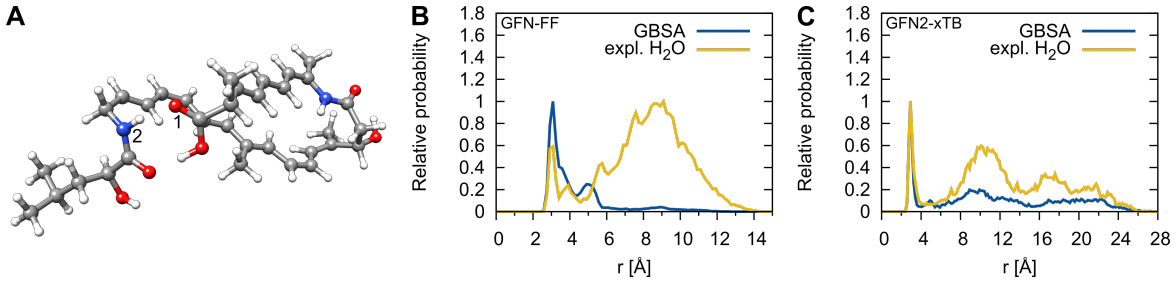

Figure C.13: (A) Gas-phase optimized structure of the energetically lowest bacillaene conformer found at the GFN2-xTB level of theory. Distance distribution functions of atoms O1 and N2 obtained from GFN-FF (B) and GFN2-xTB (C) MD simulations employing the implicit GBSA solvation model, and from an GFN-FF and GFN2-xTB MD simulation of the QCG cluster containing 100 water molecules.

C.5.4 Molecular Dynamics

The broad field of QCG applications also includes the impact of explicit solvation effects on geometries during MD simulations. Depending on the solvation model, qualitatively different geometries might result in solution and change the course and outcome of MD simulations. This is illustrated here in MD simulations the natural antibiotic bacillaene[194] applying two different solvent models. First, 100 explicit water molecules were added with QCG to the gas-phase optimized geometry and 1 ns GFN-FF and GFN2-xTB MD simulations at 298.15 K were performed under the application of an external wall potential. For comparison, MD simulations with equivalent settings were run using the implicit GBSA model. As a measure for the here relevant folding process, radial distribution functions (RDFs) of the O1-N2 interatomic distance (Figure C.13(A)) were calculated from GFN-FF and GFN2-xTB trajectories. These RDFs show the frequency of occurrence for the intramolecular hydrogen bonds that are mainly responsible for the ring-shaped geometry in the gas phase. The RDFs computed with GFN-FF (Figure C.13(B)) and GFN2-xTB (Figure C.13(C)) show similar trends. Distances below 3 Å indicate that two intramolecular hydrogen bonds are present, meaning the structure is similar to the gas phase. The second peak shortly above 3 Å shows a small elongation of the O–N distance, which is attributed to the cleavage of the H-bond between O1 and N2 under preservation of the second intramolecular hydrogen bond. Other peaks occurring up to 6 Å show the formation of intramolecular hydrogen bonds that were not present at the beginning, indicating a significant change in the structure. Distances of more than 6 Å indicate the dissociation and the separation of the two ends of the bacillaene chain. In general, GFN2-xTB yields a more elongated conformation compared to GFN-FF, reflected in a non-vanishing RDF beyond 14 Å. The MD results with GBSA reveal that in both simulations the intramolecular hydrogen bonds are predominantly formed yielding a closed structure. In the GFN2-xTB/GBSA MD-simulation, a noteworthy occurrence of structures with no intramolecular HBs is obtained, which is strongly amplified by explicit water molecules. The same picture holds for the GFN-FF. Hence, we conclude that the explicitly modeled solvent molecules lead mainly to an open structure of bacillaene and favor intermolecular over intramolecular HBs. Similar observations were reported in a recent study of methyl lactate in explicit solvent clusters.[363]

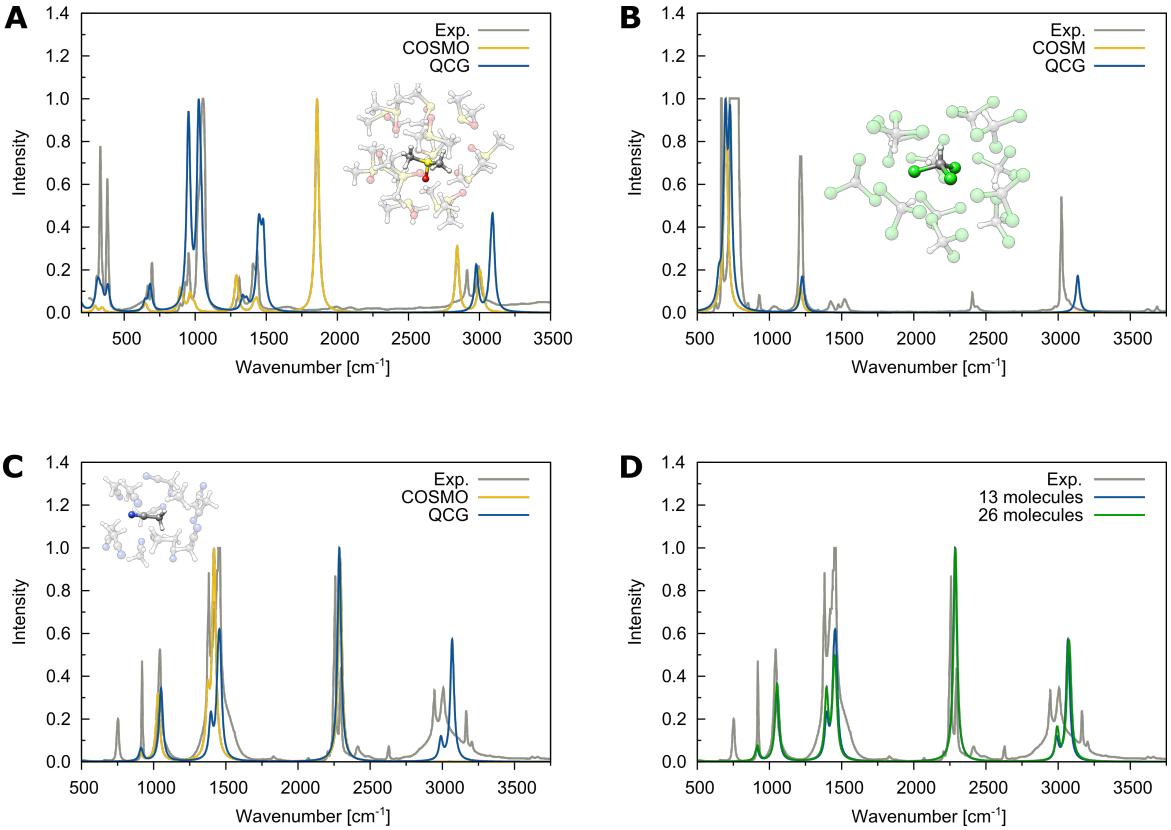

Figure C.14: IR-spectra of liquid DMSO (A), $CHCl_3$ (B), and CH_3CN (C) computed at the B3LYP-3c level of theory from explicit QCG clusters and with the COSMO model in comparison to experimental data. Different QCG cluster sizes are investigated in (D) for CH_3CN.

C.5.5 IR-spectra

The QCG algorithm was already successfully applied for IR spectra calculations in solution.[363,364] Therein, it was demonstrated that solvation effects play an important role in liquid-phase IR spectra. Continuing on this work, the QCG approach was applied for IR spectra calculations of organic liquids in comparison to experimental data. The COSMO model was additionally tested as an alternative. For the explicit description with QCG, clusters consisting of one solvent shell were generated, and the energetically lowest structure from the ensemble search was then re-optimized at the B3LYP-3c level of theory. To minimize finite cluster size effects, the vibrations of the central solute molecule were separated from the vibrations of the surrounding molecules by increasing their atomic mass to shift the vibrational frequencies to the low-frequency region. For the COSMO approach, input geometries for the calculation of liquid phase IR spectra were generated by optimizing gas-phase geometries on the B3LYP-3c level of theory employing the COSMO model. The results for different organic liquids are shown in Figure C.14. Taking into account only the most populated cluster is a good approximation. In agreement with Refs. 363,365, only the most populated cluster was taken into account for IR spectra calculation, as the effect of considering the entire SE is rather small. The IR spectra computed with the COSMO model show peaks that are not present in the experimental

Table C.3: Spectral matchscore (in %, identical spectra yield 100%) between the computed IR-spectra at the B3LYP-3c level employing the QCG and COSMO approaches and the experimental spectra for liquid DMSO, $CHCl_3$, and CH_3CN.

COSMO	Molecule	QCG
56.7	DMSO	68.7
42.2	$CHCl_3$	67.0
61.2	CH_3CN	60.8

or the QCG cluster spectrum, e.g., in the case of liquid DMSO (Figure C.14(A)) a peak of highest intensity occurring at ~1850 cm^{-1}. Moreover, the C-Cl (Figure C.14(B)) and C-H (Figure C.14(C)) stretching vibration signals are incorrectly almost not visible in the case of chloroform and acetone. In contrast, the QCG cluster ansatz yields peaks close to the experimental ones regarding position and intensity. Furthermore, changes in symmetry due to the solvent surrounding cannot be seen with COSMO but are captured by QCG. For example, the asymmetric C-Cl stretching vibrations at around 750 cm^{-1} (Figure C.14(B)) are different. For COSMO, the system has high C_{3v} symmetry leading to fewer signals than in the QCG cluster calculation, where each chlorine atom is slightly different. This leads to a loss of symmetry and multiple peaks for the C-Cl stretching and bending vibrations in better agreement with the experimental IR spectrum. Adding more explicit solvent molecules than one solvent shell has a minor influence, which is shown exemplarily for liquid acetonitrile (Figure C.14(D)). The frequencies resulting from the differently sized clusters are mostly identical, and the intensities differ only slightly, which is consistent with previous findings in Ref. 364. The overall good performance of the QCG approach is quantified by the respective spectral matchscore (for definition see Ref.[352]) reported in Table C.3. For DMSO and $CHCl_3$, the IR spectra computed for QCG clusters yield an improvement by 12.0 % and 24.8 %, respectively, over the COSMO approach. For CH_3CN, the two different solvation models show similar matchscores because COSMO performs better at around ~1450 cm^{-1} while QCG provides the more realistic spectrum for the C-H stretching vibrations at ~3000 cm^{-1}. Similar differences in IR spectra between implicit and explicit solvent models were also found in another study,[366] where the placement of solvent molecules was done manually. The automated procedure of QCG allows for the efficient study of different clusters of various sizes.

C.5.6 Solvation Free Energies

With QCG, it is possible to compute solvation free energies in a supermolecular fashion as described in Section C.2.2. With the here discussed first proof-of-principle examples, we want to evaluate the performance of QCG in comparison to the (in our opinion) most accurate implicit model available (COSMO-RS[192]) and experimental values. Therefore, δG_{solv} values were computed for 45 different systems at 298 K, combining polar and apolar organic molecules of different sizes. These are part of the SMD fitset[145] and are assumed to be also part of other fitsets for implicit solvation models. Each QCG calculation was performed ten times, and the arithmetic mean was taken. The ensembles were generated with the NCI-MTD run-type employing the default settings (e.g., an MTD time of 10 ps). For every solute, 25 solvent molecules were modeled explicitly, ensuring that at least the first solvation

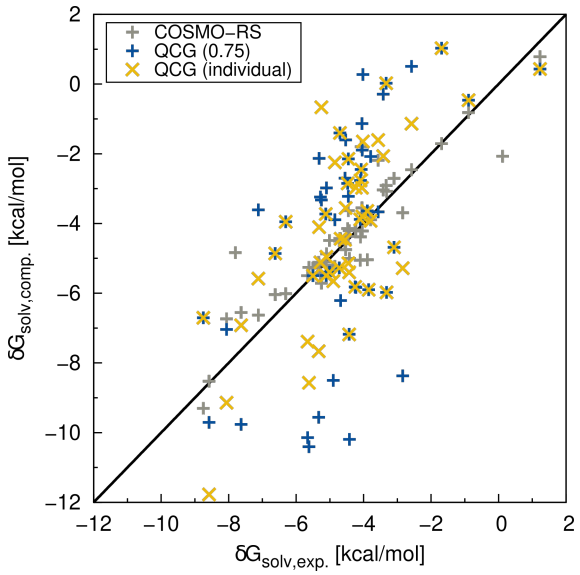

Figure C.15: Correlation plot of δG_{solv} values of 33 small organic molecules computed with QCG and COSMO-RS in comparison to experimental values. QCG values are averaged over 10 individual runs. They are given for a global scaling factor of the translational and rotational entropy of 0.75 and for an empirically adjusted solvent specific one.

shell is completely covered. For water as a solvent, the number was increased to 40 molecules to ensure a complete first solvent shell. Figure C.15 shows the correlation between QCG/COSMO-RS δG_{solv} values and the respective experimental results.[145] The corresponding statistical errors are given in Table C.5. COSMO-RS performs very well with an MD of 0.21 kcal mol^{-1} and an MAD of 0.49 kcal mol^{-1} , respectively. The solvation free energies computed with QCG have an MD of −0.07 kcal mol^{-1} and an MAD of 2.34 kcal mol^{-1} . Referring to the analysis of the statistical errors (*cf.* section C.5.1) it is clear that observed scattering from the growth and ensemble steps is also inherent in the computed δG_{solv} values (see SI for further details). Taking this into account, and regarding the fact that QCG subtracts huge energies in a ”brute force” approach, the reasonable agreement with the experiment seems encouraging. Even though the accuracy of COSMO-RS can not be reached, the already reasonable δG_{solv} values computed with QCG represent a promising starting point for further improvements, especially because of the universal applicability to arbitrary solute–solvent combinations. Currently, the following sources of error can be addressed. First, in contrast to other hybrid cluster continuum models, QCG generates two different sets of solute and solvent clusters and thus, the errors within the electronic energies and thermostatistical contributions occur twice. Also, the GFN2-xTB error for the interaction energies does not cancel out completely between the solute–solvent and the reference–solvent ensemble. Second, the ensemble might not be of similar quality for every system, because an MTD time of 10 ps was chosen as a compromise between computational costs and accuracy. Longer simulation times should be tested comprehensively in the future. Third, using only solute–solvent clusters that are at least populated by 10 % for the calculation of the solvation free energy leads to an error of the conformational entropy and thus the conformational free energy G_{conf}. For example, a small cluster of three chloroform molecules has approximately a G_{conf} value of −5.23 kcal mol^{-1} at 298.15 K, while taking only clusters that are populated by

Table C.4: Adjusted scaling factors for the rotational and translational entropy contributions of different solvents.

Solvent	Scaling factors
Benzene	0.65
Water	0.75
CH_3CN	0.85
DMSO	0.90

Table C.5: Statistical measures (in kcal mol^{-1}) for the δG_{solv} values of small organic molecules computed with QCG/NCI-MTD and COSMO-RS in comparison to experimental values. QCG values are given for a scaling factor of the translational and rotational entropy of 0.75, and for an empirically adjusted solvent-specific one.

	QCG(0.75)	QCG(individual)	COSMO-RS
MD	−0.07	−0.02	0.21
MAD	2.34	1.57	0.49
RMSD	2.86	2.12	0.75
SD	2.89	2.14	0.73

at least 10 % yield a lower G_{conf} of −1.33 kcal mol^{-1}. However, partial error compensation of solute–solvent and solvent–solvent ensembles is expected. Including more solute–solvent clusters would lead to a much higher computational expense. Fourth, the rather simple GBSA model applied for electrostatic bulk screening is computationally efficient but might introduce further errors. A future implicit solvent model that describes the electrostatic response of a solvent and accounts for polarization should improve the description of the bulk. Lastly, the scaling factor for the translational and rotational entropy contributions was empirically determined to be 0.75 as an average over many different solvents. An improvement can be obtained by adjusting it for different solvents (Table C.4), leading to a 0.77 kcal mol^{-1} lower MAD (Table C.5). Various technical aspects can be improved, thereby increasing the computational cost, e.g., increasing the MTD times, or including also less populated solute–solvent clusters. Additionally, $\Delta\delta G_{\mathrm{solv}}$ for reactions may be easier to compute than absolute δG_{solv} due to error cancellation.

C.6 Conclusion

We developed and tested an automated and broadly applicable model for a QM description of explicit solvation. This procedure, termed QCG, is based on the GFN-FF and GFN2-xTB methods in combination with xTB-IFF, giving access to fast geometry optimizations, MD simulations, and docking steps. The conformational space exploration of the generated molecular clusters is conducted by the NCI-MTD algorithm of the CREST program. This enables systematic improvability of the QCG approach by extending the simulation time within the conformer search for a smaller statistical error and structures with lower energy in the ensemble. The QCG algorithm includes only very few empirical parameters (wall potential, translational/rotational scaling), as most of the required parameterization for efficient treatment is inherent in the underlying QM/FF methods. The presented

approach is unique in regard to the fully automated cluster growth and ensemble generation of arbitrary solute–solvent combinations. We tested the QCG approach on a large variety of chemical systems, ranging from small organic molecules to large anti-cancer drugs in eutectic solvents. We individually analyzed the different underlying steps (`-grow`, `-ensemble`, `-gsolv`) in the workflow and found that the reproducibility in terms of molecular structures and energies is good for small cluster sizes. Increasing their size leads to notable deviations in terms of structures and energies between the same calculations performed multiple times. The incomplete sampling of a cluster's phase space was determined as the main source of error. Nevertheless, the increase of simulation times during the conformer sampling (ensemble generation) reduced the statistical energy error significantly. We showed that QCG can be straightforwardly applied in microsolvation studies. Moreover, since many computed properties may benefit from including explicit solvent molecules, and QCG offers a simple and automated way to generate these structures that can be used, e.g., for simulating geometries and IR spectra in solution. Here, significant improvements were observed compared to the COSMO model. QCG further represents a physically reasonable procedure for the calculation of solvation free energies by including all terms in δG_{solv} explicitly. Even though the accuracy of the established, highly parameterized implicit COSMO-RS model was not reached, reasonable results coupled with universal applicability are promising for future improvements. In conclusion, the new, freely available QCG tool can help to investigate and understand solvation effects on a molecular level. Due to its computational demand, QCG is not meant to replace existing, efficient continuum models. QCG establishes an alternative solvation tool that is capable of obtaining reasonably accurate results for complex molecular systems where implicit methods reach their limits. The universal applicability to arbitrary solute–solvent combinations is a unique feature that is yet missing in the portfolio of solvation tools, and we hope that QCG will be useful for computational chemistry.

Supporting Information

The Supporting Information is available free of charge at
https://pubs.acs.org/doi/10.1021/acs.jctc.2c00239.

- Computational results, availability, and statistical error measures (PDF)
- All relevant structures in xyz format (ZIP)

Conflict of Interest

There are no conflicts to declare.

Acknowledgements

This work was supported by the DFG in the framework of the "Gottfried-Wilhelm-Leibniz" prize to S.G. S.S. thanks the "Fond der chemischen Industrie (FCI)" for financial support. The authors thank F. Bohle, M. Bursch, S. Ehlert, D. Menche, and M. Reuter-Schniete for helpful input and discussions.

APPENDIX D

Conformational energies of biomolecules in solution: Extending the MPCONF196 benchmark with explicit water molecules

Christoph Plett,* Stefan Grimme,* Andreas Hansen*

Reprinted (adapted) with permission‡ from:
C. Plett, S. Grimme, and A. Hansen, J. Comput. Chem. **45** (2024) 419.

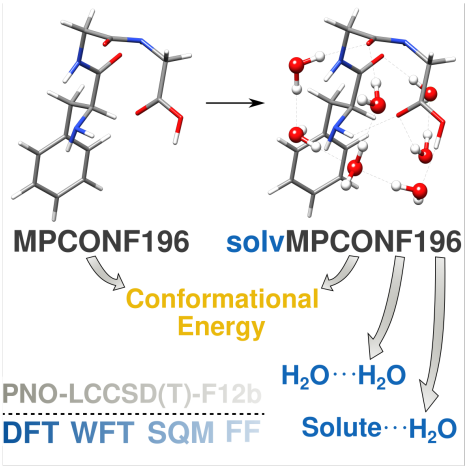

Figure D.1: Associated Table of Contents graphic for publication in Journal of Computational Chemistry.

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

Abstract

A prerequisite for the computational prediction of molecular properties like conformational energies of biomolecules is a reliable, robust, and computationally affordable method, usually selected according to its performance for relevant benchmark sets. However, most of these sets comprise molecules in the gas phase and do not cover interactions with a solvent, even though biomolecules typically occur in aqueous solution. To address this issue, we introduce a with explicit water molecules solvated version of a gas-phase benchmark set containing 196 conformers of 13 peptides and other relevant macrocycles, namely MPCONF196 [J. Řezáč *et al.*, *JCTC* **2018**, *14*, 1254–1266.] and provide very accurate PNO-LCCSD(T)-F12b/AVQZ' reference values. The novel solvMPCONF196 benchmark set features two additional challenges beyond the description of conformers in the gas phase: conformer–water and water–water interactions. The overall best performing method for this set is the double hybrid revDSDPBEP86-D4/def2-QZVPP, yielding conformational energies of almost coupled cluster quality. Furthermore, some (meta-)GGAs and hybrid functionals like B97M-V and ωB97M-D with a large basis set reproduce the coupled cluster reference with an MAD below 1 kcal mol^{-1} . If more efficient methods are required, the composite DFT-method r^{2}SCAN-3c (MAD of 1.2 kcal mol^{-1}) is a good alternative, and when conformational energies of polypeptides or macrocycles with more than 500-1000 atoms are in focus, the semi-empirical GFN2-xTB or the MMFF94 force field (for very large systems) are recommended.

D.1 Introduction

Solute-solvent interactions play a crucial role for the properties of molecules in solution such as the reactivity toward certain reaction pathways, the association of molecules, and conformational energies.[61,62,367,368] Besides industrial and laboratory processes like stereoselective reactions in water[292] or various catalytic reactions,[369,370], the solvent effect for biomolecules is of special interest as it can influence, for example, the folding and dynamics of proteins[294,295] and their binding to ligands.[371,372] Moreover, the conformations of a protein in solution can deviate from the gas-phase ones and can have different properties.[373,374] Thereby, the interaction with water molecules is especially important since they typically occur in the aqueous phase, which is also of major importance for living.[375] To understand such biomolecular processes, computational studies became a valuable tool that have already been applied to enlighten, for example, protein folding, protein-drug interactions, ligand transport phenomena, and ion channel activity.[376–378] Thereby, many different methods can be chosen from the classes of Wave Function Theory (WFT), Density Functional Theory (DFT), semi-empirical quantum mechanical (SQM) methods, and Force Fields (FF) that provide different cost–accuracy ratios limiting, for example, the application of highly accurate methods such as coupled cluster with perturbative triples (CCSD(T)) to relatively small molecular structures.[36] To find suitable methods for certain applications and to quantify errors, benchmark sets composed of realistic test cases are used. Prominent examples are the GMTKN55,[71] the NCI Atlas,[77–80,379] and the MGCDB84[70] benchmark set. However, besides a few benchmark studies that focus on water–water interactions,[81,380] most of these benchmark sets consider isolated gas-phase molecules and neglect solvent effects that have to be additionally modeled for systems in solution.[97] This can be done with computationally efficient implicit solvent models, but they may be too inaccurate for, e.g., the prediction of conformational energies, as they approximate the solvent only as a continuous

potential.[153,381] The inclusion of explicit solvent molecules in either a full explicit treatment or microsolvation approach can increase the accuracy, but they also increase the computational costs significantly.[68,157,382,383]

In this study, we introduce a benchmark set that allows the evaluation of computational, atomistic methods for biomolecules, especially proteins, in aqueous solution. Therefore, we extend the MPCONF196 gas-phase benchmark set[76,384] composed of conformers of different peptides and macromolecules by adding explicit water molecules. Due to recent advances and insights into basis set convergence and local coupled cluster approximations,[385] we can provide very accurate conformational reference energies for this newly compiled solvMPCONF196 and the original MPCONF196 benchmark set, where large molecules were treated with more approximate local coupled cluster methods. In the following, we first describe the generation of the solvMPCONF196 structures and the methods used for the calculation of the conformational energies. Thereafter, the solvMPCONF196 is presented, and the included interactions are discussed. Finally, an assessment of common DFT, WFT, SQM, and FF methods is performed, and some general conclusions are drawn.

D.2 Methodology

D.2.1 Geometries

The gas-phase geometries were taken from Ref. 76. To obtain the explicitly solvated structures, the directed docking feature of the automated Interaction Site Screening (aISS)[386] algorithm implemented in *xtb* version 6.6.0[84] was employed via the Quantum Cluster Growth (QCG) algorithm[289] implemented in a modified *CREST*[44] version 2.12 to saturate different interaction sites of each conformer. Thereby, the original conformer gas-phase geometry was fixed and the added solvent molecules were subsequently optimized with r^2SCAN-3c[129] implemented in ORCA 5.0.4[387] employing tight convergence thresholds and the implicit CPCM water solvation model.[154,388] Additionally, full r^2SCAN-3c optimizations were performed to investigate the effect of solute relaxation. Further details are given in Sec D.3.1.

D.2.2 Conformational energies

The reliable calculation of the usually rather small energy differences between different conformers of a molecule requires accurate treatment of both (mostly) covalent short-range interactions and medium- and long-range intramolecular noncovalent interactions (NCIs) such as H-bonds and London dispersion. For the bio-organic uncharged closed-shell molecules studied here, the CCSD(T) method yields highly accurate conformational energies, provided that the basis set used is large enough. In the case of the structures explicitly solvated with water (solvMPCONF196), there are additional intermolecular H-bonds, whose accurate description also requires large basis sets with diffuse functions. In Ref. 385, it was shown that for safely converged relative energies, a complete basis set (CBS) extrapolation with aug-cc-pVQZ/aug-cc-pV5Z would be necessary and that CBS extrapolations with smaller basis sets can lead to erroneous extrapolated limits. However, the CCSD(T) method with its $O(N^7)$ scaling behavior of computational time with system size N is computationally feasible only for the smaller subset of MPCONF196 and then only with moderately large basis sets based on which CBS extrapolated reference values were generated for this subset in the original publication.[76] For the larger subset (Figure D.2), Řezáč *et al.* applied the same CBS extrapolation, but using

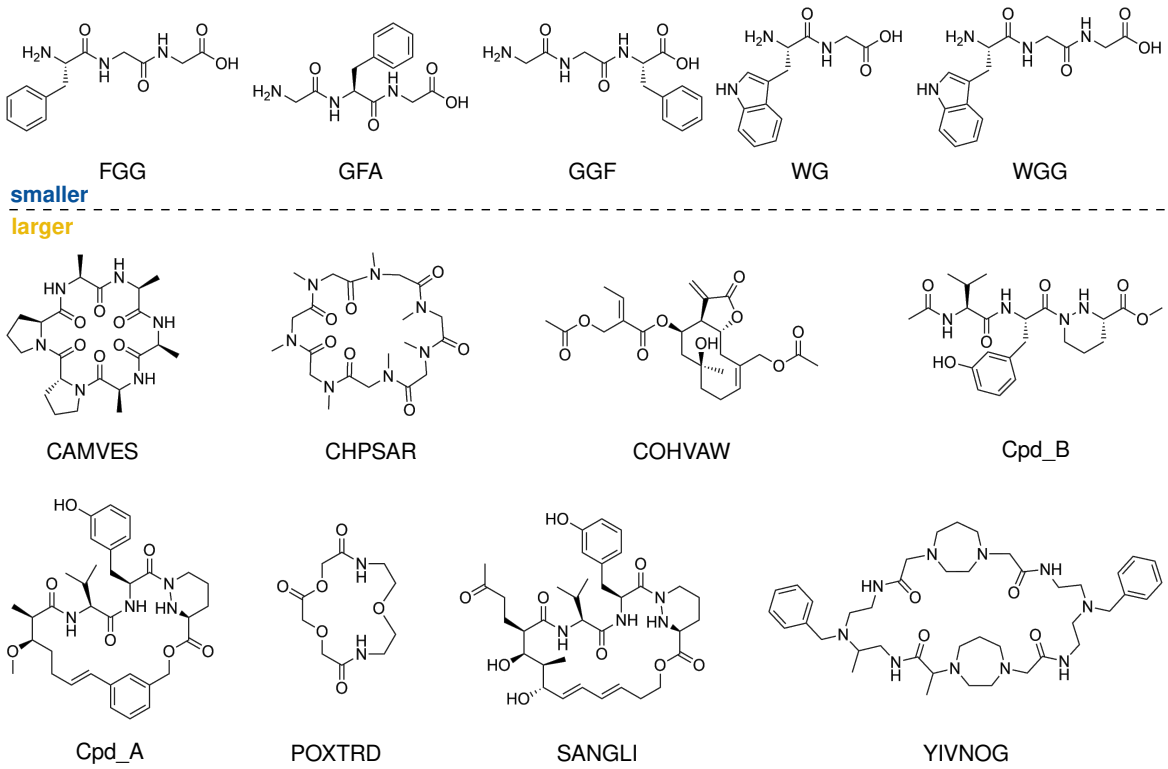

Figure D.2: Lewis structure of the molecules contained in the MPCONF196 and solvMPCONF196 divided according to Ref. 76 in a smaller and larger subgroup combining the original medium and large subgroups.

DLPNO-CCSD(T),[389,390] a much more efficient local implementation of CCSD(T), to obtain the energies. However, CBS extrapolation with local correlation methods is even more error-prone than with canonical CCSD(T), at least if the employed basis sets are not already very large and the local threshold settings were not chosen sufficiently tightly.[385,391] Additionally, because the original reference values may also suffer from these errors to some extent, and only semi-canonical triples were used, we decided to generate reference values not only for the newly introduced solvMPCONF196 set but also to revise those of the original gas-phase MPCONF196 set. Following the findings of Ref. 385, we used for this purpose a state-of-the-art local CCSD(T) implementation with explicit correlation and tight threshold settings (PNO-LCCSD(T)-F12b[195–197] with `tight` domain settings as implemented in Molpro (2022.3 release))[392,393] together with a modified aug-cc-pVQZ basis (cc-pVQZ for H to reduce the residual basis set superposition error (BSSE); hereafter abbreviated as AVQZ'). The explicit correlation allows to avoid a CBS extrapolation and still yields energies very close to the CBS limit already with the AVQZ' basis set. The use of iterative triples and tight domain settings ensures that the NCIs are also described very accurately and that the BSSE is negligibly small even for the intermolecular H-bonds in the solvated structures, for which we used the identical setup to calculate reference values.

However, this reference protocol is computationally too demanding for the systems of the larger subset, for which we used a slightly less accurate but much more efficient setup (PNO-LCCSD(T)-F12b/AVTZ' with `default` domain settings) and minimized the additional error *a posteriori* by scaling the triples

contributions (see section S2 in the SI for details).[385] The analogous values for the smaller subset are virtually identical to the corresponding PNO-LCCSD(T)-F12b/AVQZ'/tight results with a mean unsigned error of 0.03 kcal mol^{-1} and 0.05 kcal mol^{-1} for the MPCONF196 and solvMPCONF196 set, respectively, and maximum unsigned error of 0.10 kcal mol^{-1} and 0.19 kcal mol^{-1} (see SI, Table S1), so that comparably accurate reference values could also be generated for the larger subset. The residual errors are slightly larger for the solvated structures due to the larger BSSE, but based on the benchmark results in Ref. 385 and the very good agreement of the AVTZ' values with the highly accurate setup on the smaller subset, we estimate the residual error of the new reference values to be about 0.1–0.3 kcal mol^{-1} for the MPCONF196 and 0.2–0.5 kcal mol^{-1} for the solvMPCONF196 benchmark set, which corresponds to a statistical discriminability of 0.04 kcal mol^{-1} and 0.05 kcal mol^{-1}, respectively, of the mean errors of more approximate methods to be tested on these benchmark sets.

The original reference values, however, differ more significantly from the newly generated MP-CONF196 reference conformational energies (RMSD = 0.37 kcal mol^{-1}, AMAX = 1.54 kcal mol^{-1}; see SI, Table S2), comparable to the error statistics of the best tested double hybrid functional (see SI, Table S3), which underlines the need for improved reference values for high accuracy DFT methods. With the newly generated reference values, an unbiased assessment of the best available density functionals and MP2-based methods should be possible for both gas-phase and solvated conformers.

For the subsequent benchmark study, DFT single-point energies were calculated with ORCA 5.0.4 using default settings. In the case of revDSD-PBEP86 and r^{2}SCAN0, the D4 correction[254] was not computed with ORCA due to missing parameters in version 5.0.4, but with the dftd4[125] standalone program. All assessed DFT and WFT methods, except for the composite schemes, were used with the def2-QZVPP basis set and matching auxiliary basis sets.[105,394] The GFN methods were applied as implemented in xtb 6.6.0[84] and the PM methods as implemented in *MOPAC2016* version 19.179L accessed via an xtb interface. DFTBD3 was employed using *DFTB+*[395] version 22.2 with the D3(BJ) dispersion correction.[123,124] All force fields except of GFN-FF were used with *Open Babel* 2.3.1.[396] All tested methods are listed in Table D.2.

D.3 Results and discussion

In Section D.3.1, the benchmark sets are discussed. An evaluation of the tested methods with respect to the reference values for the MPCONF196 and its solvated version solvMPCONF196 is given in Section D.3.2. Thereby, DFT/WFT methods in the large def2-QZVPP basis are evaluated first, followed by the more efficient composite-DFT methods, and finally SQM and FF methods.

D.3.1 Benchmark set and the effects of solvation

The original MPCONF196 contains 196 conformers of the 13 peptides and other macrocycles shown in Figure D.2. For clarity, the classification into "smaller" and "larger" conformers is made analogous to Ref. 76, combining the original "medium" and "larger" molecules according to our applied coupled cluster levels explained in Sec D.2.2. The structures exhibit typical organic functionalizations like amine, keto, alcohol, and carboxy groups that allow for intramolecular and intermolecular hydrogen bonding. For each molecule, 15 conformers (16 for GFA) were considered in Ref. 76 that show substantially different intramolecular hydrogen bonds and differ by up to 40 kcal mol^{-1} for the larger and more flexible structures like CAMVES, CHPSAR, and Cpd_B.

Table D.1: Number of conformers, added water molecules, and atoms per molecule of the systems composed in the solvMPCONF196 benchmark set.

Molecule	Water molecules	Number of atoms
FGG	7	58
GFA	6	58
GGF	6	55
WG	6	52
WGG	6	59
CAMVES	8	92
CHPSAR	9	97
COHVAW	8	86
Cpd_B	9	91
Cpd_A	8	116
POXTRD	9	55
SANGLI	9	125
YIVNOG	7	137

For the solvMPCONF196, we added water molecules to these structures in a way that as many functional groups as possible are solvated while keeping the number of water molecules small. This ensures that the water–water hydrogen bonding is reduced to the most important interactions and that diverse water–solute interactions remain an important part of the solvated benchmark structures. The number of water molecules used for the conformers of each molecule and the resulting total number of atoms are given in Table D.1. Typical solvated structures obtained upon water addition are characterized by hydrogen bonds between water as hydrogen bond donor and an oxygen atom of the conformer as acceptor (Figure D.3, A and C). Analogous interactions can be seen for amine

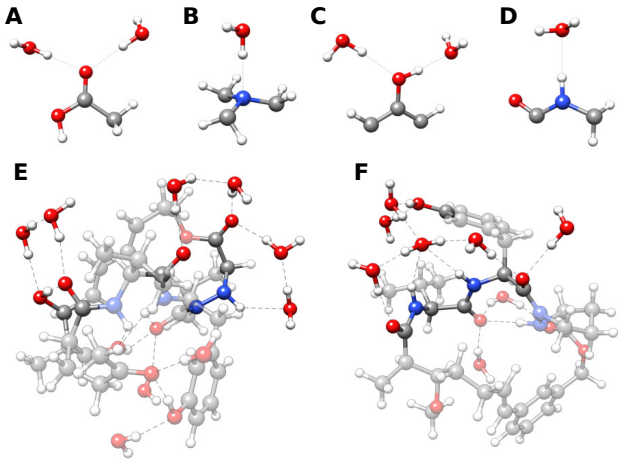

Figure D.3: Examples of bonding motifs of explicit water molecules for the structures composed in the solvMPCONF196 benchmark set. A-D are cutouts of typical H-bonded structures. E is SANGLI (conformer f). F is Cpd_A (conformer n).

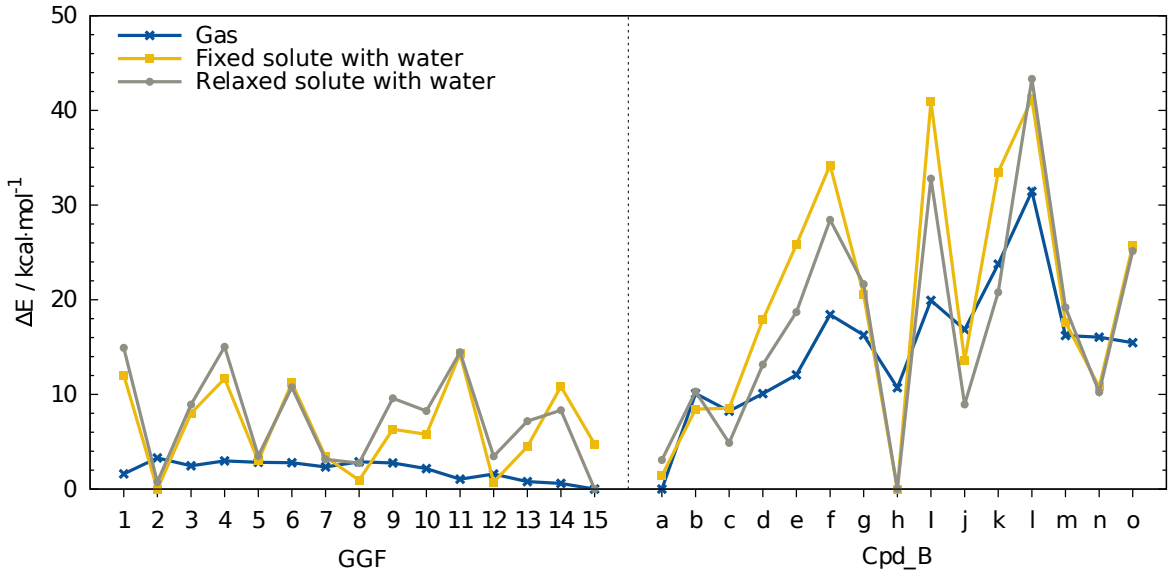

Figure D.4: r^{2}SCAN-3c conformational energies of the gas-phase and solvated conformers of GGF and Cpd_B. For the with water molecules solvated structures, the energies for the fully optimized structures and the structures with fixed gas-phase conformer geometries are given.

functional groups of the solute acting as a single hydrogen bond acceptor (Figure D.3, B). Moreover, the conformer can act as a hydrogen bond donor with H atoms bonded to nitrogen or oxygen (Figure D.3, C and D). Finally, various combinations and deviations of the described H-bond motifs occur if functional groups of the solute are spatially close (two examples are shown in Figure D.3, E and F).

To analyze the different conformational energy rankings for the MPCONF196 and the solvMPCONF196, r^{2}SCAN-3c conformational energies are exemplarily depicted for a smaller (GGF) and a larger (CAMVES) molecule in Figure D.4. The composite DFT method was chosen due to its good cost–accuracy ratio (see Sec D.3.2). Besides the conformational energies of the solvMPCONF196 containing water molecules and the solutes with the gas-phase (MPCONF196) geometries, also the energies after a full optimization including the solute are also depicted. It becomes evident that the smaller conformers have rather low energy differences of up to 4 kcal mol^{-1} in the gas phase, while the larger and more flexible conformers differ by up to 40 kcal mol^{-1}. After adding explicit water molecules, these energy differences increase strongly for the smaller conformers to more than 15 kcal mol^{-1} but only relatively moderately for the larger molecules. Furthermore, the additional water molecules often change the conformer ranking, and the energetically favored ones in the gas phase do not necessarily correspond to the low-energy conformers upon solvation. One reason for the differences is the H-bonds that are formed between water molecules and the solute. In the gas phase, the energy ranking is dominated by intramolecular hydrogen bonds. Energetically favored gas-phase conformers typically show the most intramolecular hydrogen bonds with strongly interacting functional groups that are saturated. Conformers with less strong intramolecular hydrogen bonds have higher gas-phase energy, and their functional groups can point outwards from the molecule. In such cases, the addition of water molecules leads to the formation of strong solute–solvent hydrogen bonds that lower the energy more than for a conformer with already saturated functional groups.
Relaxing also the solute geometry yields overall only small changes in the relative energies. For some

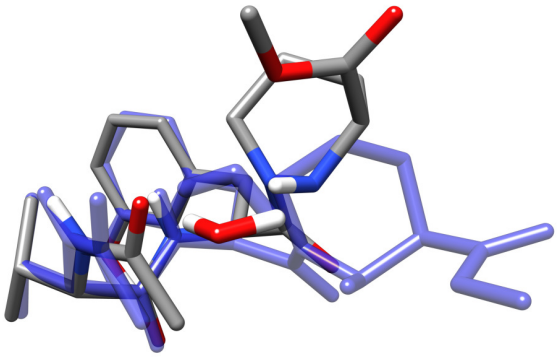

Figure D.5: Overlay of the Cpd_B (conformer a) geometries for the fixed gas-phase solute geometry (colored by heteroatom) and the fully optimized structure (colored in blue). Hydrogen atoms bound to carbon and many water molecules are omitted for clarity.

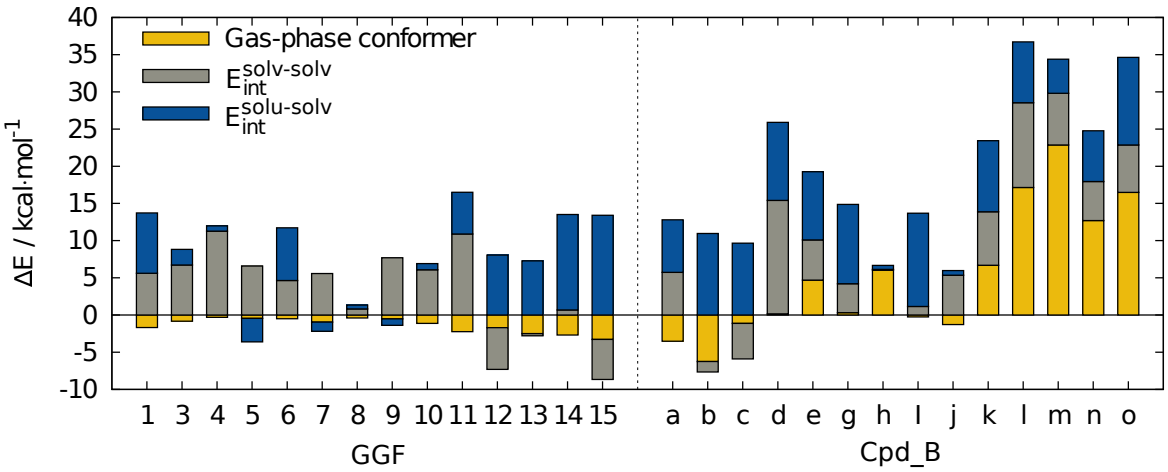

Figure D.6: Energy contributions to the r^{2}SCAN-3c conformational energy of the solvMPCONF196 benchmark set. Shown are the solute–solvent interaction, the gas-phase conformer, and the water-shell energy differences that add up to the total conformational energy.

cases, the structures do not change a lot, like for Cpd_B conformer a (RMSD of 0.2 Å), leading to the small energy difference of about 2 kcal mol^{-1} upon optimizing the whole structure. However, depending on the conformer geometry, also larger energy changes (e.g., 16 kcal mol^{-1} for conformer k of Cpd_B) can occur. For such structures, water molecules can move between two conformer strands upon full optimization, leading to a more open structure (Figure D.5) and thus a significant energy lowering of the system. This observation agrees with the general trend that for gas-phase conformers, relatively compact structures are energetically favored due to intramolecular dispersion interactions, which are quenched in solution, leading to more extended structures.[397] However, the solvMPCONF196 contains only structures where the solute has the same geometry as in the MPCONF196 gas-phase benchmark to allow a direct comparison between gas-phase and solvated results and a clear assignment of method errors. To further analyze the conformational energies of the solvMPCONF196, the three energy contributions are shown in Figure D.6 exemplarily for the

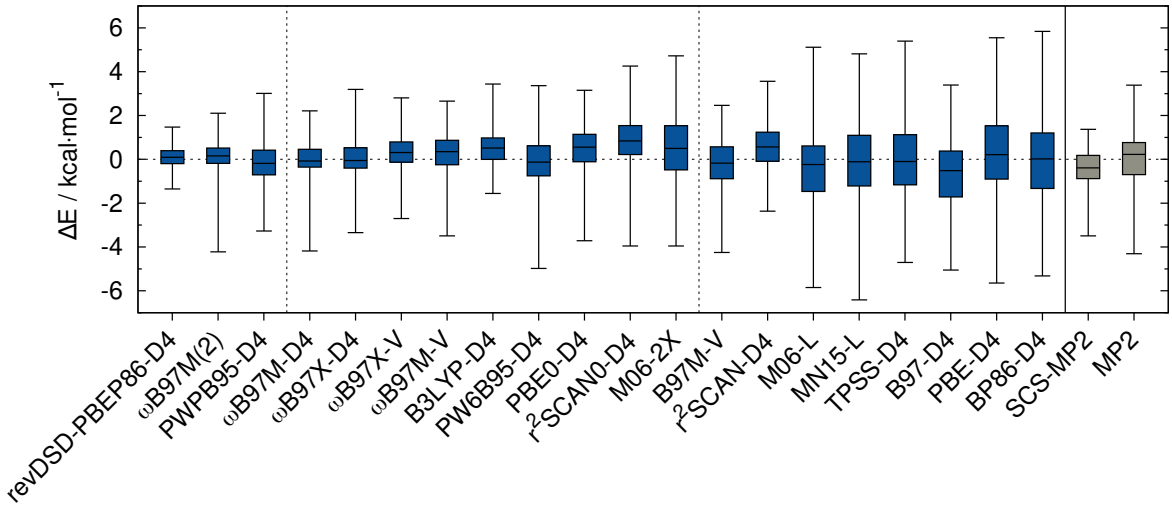

Figure D.7: Boxplot for various DFT and WFT method assessed on the solvMPCONF196 benchmark set. All results were obtained with the def2-QZVPP basis set.

conformers of GGF and Cpd_B. Two additional factors become important besides the conformational energy of the isolated conformer: the interaction energy of the solute and the solvent ($E_{int}^{solu-solv}$), and the solvent–solvent interactions $E_{int}^{solv-solv}$. For a few cases, the dominant part of the conformational energies is the difference between the pure solute (gas-phase) energies like for Cpd_b (m). Further, there are cases where the conformers differ mainly by solute–water (*e.g.,* GGF (14) and Cpd_B (I)), or water–water interactions (*e.g.,* GGF (4) and Cpd_B (j)). However, in many cases, at least two of those factors significantly impact the energy ranking. On average over all considered conformers, these three parts contribute about equally to the conformational energies and thus, for predicting conformational energies of such biomolecules in solution, a reliable method should be able to describe solute–water, and water–water interactions as well as conformational gas-phase energies accurately.

D.3.2 Method evaluation

In the following, the performance of various DFT, WFT, SQM, and FF methods is evaluated for the solvMPCONF196 and the MPCONF196 with respect to the PNO-LCCSD(T)-F12b/AVQZ' reference conformational energies. Mean deviations (MDs), mean absolute deviations (MADs), root-mean-square deviations (RMSDs), and absolute maximal deviation (AMAX) as defined in the Supporting Information are given in Table D.2. Since the conformer ranking is important for practical applications, Spearman coefficients are also shown, which represent the correlation of the conformer ranks. The corresponding statistics for the MPCONF196 with PNO-LCCSD(T)-F12b/AVQT reference values are given in the SI (Table S1, Figures S1, and Figure S2).

DFT/WFT evaluated with a large basis set

First, the performance of the DFT/WFT methods combined with the def2-QZVPP basis set is evaluated. A graphical representation of their performance on the solvMPCONF196 benchmark is given as a boxplot in Figure D.7. Considering the mean conformational energy of 9.9 kcal mol^{-1} for the whole

Table D.2: Error statistics in kcal mol^{-1} for the solvMPCONF196 with respect to the PNO-CCSD(T)-F12b reference values that yield a mean conformational energy of 9.9 kcal mol^{-1}.

Method	MD	MAD	RMSD	AMAX	Spearman
revDSD-PBEP86-D4[125,199]	0.11	0.37	0.49	1.47	0.9949
ωB97M(2)[398]	0.15	0.51	0.74	4.22	0.9916
PWPB95-D4[399]	−0.17	0.71	0.93	3.27	0.9897
ωB97M-D4[400]	0.06	0.54	0.75	4.18	0.9924
ωB97X-D4[130,400]	0.14	0.62	0.89	3.34	0.9900
ωB97X-V[401]	0.35	0.69	0.89	2.80	0.9881
ωB97M-V[402]	0.33	0.76	0.99	3.50	0.9903
B3LYP-D4[200,201]	0.54	0.82	1.06	3.44	0.9873
PW6B95-D4[202]	−0.05	0.84	1.12	4.98	0.9854
PBE0-D4[203]	0.57	0.90	1.13	3.71	0.9870
r^{2}SCAN0-D4[403]	0.92	1.05	1.38	4.25	0.9873
M06-2X[204]	0.50	1.30	1.65	4.72	0.9803
B97M-V[205,206]	−0.14	0.85	1.07	4.25	0.9814
r^{2}SCAN-D4[222]	0.61	0.96	1.26	3.56	0.9854
M06-L[223]	−0.41	1.24	1.59	5.85	0.9792
MN15-L[404]	−0.14	1.37	1.72	6.42	0.9714
TPSS-D4[405]	−0.01	1.39	1.78	5.39	0.9779
B97-D4[406]	−0.70	1.42	1.86	5.05	0.9795
PBE-D4[407]	0.23	1.44	1.91	5.64	0.9749
BP86-D4[207–209]	−0.06	1.45	1.86	5.84	0.9730
SCS-MP2[408]	−0.43	0.72	0.92	3.49	0.9889
MP2[116]	0.11	0.90	1.19	4.31	0.9865
r^{2}SCAN-3c[129]	−0.24	1.18	1.68	6.01	0.9768
ωB97X-3c[130]	1.05	1.27	1.70	6.47	0.9862
B97-3c[409]	−0.05	1.59	2.15	5.71	0.9773
PEBh-3c[210]	2.59	3.44	4.08	10.14	0.9552
HF-3c[128]	1.85	3.44	4.35	16.42	0.9341
GFN2-xTB[49]	0.12	3.22	4.45	12.53	0.9338
GFN1-xTB[48]	−0.30	3.86	4.76	13.33	0.9209
DFTB3[135]	−3.60	4.47	5.41	14.11	0.9128
PM6-D3H4X[133,410]	−0.27	4.59	5.81	20.33	0.8723
PM7[134]	−1.31	5.32	6.47	21.48	0.8739
MMFF94[211–215]	2.09	4.41	6.08	25.29	0.8456
GFN-FF[50]	−6.42	8.01	9.63	27.90	0.6660
GAFF[411]	−8.12	9.67	12.01	32.90	0.6109
Ghemical[412]	−36.47	53.75	68.97	184.41	−0.2906
UFF[137]	−55.95	74.43	97.40	307.25	−0.3387

solvMPCONF196 benchmark set, all tested DFT/WFT methods yield reasonably small MADs (up to 1.5 kcal mol^{-1}) and RMSDs of up to 1.9 kcal mol^{-1}. The corresponding values for MPCONF196 are 1.2 kcal mol^{-1} and 1.7 kcal mol^{-1}, respectively, see SI. Also, the Spearman coefficients are close to 1 and generally better than for MPCONF196 as the energy differences between different conformers increase upon adding water, which leads to a clearer energetic separation of the different conformers. Considering the different classes of DFT functionals, performance trends according to Jacob's ladder are observed.[119]

For the (meta-)GGAs, most of the functionals (BP86-D4, PBE-D4, B97-D4, TPSS-D4, MN15-L, and M06-L) perform similarly with MADs between 1.5 kcal mol^{-1} and 1.2 kcal mol^{-1}. The two best performers in this class are r^{2}SCAN-D4 and B97M-V with MADs below 1 kcal mol^{-1}. Even though r^{2}SCAN-D4 has a shift (MD = 0.6 kcal mol^{-1}), it shows the smallest number of outliers among the tested (meta-)GGAs (AMAX=3.6 kcal mol^{-1}), leading to the overall good MAD. Considering the hybrid DFT methods, M06-2X yields the largest MAD for the solvated structures. This is followed by r^{2}SCAN0-D4, which shows with an MD of 1.1 kcal mol^{-1} an even larger shift than its meta-GGA equivalent r^{2}SCAN-D4. Generally, the inclusion of Fock exchange shifts the mean conformational energies to larger values (MDs > 0 kcal mol^{-1} for most of the tested hybrid functionals), which is more pronounced for the solvated structures compared to the gas-phase ones. Thereby, a significant increase of the conformational energies is usually observed for the larger structures, e.g., by 0.7 kcal mol^{-1} when going from PBE to PBE0. PBE0-D4 performs similarly well as PW6B95 and B3LYP-D4 with MADs between 0.8 kcal mol^{-1} and 0.9 kcal mol^{-1}. Even more accurate values can be achieved with the ωB97X-V functional or its reparameterized D4 variant.[130] Only ωB97M-D4 yields a lower MAD close to 0.5 kcal mol^{-1} with a very good Spearman coefficient. Looking at the MP2 methods, SCS-MP2 performs better than MP2 with an MAD of 0.7 kcal mol^{-1}, similar to the better hybrid functionals, but worse than double hybrids, which yield the best results. ωB97M(2) improves upon ωB97M-V, on which it is based, yielding values similar to ωB97M-D4. The overall best performer is revDSD-PBEP86-D4 with an MAD of 0.4 kcal mol^{-1} for the solvated structures (0.3 kcal mol^{-1} for the gas-phase structures), which corresponds to only 4 % of the mean conformational energy. This method is able to predict almost the complete reference conformer ranking correctly and shows no outlier of more than 1.5 kcal mol^{-1} from the coupled cluster reference.

To allow a better comparison of the method performances for the gas-phase and the solvated structures, the Pearson correlation coefficients of DFT/WFT methods in a large def2-QZVPP basis are depicted in Figure D.8. It is evident that the double hybrid, the hybrid functionals (except M06-2X), and (SCS-)MP2 methods perform similarly well for the gas-phase and solvated structures, with only a small loss in accuracy when adding explicit water molecules. This changes for the (meta-)GGAs, where the error increases significantly for the solvated structures. Only B97M-V and r^{2}SCAN-D4 perform similarly for the two sets. A possible reason for the worse performance of (meta-)GGAs for the solvated structures is that these polar systems, composed of many intermolecular hydrogen bonds, are more prone to the self-interaction error and thus worse described with a (meta-)GGA.[36,413]

To estimate the importance of the basis set, different "def2" basis sets were employed with the B97M-V method. The resulting deviations from the coupled cluster values are depicted in Figure D.9. It is evident that using a small basis set like def2-SVP is neither sufficient for conformational energies of the gas-phase nor the solvated structures. At least triple ζ basis sets should be used for DFT and WFT. This is especially important for the solvated structures, where the difference between def2-SVP and def2-TZVP is even larger compared to the gas-phase energies. Here, additional errors like intermolecular BSSE occur, which are significantly reduced when using the def2-QZVPP basis

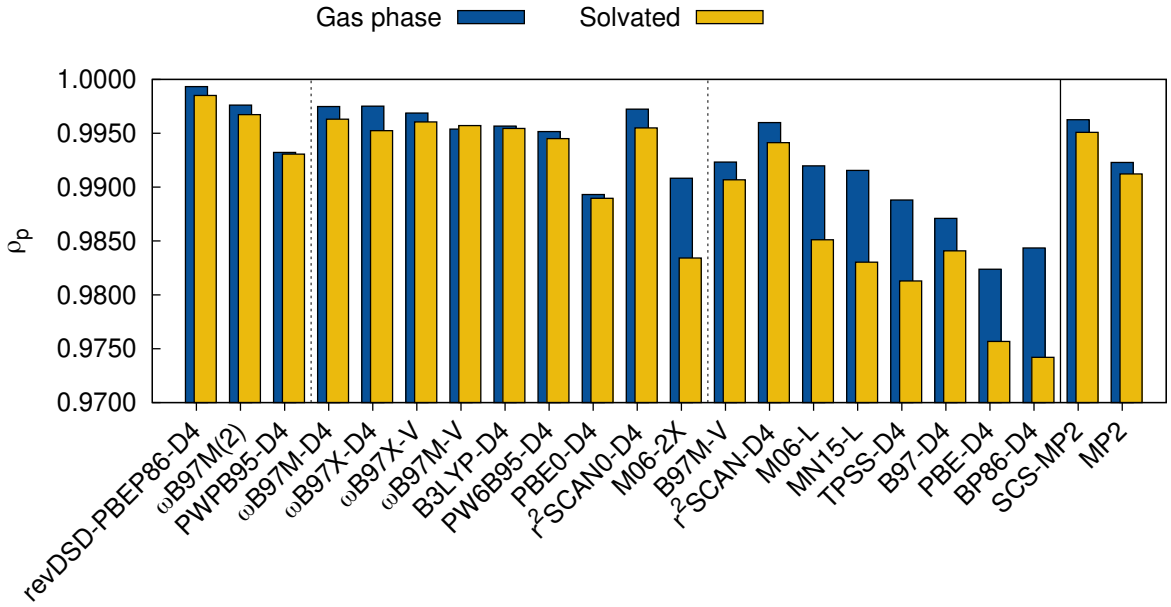

Figure D.8: Pearson correlation coefficient of the tested DFT/WFT methods for the gas-phase and the solvated structures. The def2-QZVPP basis set was employed for all methods.

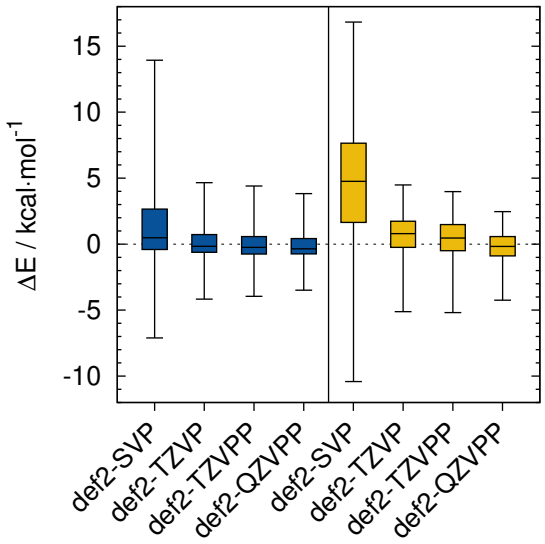

Figure D.9: Boxplot of the B97M-V deviations for different basis sets.

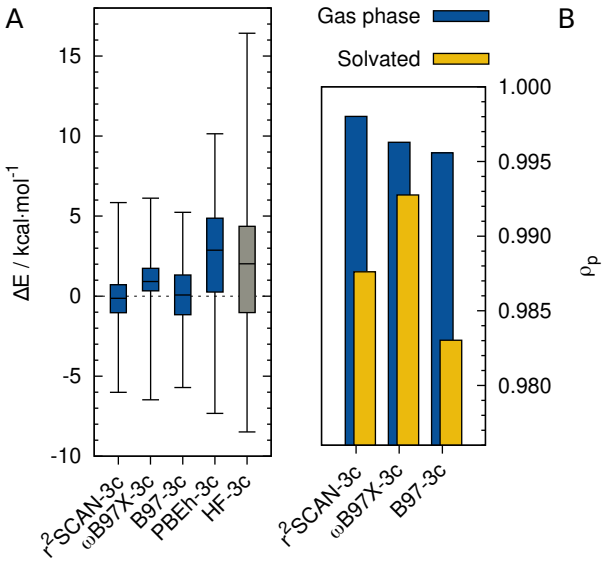

Figure D.10: (A) Boxplot of the different composite method results for the solvated geometries. (B) Pearson correlation coefficient for the best performing composite methods for the gas-phase structures and the solvated ones.

instead of the def2-SVP basis set. Thus, def2-QZVPP can be applied safely with DFT, which may be different for MP2 as WFT converges generally more slowly with the basis set size. For the gas phase set, even the def2-TZVP basis can be recommended for DFT calculations.

Composite methods

When computationally more efficient methods are required, e.g., for screening many different conformers or computing larger molecules like small proteins, composite DFT and WFT methods can be used. The results of different composite methods are shown as a boxplot together with the Pearson correlation coefficient in Figure D.10. Here, the worst performers are HF-3c and PBEh-3c with MADs of about 3.4 kcal mol^{-1} (2.3 kcal mol^{-1} and 1.6 kcal mol^{-1} for the gas-phase structures, respectively). These two older methods employ smaller basis sets (MINIX for HF-3c and a modified def2-SVP for PBEh-3c) that lead to larger errors, especially for the solvated structures as discussed above. The intermolecular BSSE can only partially be cured by the applied correction terms. B97-3c employing the def2-mTVP basis set performs reasonably well with an MAD of 1.6 kcal mol^{-1} for solvMPCONF196. ωB97X-3c performs better even though a polarized valence double ζ (vDZP) basis set is employed. However, this basis set was shown to have only a small BSSE of approximately def2-TZVP quality.[130] The best MAD of 1.2 kcal mol^{-1} among the tested composite methods has r^{2}SCAN-3c yielding results similar to good (meta-)GGAs in the large def2-QZVPP basis set for the solvated structures at a small fraction of computation time (see Sec D.3.2). Furthermore, r^{2}SCAN-3c yields remarkably good results also for the gas-phase structures with an MAD below 0.5 kcal mol^{-1} close to the best hybrid functional with a def2-QZVPP basis set.

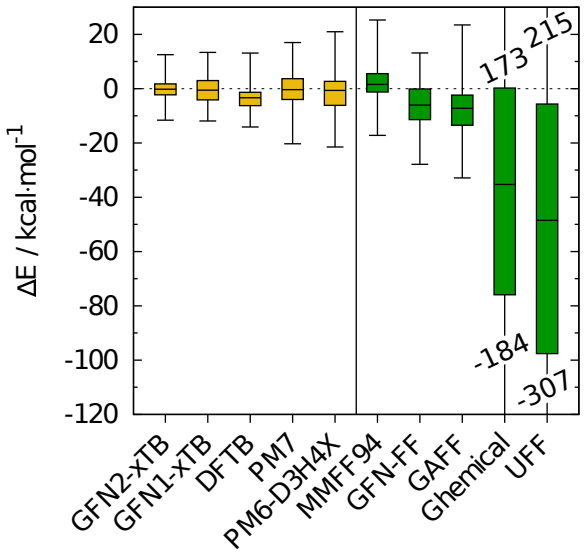

Figure D.11: Boxplot for the tested semi-empirical and FF methods assessed on the solvMPCONF196 benchmark set.

Semi-empirical and FF methods

If typical proteins or other biomolecules should be modeled, their size usually prevents the application of standard DFT/WFT methods, especially for conformer screening, and only semi-empirical and FF methods are computationally feasible. Therefore, we tested some well-established methods from these classes and depicted their results in Figure D.11. For the semi-empirical methods, PM7 and PM6-D3H4X perform worst with MADs of 5.3 kcal mol^{-1} and 4.6 kcal mol^{-1} for solvMPCONF196. The DFTB3 method shows significantly lower scattering and fewer outliers. However, it yields a large mean shift of −3.6 kcal mol^{-1} leading to a similar MAD and overall too low conformational energies. The best SQM methods, GFN1-xTB and GFN2-xTB, show the smallest scattering and MADs of 3.9 kcal mol^{-1} and 3.2 kcal mol^{-1} , which is a good result considering that a minimal basis set is applied.

For the FFs, the general Ghemical and UFF force fields that can be applied across the periodic table show very poor Spearman coefficients close to zero, indicating randomly ordered conformers compared to the reference. Moreover, they yield extremely large MADs of more than 53 kcal mol^{-1} and 74 kcal mol^{-1} and are thus not applicable for ranking conformers of biomolecules. GFN-FF performs best among the generally applicable force fields and even outperforms GAFF, which was specially parameterized for bio-organic systems. MMFF94, which has also been optimized for biomolecules, shows the best MAD of 4.4 kcal mol^{-1} among the tested force fields and is only slightly worse than the best SQM methods. Notably, it yields astonishingly good results for the gas-phase conformers with an MAD below 1 kcal mol^{-1} similar to DFT/WFT quality even though the RMSD is much higher (SI Table S1) due to some outliers, especially for SANGLI.

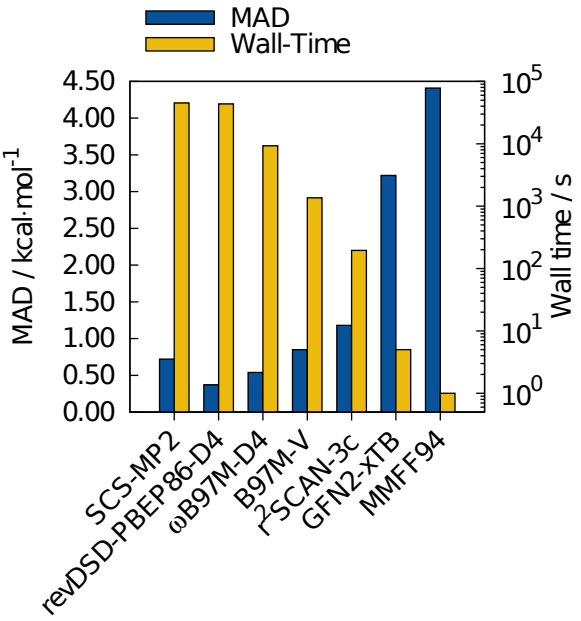

Figure D.12: MAD for the solvMPCONF196 benchmark set and wall times for YIVNOG (conformer a) solvated with water computed on 14 cores of an Intel® Xeon® CPU E5-2660 v4 @ 2.00GHz. Shown is the best performer of each method class.

Timings

To allow a fair comparison among the different classes of methods, not only the errors in predicting conformational energies have to be discussed, but also the respective computational timings. Therefore, the wall time and MAD for every best performer of each class are depicted in Figure D.12. While SCS-MP2 and the double hybrid reDSD-PBEP86-D4 have similar computational costs, they significantly differ regarding their MADs, with SCS-MP2 showing almost twice as large errors. Also, ωB97M-D4 yields better results than SCS-MP2 while saving almost a factor of five in computation time. B97M-V provides reasonable accuracy with an MAD below 1 kcal mol^{-1} but is almost seven times faster than the ωB97M-D4 hybrid. In practice, often larger structures with many conformers have to be screened. For this application, we recommend the composite method r^{2}SCAN-3c, which is even seven times faster than B97M-V/def2-QZVPP yielding an MAD close to 1 kcal mol^{-1} . However, proteins and large biomolecules or extensive conformer and structure screenings are often too expensive for DFT/WFT methods. Then, the semi-empirical GFN2-xTB method becomes a reasonable choice as it is more than one order of magnitude faster but still provides reasonable accuracy. For very large systems, we recommend MMFF94 or GFN-FF.

D.4 Conclusion

We extended the MPCONF196 conformer benchmark set toward systems in solution by solvating the 196 conformers of 13 different peptides and organic macrocycles with explicit water. Further, we provide accurate PNO-LCCSD(T)-F12b/AVQZ' reference values for the resulting solvMPCONF196 and the original MPCONF196 set and evaluated 16 different DFT methods, (SCS-)MP2, five SQM,

and five FF methods on those. Generally, for conformational studies of proteins, at least a few explicit water molecules might be considered because the solute–water and water–water interactions can have large contributions to the conformational energies. Also, the conformer structure can change under the influence of explicit water molecules, and thus, full structure optimizations are required if gas-phase conformers are solvated. For this, we recommend programs like the QCG[289] that can add water molecules and perform conformational searches automatically. As the conformer generation of larger biomolecular systems can become computationally very expensive, efficient methods like GFN2-xTB or, if this is still too expensive, GFN-FF and MMFF94 are recommended. For further refining the conformational ranking and sorting out structures, r^2SCAN-3c is well suited as it provides a very good cost–accuracy ratio. If highly accurate conformational energies are required, ωB97M-D4 or revDSD-PBEP86-D4 with a def2-QZVPP basis set may be used for a final conformer ranking. Thereby, particular attention has to be paid to the choice of the basis set as the explicit water molecules introduce intermolecular interactions, which make the systems prone to BSSE.

Supporting information

The Supporting Information is available free of charge at
https://onlinelibrary.wiley.com/doi/full/10.1002/jcc.27248.

- Statistical error measures, details on the reference value computation, and method assessment for the MPCONF196 (PDF)
- Structures of the benchmark sets in xyz format (ZIP)
- Relevant energies (xlsx)

Conflict of interest

There are no conflicts to declare.

Acknowledgements

This work was financially supported by Merck KGaA. The authors thank M. Bursch, and T. Gasevic for helpful discussions. The technical support from H.-J. Werner concerning the Molpro calculations is greatly appreciated.

APPENDIX E

ONIOM meets *xtb*: efficient, accurate, and robust multi-layer simulations across the periodic table

Christoph Plett,* Abylay Katbashev,* Sebastian Ehlert,† Stefan Grimme,* Markus Bursch‡

Reprinted (adapted) with permission§ from:
C. Plett, A. Katbashev, S. Ehlert, S. Grimme, and M. Bursch, Phys. Chem. Chem. Phys. **25** (2023) 17860.

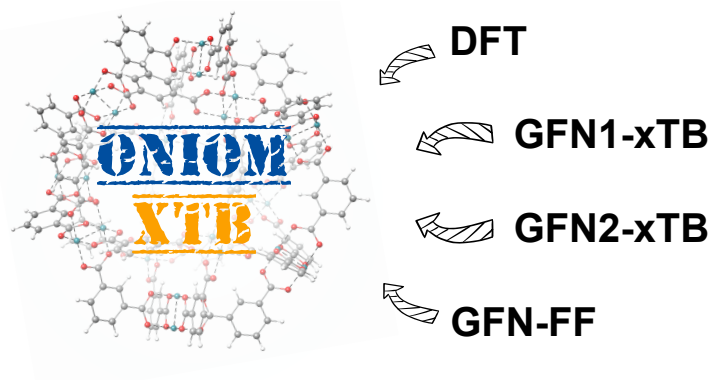

Figure E.1: Associated Table of Contents graphic for publication in Physical Chemistry Chemical Physics.

*Mulliken Center for Theoretical Chemistry, University of Bonn, Beringstr. 4, D-53115 Bonn, Germany
†Microsoft Research AI4Science, Evert van de Beekstraat 254, 1118 CZ Schiphol, The Netherlands
‡Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, 45470 Mülheim an der Ruhr, Germany.

Abstract

The computational treatment of large molecular structures is of increasing interest in fields of modern chemistry. Accordingly, efficient quantum chemical approaches are needed to perform sophisticated investigations on such systems. This engaged the development of the well-established "Our own N-layered Integrated molecular Orbital and molecular Mechanics" (ONIOM) multi-layer scheme [L. W. Chung *et al., Chemical Reviews*, 2015, **115**, 5678–5796]. In this work, we present the specific implementation of the ONIOM scheme into the *xtb* semi-empirical extended tight-binding program package and its application to challenging transition-metal complexes. The efficient and broadly applicable GFN*n*-xTB and -FF methods are applied in the ONIOM framework to elucidate reaction energies, geometry optimizations, and explicit solvation effects for metal-organic systems with up to several hundreds of atoms. It is shown that an ONIOM-based combination of density functional theory, semi-empirical, and force-field methods can be used to drastically reduce the computational cost and thus enable the investigation of huge systems at almost no significant loss in accuracy.

E.1 Introduction

Current developments and modern techniques for synthesizing and analyzing molecules allow chemists to study increasingly large and complex systems including, e.g., macromolecules, reaction networks, or supramolecular complexes.[414–416]. Besides a long-standing emphasis on protein structures and their functions in the macro-molecular regime,[417], large metal-organic structures such as metal-organic polyhedra (MOP)[187,418] or molecular machines[419,420] gain growing attention and find application in various fields like drug-delivery processes,[421,422] functional materials,[423] catalysis,[424,425] or fuel storage.[267] Such metal-organic systems can also reach impressive sizes of hundreds to thousands of atoms and may involve moieties with challenging electronic structures. Besides advancing experimental techniques, computational methods and workflows became a valuable utility for a reliable description of large molecular systems enabling deeper insights into basic properties and mechanisms.[16,426] However, the application of accurate quantum chemical (QC) methods like density functional theory (DFT) and wave function theory (WFT) methods to extended systems is often limited due to the rapidly increasing computational costs with the system size.[427]

This encouraged the continuous development of force field (FF) and semi-empirical quantum mechanical (SQM) methods that are routinely applied to systems of hundreds up to thousands of atoms.[266,428–430] More recent developments in this field are the GFN*n*-xTB and force-field methods[84], GFN1-xTB,[48] GFN2-xTB,[49] and GFN-FF[50], which have become widely used and well-established tools for the treatment of large systems.[431,432] In contrast to most other semi-empirical approaches and FFs, the GFN methods are consistently parameterized for all elements up to Rn and are thus applicable to a large chemical space throughout the periodic table. These methods are already routinely used in black-box approaches for conformer sampling[44], energetic sorting of structure ensembles,[35] or docking algorithms such as the recent aISS[386] method, with the latter being specifically suited for molecule-binding to large structures.[267] Nevertheless, specifically complex and challenging electronic structures often require an even more accurate direct or post-processing description at a more sophisticated DFT or WFT level.[36]

To close this gap, various multi-layer schemes have emerged. They rely on the fact that treating the whole system highly accurately is often unnecessary, but rather a small reactive part or interaction

site is of importance.[83] For multi-layer schemes, computationally demanding but accurate methods are applied only to a small region of interest, while non-negligible environmental effects are treated with computationally cheaper methods. Crucial environmental influences include, e.g., backbone strain in proteins[433], supramolecular stabilization of reactive species,[434,435] reaction mechanisms including large structurally strained systems,[436] or explicit solvation.[437,438] Besides well-known QM/MM schemes,[439] the so-called "Our own N-layered Integrated molecular Orbital and molecular Mechanics" (ONIOM) method[83,440] proved to be a valuable tool for multi-layer simulations. ONIOM was already successfully applied for modeling organic reactions such as Friedel–Crafts reactions[441], metal-organic catalysis,[442] and zeolite reactivity to name only a few.[443] In this context, the GFN methods represent a perfect match to simulate the outer region within the ONIOM scheme due to their broad parameterization and reasonable accuracy.[266,444–446]

Accordingly, we present here the implementation of the ONIOM scheme in the open source *xtb* program that allows the direct combination of the GFN methods with any desired DFT or WFT method via an interface to the prominent *ORCA*[387,447] or *TURBOMOLE*[448,449] program packages in an easily applicable and intuitive way. This enables the efficient and accurate treatment of not only organic structures but also of molecules containing, e.g., transition metals, heavy main group elements, or even lanthanoids.[266,446] Besides the theoretical excursion into the ONIOM approach and its implementation in *xtb*, we demonstrate its efficient application by investigating molecular geometries and energies of challenging metal-organic structures.

E.2 Theoretical overview and implementation

In principle, multi-layer techniques used to intermix different levels of QC theory are either additive or subtractive. Additive schemes such as QM/MM or QM/QM were established first and rely on coupling terms between the two regions treated with different methods.[450,451] Subtractive schemes like ONIOM do not require additional coupling terms in the Hamiltonian and are thus easier to implement and allow the straightforward mixing of any methods. For the simplest form of the ONIOM scheme, the whole system is divided into two layers: an inner and outer region. These regions are typically treated separately with a high- ($E^{\text{high}}_{\text{inner}}$) (e.g., DFT) and low-level ($E^{\text{low}}_{\text{outer}}$) method (e.g., SQM or FF) and consequently fused to obtain the ONIOM energy expression:

$$E_{\text{ONIOM}} = E^{\text{high}}_{\text{inner}} + E^{\text{low}}_{\text{outer}}. \tag{E.1}$$

The low-level description of the outer region is obtained by the subtraction of the energy of the inner region ($E^{\text{low}}_{\text{inner}}$) from the energy of the whole system ($E^{\text{low}}_{\text{whole}}$) calculated with the corresponding low-level approach:

$$E^{\text{low}}_{\text{outer}} = E^{\text{low}}_{\text{whole}} - E^{\text{low}}_{\text{inner}}. \tag{E.2}$$

In this implementation, no electrostatic embedding is employed for the calculation of the inner region.

If requested, solvation effects can be accounted for by a simplified solvation embedding. Thereby, only $E^{\text{low}}_{\text{whole}}$ is computed embedded in an implicit solvent model while the inner region is always computed in the gas phase.

E.2.1 ONIOM boundary

The artificial separation of a chemical system into different regions and the creation of the corresponding boundaries between them can be challenging. The trivial case is to deal with boundaries that include only non-covalent interactions. In contrast, treating large molecules with a coupling scheme like ONIOM usually leads to covalent bond breaking. In such cases, the artificial partitioning of the system leads to the formation of radicals or dangling bonds at the boundaries. To address this problem, a number of different approaches were developed[83,216] such as the link atom (LA) approach[452,453] or the frozen localized orbitals[454] approach. Herein, we consider only the former, which is implemented within the *xtb* software in the course of this work.

The general idea behind the link atom saturation technique is the capping of the cleaved bonds to fill the incomplete valence shells of the interface atoms of the partitioned inner region. This is done via the introduction of a dummy atom, usually hydrogen, on the bond vector between the boundary atoms. The exact position of a link atom is derived as:

$$\vec{R}_{\mathrm{LA}} = \vec{R}_{\mathrm{inner}} + k(\vec{R}_{\mathrm{outer}} - \vec{R}_{\mathrm{inner}}) \tag{E.3}$$

where $\vec{R}_{\mathrm{inner}}$ and $\vec{R}_{\mathrm{outer}}$ are the coordinates of the atoms in the inner and outer regions with mutual bonds that split during the ONIOM procedure, and k is the distance scaling factor. Conventionally, k is taken as a constant determined from the predefined element-specific distances between the inner region atom (IA) - outer region atom (OA) and IA-H atoms listed in the supporting information. For cases without listed values, a default value for k is chosen as the ratio of the carbon–carbon/carbon–hydrogen bond distances. As an alternative, we implemented the possibility to make the scaling factor k dependent on the actual values for the IA-OA distance that are present in the system instead of tabulated values. The corresponding derivatives are given in the SI. Figure E.2 demonstrates the ONIOM interpolation scheme, where the chemical system is first partitioned by cleaving its C-C σ bonds, and then saturated with hydrogen atoms via the link atom approach.

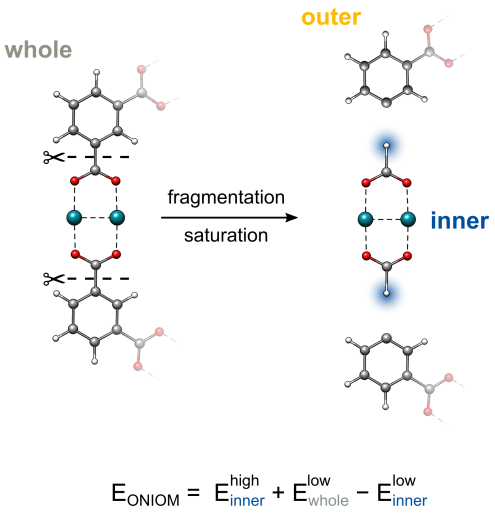

Figure E.2: The fragmentation and subsequent saturation procedures within the ONIOM framework. Capping hydrogen atoms are highlighted in blue.

E.2.2 Topology

One of the distinguishing features of the *xtb* ONIOM implementation is the usage of the topology information generated with the corresponding GFN method to automate and validate the partitioning and saturation. As cutting double or triple bonds can lead to problematic systems, *xtb* uses its internal bonding topology data to prohibit the cleavage of higher-order bonds. Moreover, an automatic charge identification routine based on partial charges calculated with the respective GFN method is used to determine the total charge of the inner region automatically.

E.2.3 Jacobian

While the ONIOM single-point energy can be evaluated directly using Equation E.1, structure optimizations and frequency calculations require energy derivatives, such as gradients or the Hessian, that need additional treatment. The major complication emerges from the forces introduced by the artificial capping atoms that have to be reassigned accurately to the corresponding real atoms. For this purpose, the Jacobian matrix is used, which is defined as

$$J(\vec{R}_{\text{inner}}; \vec{R}_{\text{whole}}) = \frac{\delta \vec{R}_{\text{inner}}}{\delta \vec{R}_{\text{whole}}} \tag{E.4}$$

where R_{inner} denotes the coordinates of the inner region atoms, while R_{whole} are those of the atoms of the entire system. Taking this correction factor into consideration, the final expression for the ONIOM gradients for the two-layer interpolation is derived from the gradients calculated with the high- ($\vec{g}^{\,\text{high}}_{\text{inner}}$) and low-level ($\vec{g}^{\,\text{low}}_{\text{whole}}$, $\vec{g}^{\,\text{high}}_{\text{inner}}$) methods and can be written as

$$\vec{g}_{\text{ONIOM}} = \vec{g}^{\,\text{low}}_{\text{whole}} - \vec{g}^{\,\text{low}}_{\text{inner}} J(\vec{R}_{\text{inner}}; \vec{R}_{\text{whole}}) + \vec{g}^{\,\text{high}}_{\text{inner}} J(\vec{R}_{\text{inner}}; \vec{R}_{\text{whole}}) \tag{E.5}$$

The formation of the Jacobian matrix requires the differentiation of Equation E.3 with respect to the real coordinates ($\vec{R}_{\text{whole}}$).

E.2.4 Implementation and availability

The ONIOM scheme implemented in the *xtb*[84,455] program package is invoked with one simple command line instruction and can be used as a standalone program with any GFN method combination. Further, it can be interfaced to *ORCA* or *TURBOMOLE* to combine DFT/WFT with the GFN methods. The application guideline and installation instructions for the *xtb* program can be found at the xtb-docs page.

E.3 Computational details

All GFN1-xTB, GFN2-xTB, GFN-FF, and ONIOM calculations were performed with the *xtb* 6.6.0 program package. DFT computations, including also the DFT part for the ONIOM scheme executed and processed by *xtb*, were performed with *TURBOMOLE* 7.5.1 or 7.6, except for the solvent clusters where *ORCA* 5.0.3 was employed. If not stated otherwise, the GBSA[181] (SQM, FF) (*xtb*, toluene),

COSMO[155,317] (*TURBOMOLE*, toluene), and CPCM[154,388] (*ORCA*, DMSO) implicit solvation models were used and will be discussed in more detail in section E.4.2.

All DFT calculations employ Ahlrichs' def2 Gaussian-type atomic orbital basis sets[105] and some modified variants developed in the framework of the respective DFT-3c composite methods PBEh-3c[210] (def2-mSVP), B97-3c[409] (def-mTZVP), and r^{2}SCAN-3c[129] (def-mTZVPP). Matching def2 effective core potentials[456] were used throughout. If not stated otherwise, the default calculation settings and convergence criteria of the respective codes were applied including the use of the resolution of the identity (RI) approximation with matching auxiliary basis sets.[394] The D4[125,254,457] London dispersion correction was used throughout.

The systematic workflow for locating transition states was adopted from the work of Dohm *et al*.[458] The double-ended Growing-String Method (GSM) by Zimmerman *et al.* was used to refine reaction paths[459–461] at the GFN2-xTB potential energy surface (PES). The resulting saddle points were further optimized with corresponding density functionals to obtain final transition state geometries. All calculations were executed on an Intel® Xeon® CPU E5-2660 v4 @ 2.00GHz machine.

E.4 Results and discussion

In the following, the application of our general ONIOM scheme implementation in *xtb* is demonstrated for selected showcases ranging from geometry optimizations of MOF cutouts, and explicit solvation of transition metal complexes to reaction mechanism exploration at metal-organic polyhedra.

E.4.1 Molecular structures

A great benefit of multi-layer approaches such as ONIOM is the drastic reduction of computation time compared to a full DFT or WFT approach. This is especially important for geometry optimizations that often represent one of the most time-consuming tasks in computational chemistry workflows.

As a demonstration, the following examples show the use of our implementation for geometry optimizations of challenging metal-organic systems. To allow for a fair assessment of the ONIOM treatment, the showcase systems were selected to still be treatable at a reasonable DFT level. Thus, a direct comparison of the multi-layer combinations and full-DFT is possible. As the size of the investigated systems limits the choice of suitable methods, the TPSS[405] functional in combination with the def2-SVP basis set was chosen as DFT approach. Nevertheless, employing the ONIOM scheme would allow for the treatment of much larger structures, where a full DFT treatment, even at this low level, would be unfeasible.

As a first example, a 484-atom large cut out of the prominent zirconium-based UiO-66 MOF, synthesized by Cavka *et al.*,[218] was examined. Various combinations of GFN and DFT methods were tested within the ONIOM scheme (Figure E.3) for optimizing the geometry. For the ONIOM calculations, the Zr-containing nodes were treated with the high-level method. The resulting method combinations are denoted as "*ONIOM(high-level method:low-level method)*" in the following. The influence of the inner region size on the ONIOM performance was tested by including one, three, and six Zr nodes. In general, the included Zr nodes were treated as one single extended inner region, not as separated subregions. As a measure of structural quality, the heavy-atom root-mean-square deviation (hRMSD) to the crystal structure is used.

Despite having the largest deviation from the reference, the atomistic GFN-FF is highly efficient in

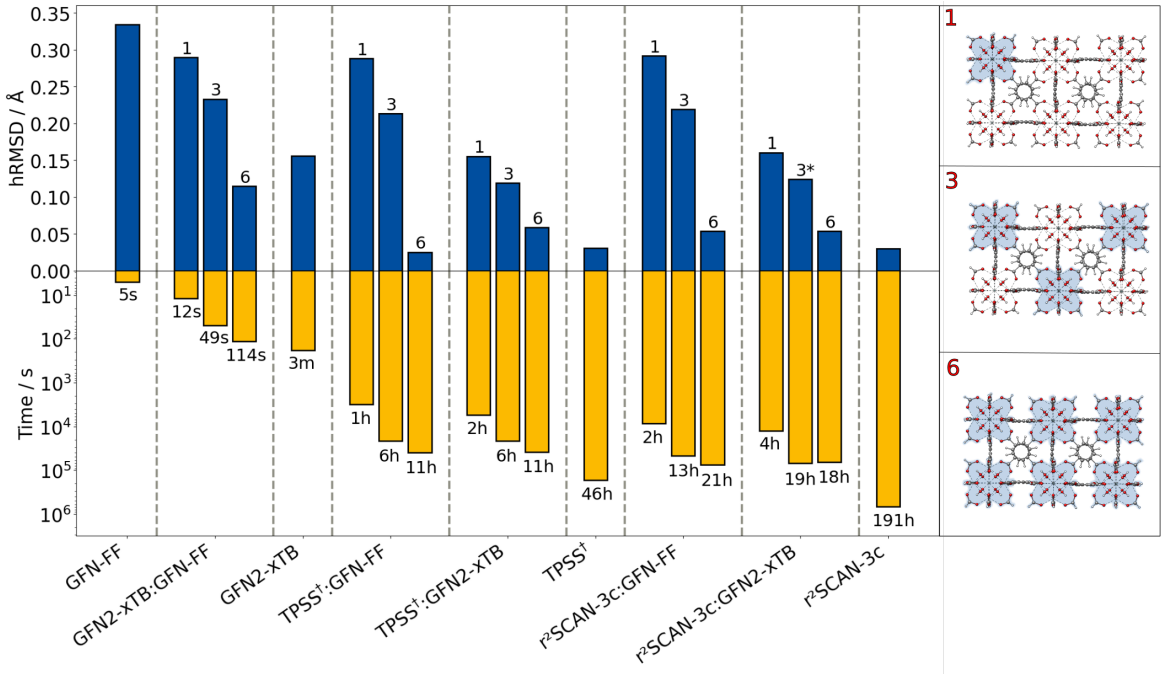

Figure E.3: hRMSD (vs. X-Ray reference) and wall time values (14 cores) for the geometry optimization of the UiO-66 polyhedron. 1, 3, and 6 are the numbers of the metal clusters included in the inner region, which is highlighted in blue. TPSS† = TPSS-D4/def2-SVP. *Extended number of optimization cycles due to the convergence issue.

terms of computational timings (5s). It performs well for the purely organic linker and is still reasonable for the Zr-moiety. The semi-empirical GFN2-xTB shows an improved structure compared to GFN-FF by still remaining computationally efficient (3 min). Applying the ONIOM(GFN2-xTB:GFN-FF) scheme yields a better hRMSD compared to a full GFN-FF optimization already with only one Zr moiety in the inner region. By extending the GFN2-xTB region to the other Zr moieties, the geometry can be improved systematically while still being computationally very efficient. Applying GFN2-xTB to each of the six Zr nodes leads to a slightly lower hRMSD compared to the full GFN2-xTB optimization. The combination of GFN2-xTB for the metal-containing moieties and GFN-FF for the purely organic linker combines the strength of both methods and yields a very reasonable result within seconds optimizing the full structure in seconds to a few minutes.

The overall lowest hRMSDs are observed for the two DFT methods (TPSS and r^{2}SCAN-3c), but with these methods, the geometry optimization takes several hours up to more than a week for r^{2}SCAN-3c. The combination of DFT methods and GFN-FF does not systematically yield improved hRMSDs over the ONIOM(GFN2-xTB:GFN-FF) approach for the inclusion of one or three Zr nodes. This may be due to asymmetrical distortions of the overall structure which is not the case when six nodes are included. Then, the ONIOM(DFT:GFN-FF) combination yields excellent agreement with the reference and the full-DFT treatments. In this case, the computation time is reduced from 46 to only 11 hours at no loss in accuracy. This reduction is even larger for the ONIOM(r^{2}SCAN-3c:GFN-FF) combination, where it is reduced from 191 hours to 21 hours. Accordingly, the ONIOM approach can save days to weeks of computation time at no relevant loss in accuracy. Noteworthy here is the lower

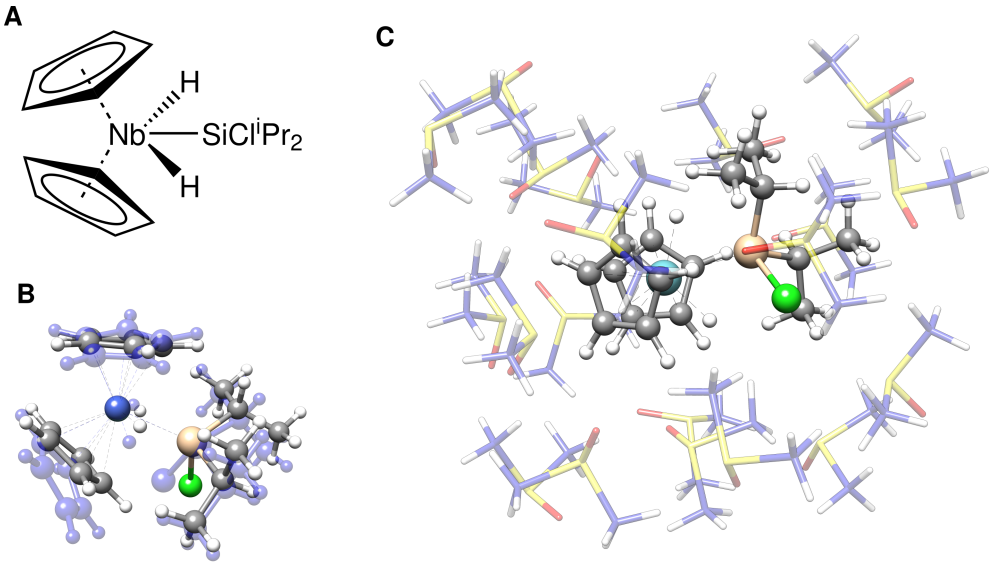

Figure E.4: (A) Lewis structure depiction of $[Cp_2Nb(H)_2(SiCl^iPr_2)]$. (B) structure overlay of the full GFN2-xTB (blue) and TPSS-D4/def2-SVP (color-coded) optimized structures. DMSO molecules are removed for clarification. (C) TPSS-D4/def2-SVP optimized geometry of $[Cp_2Nb(H)_2(SiCl\mathit{i}Pr_2)]$ dissolved by 20 DMSO molecules.

computational time for the inner region including six nodes with the ONIOM(r^{2}SCAN-3c:GFN2-xTB) scheme compared to that including three nodes. This can be explained by a faster convergence of the geometries resulting from the consistent DFT treatment of the zirconium-centered moieties.

Another possible use case of our implementation is the modeling of explicit solvation. If an implicit solvation model becomes insufficient to describe the solvent effect, it is necessary to include explicit solvent molecules, which can increase the computational costs tremendously. To still include a sufficient amount of explicit solvents to describe the system accurately, the ONIOM scheme can be used to employ a DFT or WFT method for the solute and a GFN method for the solvent. In this case, no covalent bonds are broken, which renders the GFN methods specifically suitable as they are designed to reproduce non-covalent interactions accurately.

As an example, the Bis(cyclopentadienyl)silylniobium complex $[Cp_2Nb(H)_2(SiCl^iPr_2)]$[217] (Figure E.4 A) was solvated with 20 DMSO molecules using the Quantum-Cluster-Growth (QCG) algorithm. [289,386] This transition metal complex has previously proven to be a specifically difficult case for common semi-empirical methods in the TMG145 benchmark study.[266] The solvated complex was optimized at three different levels, TPSS-D4(CPCM)/def2-SVP (Figure E.4 C), GFN2-xTB(ALPB), and ONIOM(TPSS-D4/def2-SVP:GFN2-xTB(ALPB)). In the ONIOM scheme, the Bis(cyclopentadienyl)silylniobium complex was chosen as the inner region. A comparison of the cut-out solute geometry optimized within the ONIOM scheme and pure GFN2-xTB is shown in Figure E.4 B. Optimizing the explicitly solvated complex with TPSS-D4/def2-SVP retains the major structure motif of the Bis(cyclopentadienyl)silylniobium complex with respect to the gas-phase TPSSh[462]-D3(BJ)-ATM reference structure.[266] Using GFN2-xTB for optimizing the whole cluster yields a qualitatively wrong, distorted structure with an RMSD of 0.26 Å of the cut-out solute. Thereby, a haptotropic shift of the cyclopentadienyl ligands occurs, and the symmetric coordination of the

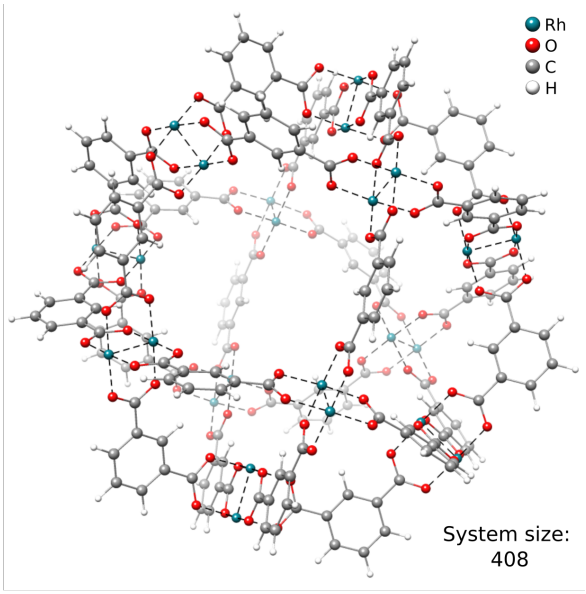

Figure E.5: The molecular structure of the DALTES metal-organic polyhedron.

terminal hydride ligands is distorted. Accordingly, GFN2-xTB is, similar to other semi-empirical and Force Field methods,[266] not able to describe this complex system accurately. Applying the ONIOM(TPSS-D4/def2-SVP:GFN2-xTB) scheme instead yields the correct structure in excellent agreement with the full DFT optimization (RMSD of 0.01 Å) at a fraction of computational cost (59 min vs 48 h 58 min). Despite the good performance of the ONIOM scheme demonstrated for this example, electrostatic embedding may enhance the results for cases with highly polar environments showing strong charge-dependent interactions. Nevertheless, the appropriate choice of the ONIOM boundary may reduce errors in this respect.

E.4.2 Electronic energies

Another key ability of the ONIOM scheme is the calculation of energies for extended structures. This is specifically useful for, e.g., reaction mechanism elucidation, where the chemical transformation is typically occurring at a relatively small reactive site.

As a first example, we discuss the Rh-functionalized metal-organic cuboctahedron illustrated in Figure E.5. In the following, it is referred to as DALTES[186] in correspondence with its reference code from the Cambridge Structural Database (CSD).[463] Herein, we consider the cyanosilylation reaction facilitated by the dirhodium paddle-wheel nodes (Figure E.6 A). The proposed mechanism depicted in Figure E.6 B was adapted from Zhang *et al.*[219]. The first step is the coordination of an aldehyde substrate to the open Rh sites of DALTES, which is followed by the nucleophilic addition of the isomerized trimethylsilyl cyanide (iso-TMSCN) to the activated carbonyl compound. Afterward, the trimethylsilyl group isomerizes into the final product, which is the rate-determining step of the catalytic reaction. Finally, the produced cyanohydrin derivative is cleaved from the metal cluster of the catalysts. To validate the ONIOM approach, a full quantum mechanical reference energy profile was calculated. Due to the size of the investigated system (404 atoms), the choice of the reference method was again limited to the (meta-)GGA/DZ level. Accordingly, the TPSS-D4/def2-SVP level

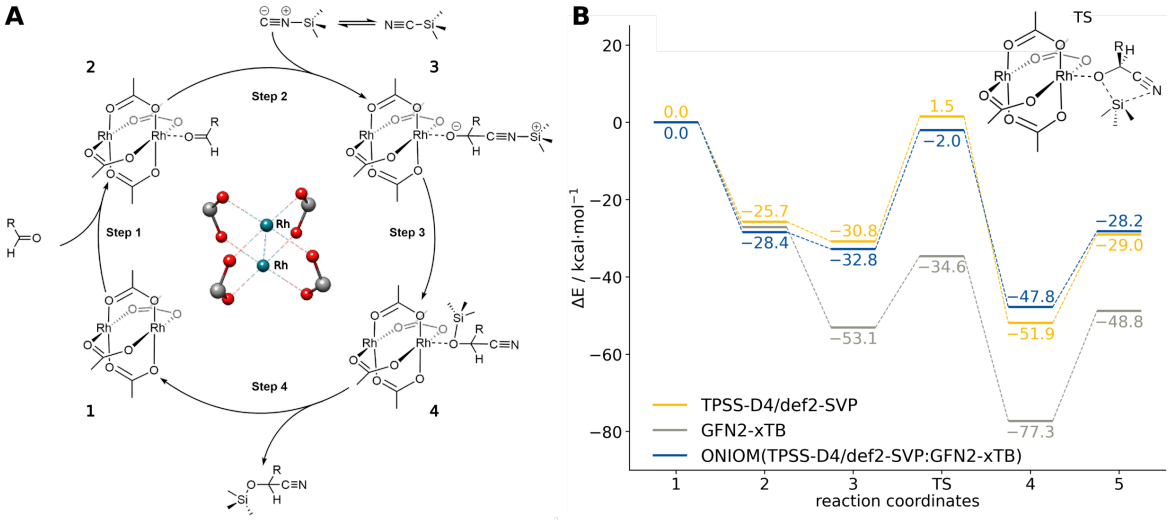

Figure E.6: (A) Schematic reaction mechanism of the cyanosilylation at an Rh-based paddle-wheel motif of the DALTES MOP. (B) The relative potential energy curve of the cyanosilylation reaction computed with TPSS-D4/def2-SVP, GFN2-xTB, and their ONIOM(TPSS-D4/def2-SVP:GFN2-xTB) combination onto the TPSS-D4/def2-SVP optimized geometries implicitly solvated in toluene.

with COSMO implicit solvation was chosen as a reference and QM component in the ONIOM approaches. The energy profile of the catalytic cycle was then recomputed at the GFN2-xTB and ONIOM(TPSS(COSMO):GFN2-xTB(ddCOSMO)) levels (Figure E.6). For the ONIOM scheme, only the respective dimetal node involved in the cyanosilylation and the corresponding reactants were included in the inner region.

The semi-empirical GFN2-xTB energy profile qualitatively agrees with the full-DFT results, but the relative energies differ by up to 35 kcal mol^{-1}. Applying the ONIOM approach instead yields very good agreement with the TPSS-D4/def2-SVP reference energy, only varying by 1-3 kcal mol^{-1}. Accordingly, the ONIOM scheme can be used to effectively converge the results by a combination of GFN2-xTB with any suitable DFT method of choice. By using ONIOM, the average computational wall time of the energy evaluations is reduced from 30 minutes to only 40 seconds. This allows choosing a significantly more accurate method in the inner region that would be unfeasible in a full DFT approach. Similar results are obtained for other methods combinations such as ONIOM(r^{2}SCAN-3c:GFN2-xTB) and ONIOM(ωB97X-3c:[130]GFN2-xTB) (see SI).

An even more challenging system is the recently synthesized porphyrin-based macromolecular spoked-wheel complex by Majewski *et al.*[220], illustrated in Figure E.7. Its size of 870 atoms renders most conventional DFT methods unfeasible, and thus we computed the bond formations at the rim of the spoked-wheel complex at the ONIOM(TPSS:GFN2-xTB(GBSA)) level. The rim of the spoked-wheel complex consists of 18 5,15-linked porphyrin units coordinated by Zn(II) or Ni(II) ions. By the addition of bis(trifluoroacteoxy)iodobenzene (PIFA), the meso-meso coupled Zn-porphyrin rings can be fused further, forming a nanobelt involving three-fold single bond coupled Zn-porphyrin units. The corresponding reaction energy of this coupling reaction was computed including the Zinc-functionalized porphyrin rings in the inner region with 432 atoms. The obtained reaction energies are shown in Table E.1. The reference method yields an overall reaction energy of

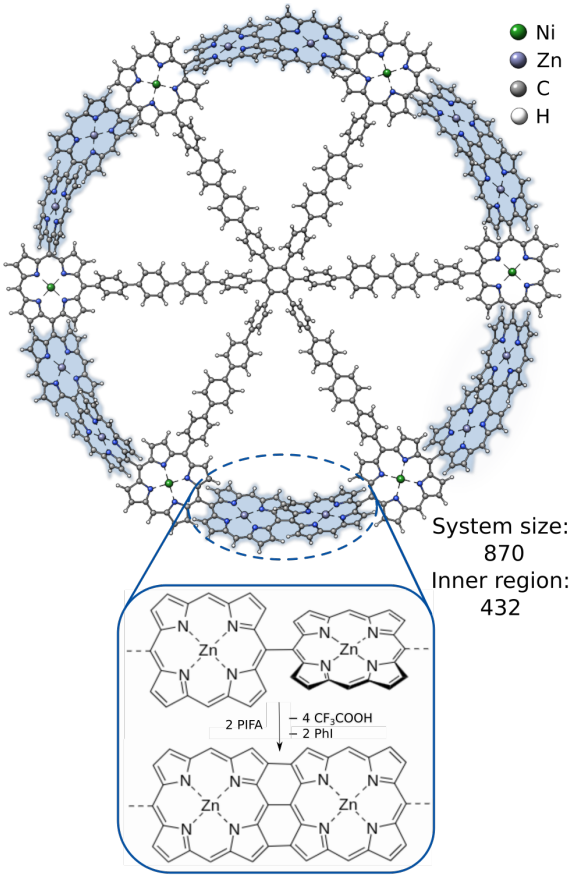

Figure E.7: The molecular structure of the 5,15-linked porphyrin nanoring and the fusion reaction between its Zn-functionalized porphyrin units.

Table E.1: The C-C coupling reaction energies of the 5,15-linked porphyrin nanoring implicitly solvated in toluene (ddCOSMO) and calculated onto the corresponding molecular geometries optimized at the same levels of theory with the corresponding wall time (28 cores) values for the self-consistent field (SCF) energy calculation. TPSS† = TPSS-D4/def2-SVP.

Methods	ΔE / kcal mol^{-1}	SCF time
ONIOM(TPSS†:GFN2-xTB)	−44.9	25 min
GFN2-xTB	−34.9	36 sec
ONIOM(GFN2-xTB:GFN-FF)	−24.4	4 sec
ONIOM(TPSS†:GFN1-xTB)	−47.0	24 min
GFN1-xTB	−37.6	39 sec
ONIOM(GFN1-xTB:GFN-FF)	−26.6	4 sec
ONIOM(TPSS†:GFN-FF)	−35.3	15 min
GFN-FF	172.6	0.4 sec

$-44.9\,\mathrm{kcal\,mol^{-1}}$. Using GFN2-xTB and GFN1-xTB yields reasonable reaction energies of -37.6 and $-34.9\,\mathrm{kcal\,mol^{-1}}$, respectively, and shows the qualitatively right trend. Using only GFN-FF yields a much too high, qualitatively wrong interaction energy. Combining it with GFN*n*-xTB within the ONIOM scheme can reduce this error drastically and gives a qualitatively correct behavior. These results show that it is possible to get good and fast results with the ONIOM(GFN*n*-xTB:GFN-FF) scheme and that if more accurate results are required for such large systems, ONIOM(DFT:GFN) can be a valuable tool as full DFT calculations are still too costly.

E.5 Conclusions

In the course of this work, the subtractive ONIOM scheme was implemented into the free, open-source *xtb* software package. Furthermore, several new implementation-specific features, such as an automatic charge and topology handling, were introduced. The utility of the ONIOM scheme in combination with the GFN family of semi-empirical and force-field methods was demonstrated exemplarily for geometry optimization and reaction energy evaluation of large metal-organic systems. Various combinations of DFT and the GFN methods were tested using the interface of *xtb* to the popular quantum chemistry packages *TURBOMOLE* and *ORCA*. It was shown that the ONIOM approach can be utilized to clearly improve on the already reasonable results of the GFN methods for challenging systems containing many transition metal atoms. By matching the strengths of the GFN methods and DFT, even critical cases that require a highly accurate quantum mechanical description can be treated. Accordingly, we demonstrated that by using the ONIOM framework in *xtb*, DFT-quality reaction energies and molecular structures can be obtained at a fraction of the computational cost of conventional DFT. This also allows the use of computationally more expensive and more accurate DFT methods. The efficient, robust, and easy-to-use implementation of the ONIOM scheme into *xtb* opens up new possibilities with regard to the treatment of very large systems containing a variety of elements across the periodic table.

Supporting information

The Supporting Information is available free of charge at
https://pubs.rsc.org/en/content/articlelanding/2023/cp/d3cp02178e.

- Details on the implementation, and detailed computational results (PDF)
- All relevant structures in xyz format (ZIP)

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The German Science Foundation (DFG) is gratefully acknowledged for financial support through a Gottfried Wilhelm Leibniz prize to S.G. and Merck KGaA. Further, S.G. and M.B. gratefully acknowledge financial support of the Max Planck Society through the Max Planck fellow program.

APPENDIX F

Toward Reliable Conformational Energies of Amino Acids and Dipeptides—The DipCONFS Benchmark and DipCONFL Datasets

Christoph Plett,* Stefan Grimme,* Andreas Hansen*

Reprinted (adapted) with permission‡ from:
C. Plett, S. Grimme, and A. Hansen, J. Chem. Theory Comput. **20** (2024) 8329.

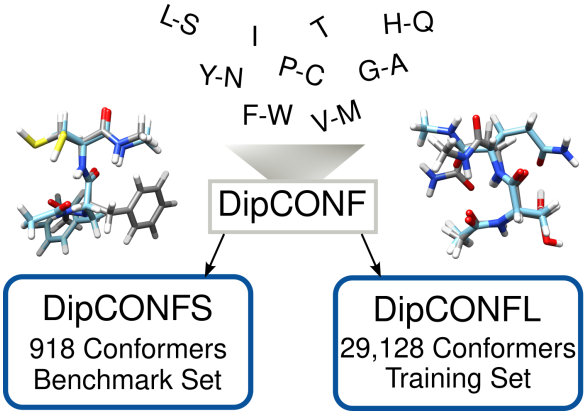

Figure F.1: Associated Table of Contents graphic for publication in Journal of Chemical Theory and Computation.

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

Abstract

Simulating peptides and proteins is becoming increasingly important, leading to a growing need for efficient computational methods. These are typically semi-empirical quantum mechanical (SQM) methods, force fields (FFs), or machine-learned interatomic potentials (MLIPs), all of which require a large amount of accurate data for robust training and evaluation. To assess potential reference methods and complement the available data, we introduce two sets, DipCONFL and DipCONFS, which cover large parts of the conformational space of 17 amino acids and their 289 possible dipeptides in aqueous solution. The conformers were selected from the exhaustive PeptideCs dataset by Andris et al. [J. Phys. Chem. B, **2022** *126*, 5949]. The structures, originally generated with GFN2-xTB, were re-optimized using the accurate r^{2}SCAN-3c density functional theory (DFT) composite method including the implicit CPCM water solvation model. The DipCONFS benchmark set contains 918 conformers and is one of the largest sets with highly accurate coupled cluster conformational energies so far. It is employed to evaluate various DFT and wave function theory (WFT) methods, especially regarding whether they are accurate enough to be used as reliable reference methods for larger datasets intended for training and testing more approximated methods like semi-empirical quantum mechanical (SQM) methods, force fields (FFs), and machine-learned interatomic potentials (MLIPs). The results reveal that the originally provided BP86-D3(BJ)/DGauss-DZVP conformational energies are not sufficiently accurate. Among the DFT methods tested as an alternative reference level, the revDSD-PBEP86-D4 double hybrid performed best with a mean absolute error (MAD) of 0.2 kcal mol^{-1} compared to the PNO-LCCSD(T)-F12b reference. The very efficient r^{2}SCAN-3c composite method also shows excellent results, with an MAD of 0.3 kcal mol^{-1} , similar to the best-tested hybrid ωB97M-D4. With these findings, we compiled the large DipCONFL set, which includes over 29,000 realistic conformers in solution with reasonably accurate r^{2}SCAN-3c reference conformational energies, gradients, and further properties potentially relevant for training MLIP methods. This set, also in comparison to DipCONFS, is used to assess the performance of various SQM, FF, and MLIP methods robustly and can complement training sets for those.

F.1 Introduction

Many biological and medical breakthroughs like the development of drugs and vaccines require understanding the role of proteins.[464,465] Thereby, elucidating the structure of a protein is often a crucial step and thus one of the most important fields of research.[25,466,467] Besides experimental techniques, computational chemistry and structural bioinformatics became more and more important for this task with different computational methods and workflows accelerating the discovery of protein structures.[468,469] They often rely on evaluating many different protein conformers to find the energetically most favorable (populated) one.[470] Due to the large size and conformational space of peptides and proteins, this leads to an immense computational demand excluding the application of accurate but costly methods like wave function theory (WFT) and density functional theory (DFT) for energy-based structure prediction of proteins.[471] Thus, more efficient methods like semi-empirical quantum mechanical (SQM) methods, force fields (FFs), or general functions for energy evaluation are typically used.[445,472,473] In addition, recent machine learning approaches like machine-learned interatomic potentials (MLIPs) showed promising results and received increasing attention.[86,474] However, these efficient methods typically include empirical approximations optimized on certain test

systems and hence rely on reference data to be trained and tested on. The generation of such data experimentally is quite elaborate,[475–477] but using reliable theoretical techniques can greatly increase the available reference data. As this requires the application of methods that are computationally too expensive for large systems like proteins, sets of smaller but comparable structures are used. Examples of conformation benchmarks are the YMPJ,[478] the MPCONF196,[76] and the 37conf8.[74] These sets are usually limited to a few molecules with 20-200 atoms and less than 1,000 conformers. Also, larger sets like the PEPCONF[479] with 755 molecules, but only six conformers per included di- and tripeptides, cannot nearly cover the whole conformational space of, e.g., the glutamine dipeptide with already tens of thousands of conformers.[221]. Another larger dataset specially designed for training MLIPs is the SPICE dataset supplemented with ωB97M-D3(BJ)[123,124,205]/def2-TZVPPD[105,106] properties.[224] Among other biologically important structures, it covers conformers of all possible dipeptides formed by the 20 natural amino acids and their common protonation states. The structures were generated with the Amber14 FF[480] with RDKit[481] in combination with molecular dynamic simulations and LBFGS energy minimization, resulting in 50 conformers per dipeptide evenly divided into high- and low-energy conformers. Recently, a computational study by Rulíšek et al. aimed to cover the whole conformational space of 20 natural amino acids and their respective dipeptides in aqueous solution, leading to the PeptideCs dataset with more than 400 million structures.[221] As generating and evaluating such a large amount of conformers is computationally extremely costly, this study was limited to SQM-optimized structures and computationally rather cheap GGA (BP86[208,209]-D3(BJ)/DGauss-DZVP[225]) conformational energies.

Here, we aim for a benchmark set that allows the selection of reliable reference methods for generating training and test sets for SQM, FF, and MLIP methods. For this purpose, we introduce the DipCONFS benchmark set, which is one of the largest sets with highly accurate coupled cluster conformational energies. It contains 918 conformers of all-natural amino acids and their respective dipeptides in neutral, non-zwitterionic states. We assess different DFT and WFT methods on the DipCONFS to evaluate whether they are accurate enough to be used as reference methods for much larger datasets. Further, we introduce the DipCOFNFL set including 29,128 r^{2}SCAN-3c[129](CPCM, water)[154] optimized conformers with accurate r^{2}SCAN-3c[129] properties to complement training and validation sets for SQM, FF, and MLIP methods like SPICE. Additionally, we employ it to evaluate commonly applied methods of these classes.

F.2 Methodology

F.2.1 Geometries

The DipCONF sets cover conformers of 17 neutral amino acids (Figure F.2) and their 289 possible dipeptides. Ionic states are excluded to avoid larger errors that may occur when employing simple methods to (zwitter)ions. The typical conformational energies for such species are rather different from those of neutrals and would require a separate benchmark study. For Histidin (H), both neutral states (protonated at either Nδ or Nϵ) are included. The N- and C-termini are always capped with acetyl or methylamine groups, respectively, to mimic the environment in a typical protein. The capping and dipeptide formation of Serine (S) and Proline (P) to form *S*, *P*, *S-P*, and *P-S* are exemplary depicted in Figure F.2. For compiling the DipCONFL set, we included 51 unique conformers per amino acid and 101 conformers per dipeptide from the PeptideCs dataset, if available. As the conformational space for some molecules is already exhausted when using 51 or 101 conformers,

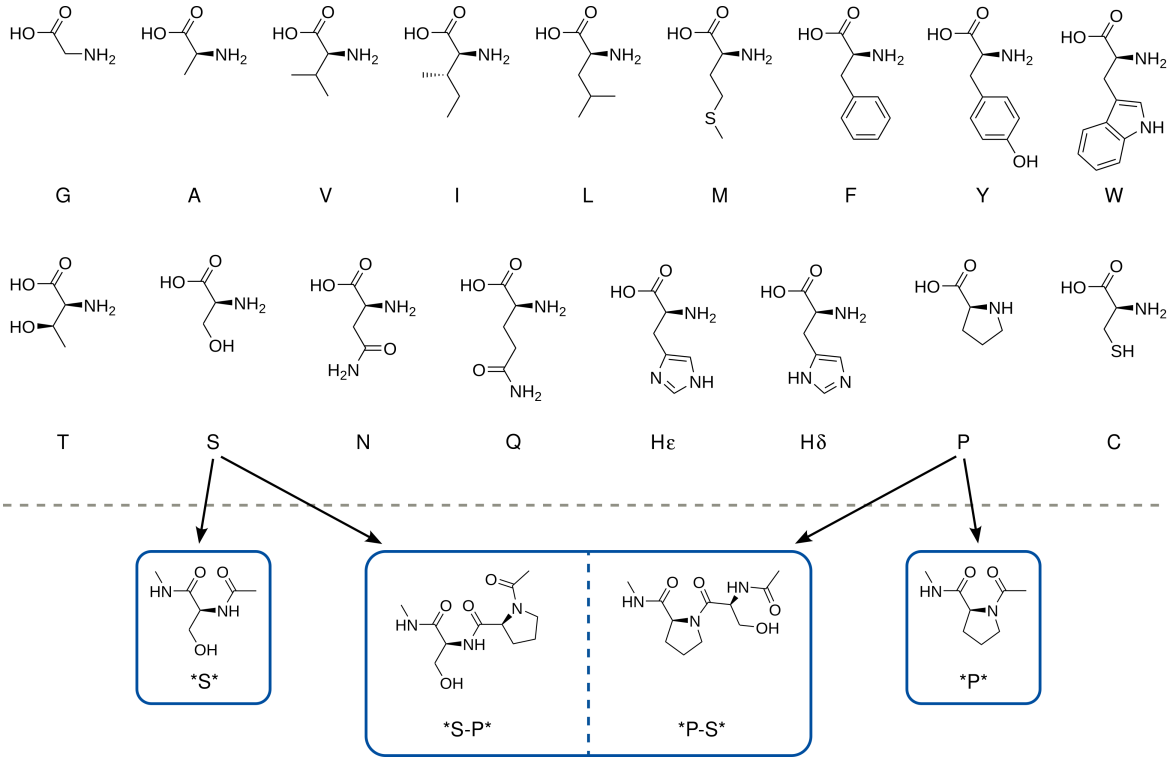

Figure F.2: All amino acids comprised in the DipCONF sets. Exemplary, for Serine (S) and Proline (P), the included structures of their monomers and the respective dimers are shown. The other amino acids were treated analogously.

respectively, including more conformers per structure would introduce an imbalance toward molecules with a larger conformational space. Further, to avoid introducing a bias into the structure selection, we included the originally lowest-energy structure of the PeptideCs database and randomly selected the other structures. Each selected structure was optimized with r^{2}SCAN-3c employing the CPCM water implicit solvation model as implemented in *ORCA* 5.0.4.[387] It has been shown that r^{2}SCAN-3c yields reliable results for conformational energies, and by employing the implicit solvation model for geometry optimizations, realistic structures of the molecules in aqueous solution are obtained. Afterward, identical conformers in the set were identified through the root-mean-square deviation of atoms excluding hydrogen (hRMSD). Structures with an hRMSD lower than 0.3 Å were replaced by a new r^{2}SCAN-3c(CPCM, water)-optimized structure from the PeptideCs dataset. This replacement was done until no identical conformers were left in the DipCONFL set. The only exceptions are two conformers of A-Q, P-S, Hδ-C, and Hϵ-C that showed (slightly) different intramolecular interaction motifs at this threshold, which were kept. Finally, r^{2}SCAN-3c reference electronic properties were computed. Solvation and thermo-statistical contributions are not part of this study, even though they can have a large impact on conformational free energies[482] and have to be included when simulating real systems. However, they are usually treated with separate models in common conformational screening workflows.[35] Besides yielding realistic structures in solution resembling typical applications, this procedure of providing gas-phase properties for structures optimized with the implicit solvent model bears another advantage: the DipCONFL conformers are not necessarily

minima on the gas-phase potential energy surface. Even though the gas-phase gradient norms only range from 16.0 kcal mol^{-1} bohr^{-1} to 56.2 kcal mol^{-1} bohr^{-1} (Supporting Information (SI), Table S2), exemplary optimizations (SI, Figure S2) show that a structure with a smaller gradient norm can almost be identical in the gas and solvent phase, while the geometries of molecules with higher norms can differ quite strongly. This suggests that the DipCONFL contains properties of structures partially close to gas-phase energy minima but also significantly differing, which is advantageous for robust training and validation of SQM, FF, or MLIP methods. Further, the unbiased selection from an exhaustive set of conformers together with the r^{2}SCAN-3c(CPCM, water) geometry optimizations differ from the structure generation of other large datasets that use SQM, FF, or MLIP methods and snapshots from MD simulations. Thus, the DipCONFL provides a unique set of structures that can be used to extend available sets like SPICE for training and validating SQM, FF, and MLIP methods. To compile the DipCONFS benchmark set, for which highly accurate PNO-LCCSD(T)-F12b/AVQZ'[195–197] are generated, three conformers per molecule were selected from the DipCONFL set: the energetically lowest, the second lowest, and one random conformer, adding up to 918 structures. The resulting energy contribution of the DipCONFS (SI, Figure S2) shows many conformers with a conformational energy of a few kcal mol^{-1} and a broad distribution covering higher conformational energies. This allows assessing a method's performance for two energetically close-lying low-energy conformers, important for correctly predicting the energetically lowest one, and for larger conformational energies selected in an unbiased (random) way. Both sets, together with their respective electronic properties, are available in the SI and on GitHub.[483] Besides xyz structures, also HDF5 files for both sets are provided, including the structures and respective electronic reference energies. Additionally, for the DipCONFL, gradients, dipole moments, atomization energies, Mulliken, and Hirshfeld partial charges are included in the HDF5 file. A detailed description of the data storage is available in the SI. Note that the atomization energies computed with (meta-)GGAs like r^{2}SCAN-3c may have larger errors than the conformational energies.[71]

F.2.2 Conformational Energies

All DFT and MP2 computations were performed with *ORCA* 5.0.4[387] employing default settings if not noted otherwise. For the r^{2}SCAN-3c geometry optimizations with the CPCM water solvation model, loose optimization thresholds were used to retain computational feasibility. Except for the composite-DFT, ωB97M-D3(BJ)/def2-TZVPPD, and BP86-D3(BJ)/DGauss-DZVP methods, the large def2-QZVPP basis set was employed with matching auxiliary basis sets for computing single-point conformational energies. The BP86-D3(BJ)/DGauss-DZVP computations were conducted analogously to the original PeptideCs dataset. The basis set was obtained from the basis set exchange database,[484] the auxiliary basis set was automatically generated,[485] and specially adjusted D3(BJ) parameters were taken from Ref. 486[226]. For the revDSD-PBEP86-D4[125,199] and ωB97X-D4[400] computations, refined D4 parameters were used.[130,487] The PM methods were employed with *MOPAC2016* version 19.179L accessed via an *xtb* interface. The GFN1-xTB, GFN2-xTB, and GFN-FF methods were applied as implemented in *xtb* 6.6.0.[84] DFTBD3 was employed using the D3(BJ) dispersion correction via *DFTB+*[488] version 22.2. All force fields except the GFN-FF were used as implemented in *OpenBabel* 2.3.1.[396] The ANI-2x MLIP[489] was accessed via the MLatom program.[490] AIMNet2[227] was used via the atomic simulation environment[491] with the "wb97m-d3_3" parameter set.

To generate highly accurate reference conformational energies for the DipCONFS set, we employed a state-of-the-art local CCSD(T) implementation with explicit correlation (PNO-LCCSD(T)-

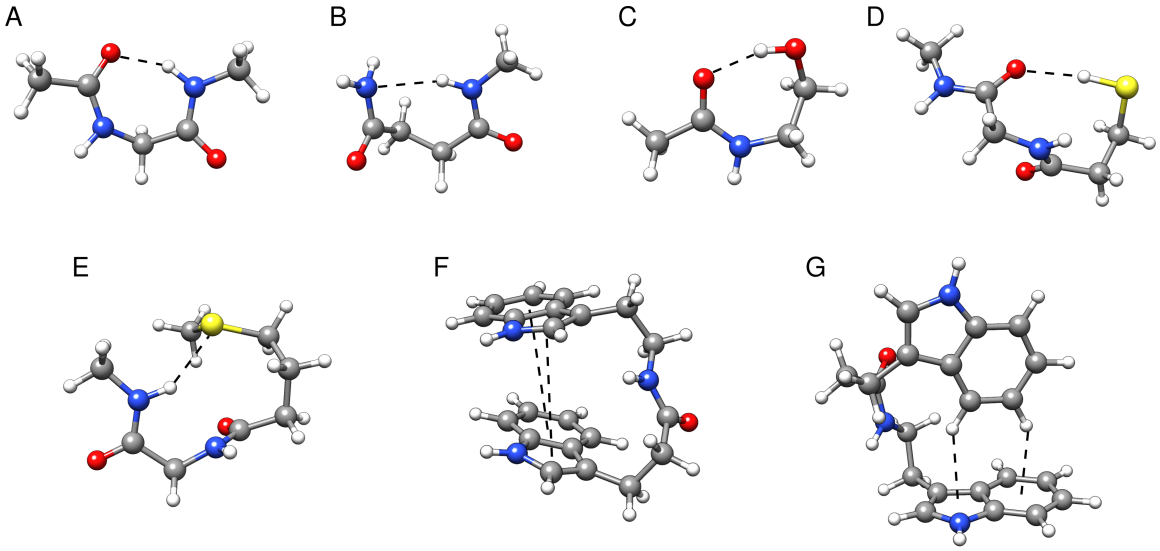

Figure F.3: Exemplary interaction motifs covered by the DipCONF sets. For clarity, only cutouts of structures from the DipCONFL were taken. Dangling bonds were capped with hydrogen atoms. H atoms are depicted white, C gray, N blue, O red, and S yellow.

F12b)[195–197] with `tight` domain settings as implemented in Molpro (2024.1 release)[392,393] together with the large aug-cc-pVQZ basis set[104] (cc-pVQZ for H to reduce the residual basis set superposition error (BSSE); hereafter abbreviated as AVQZ'). This computationally rather expensive reference level (about 1-2 days wall time (48 CPU processes) per conformer) reliably achieves a very small residual error of approximately 0.1 kcal mol^{-1} for conformational energies of organic compounds and was found to be superior to other explicitly correlated local coupled implementations in terms of efficiency, accuracy, and robustness.[198] This allows us to meaningfully evaluate the performance of also the best DFT methods for the usually quite small energy differences of the dipeptide conformers. Since the here introduced DipCONFS set comprises a much larger number of highly accurate conformer energies than in commonly used sets like the PCONF,[492] it may also be very helpful for developing (e.g., to fit parameters) and testing new approximate QC methods. All single-point energy calculations are carried out on the same r^{2}SCAN-3c(CPCM, water) optimized structures.

F.3 Interaction Motifs

The conformational energies of the DipCONF sets are dominated by intramolecular non-covalent interactions (NCIs). Especially, hydrogen bonds are crucial for the formation of different conformers as a typical H-bond in a protein has a strength of about 5-6 kcal mol^{-1} ,[228] but as there is a variety of different amino acids, also other NCI motifs occur. Examples of typical NCIs included in the DipCONF sets are shown in Figure F.3. The most frequently observed intramolecular interaction in the DipCONF sets is the N-H$\cdots$O hydrogen bond (Figure F.3 A). In principle, it can be formed purely by the backbone of every peptide. A second NCI, which is in principle also possible for every peptide, is the N-H$\cdots$N hydrogen bond (Figure F.3 B). It occurs less often in the DipCONF sets as the backbone strain usually does not allow for an optimal interaction. However, introducing additional

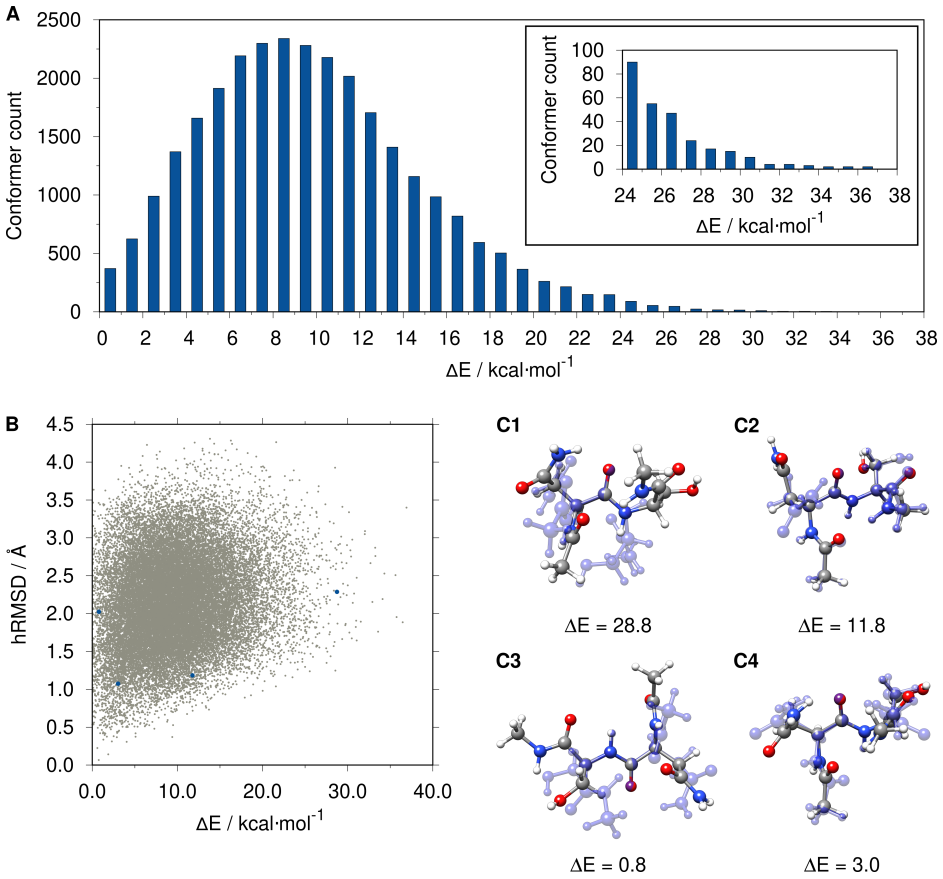

Figure F.4: (A) Conformational energy distribution of the DipCONFL set. Conformers were counted in 1 kcal mol^{-1} steps. (B) Correlation plot of the conformational energy and hRMSD with respect to the energetically lowest conformer. Structures C1-C4 are marked as blue dots. (C1)-(C4) Overlays of different conformers of *N-S* with relative energies in kcal mol^{-1}. H atoms are depicted white, C gray, N blue, O red, and S yellow.

amide or N-heterocyclic functional groups with amino acids like Asparagine, Glutamine, Tryptophan, or Histidine increases the possibility of finding such an interaction motif. Further hydrogen bonds with O-H, N-H, and S-H donors and varying acceptor atoms can occur if the respective amino acids are present. Some examples are shown in Figure F.3 C, D, and E. Despite hydrogen bonds, other NCI motifs like those of aromatic systems (Figure F.3 F, and G) are included if the peptide contains aromatic side chains. As various interaction motifs coupled with different backbone torsions are possible for a single structure, many different conformers with a large range of conformational energies are obtained.

To elucidate the connection of the conformational energy and the interaction motifs, Figure F.4 A depicts the conformational energy distribution of the DipCONFL, Figure F.4 B its correlation with the intra-ensemble hRMSD, and Figure F.4 C overlays selected conformers with their respective relative energies. The energy differences between the DipCONFL conformers range from 0 to 37 kcal mol^{-1} and correspond to a Poisson-type distribution with a maximum at about 8 kcal mol^{-1} (Figure F.4 A). From Figure F.4 B, it becomes obvious that the conformational energy correlates rather weakly with the

hRMSD. In fact, the largest hRMSDs of more than 4.0 Å lead to energy differences between only 5 and 20 kcal mol^{-1}, showing that a large hRMSD does not necessarily lead to a large energy difference. For example, two strongly deviating structures can either have completely different energies (e.g., Figure F.4, C1) if the amount and strength of NCIs differ, or they can have very similar energies (e.g., Figure F.4, C3) if different NCIs are formed that are similar in strength. On the other hand, a small hRMSD does not exclude large energy differences, even though the largest conformational energies (>30 kcal mol^{-1}) are not observed for hRMSDs smaller than 1.0 Å. Thus, a low hRMSD can lead to both large (e.g., Figure F.4, C2) or low (e.g., Figure F.4, C4) conformational energies. For example, a bond rotation only weakly affects the hRMSD, but changes the acceptor/donor of a hydrogen bond or even causes its loss. In passing, it is noted that also structures with fewer intermolecular interactions can become important as the additional unoccupied interaction sites can interact with solvent molecules. This can have a strong influence on the energy ranking, making sometimes higher-energy gas-phase conformers the energetically favored ones in solution [482].

F.4 Results and Discussion

In the following, different DFT and WFT methods are evaluated on the smaller DipCONFS benchmark set, and their suitability as a reference method for larger datasets like the extended DipCONFL dataset is examined. Subsequently, SQM, FF, and MLIP methods are assessed on both DipCONF sets, and their respective error sources are discussed.

F.4.1 DFT and WFT Performance

Comparing different DFT and WFT methods tested on the DipCONFS benchmark set (Table F.1 and Figure F.5), the trend of Jacobs ladder holds true,[119] even though rather accurate results can be achieved with methods of each tested class. The conformational energies of the evaluated WFT methods show that SCS-MP2 (MAD of 0.4 kcal mol^{-1}) improves on MP2 (MAD of 0.6 kcal mol^{-1}) but does not reach the best performers of the different DFT classes. The double hybrid revDSD-PBEP-D4/def2-QZVPP yields the overall lowest errors with an MAD of 0.2 kcal mol^{-1} and no outliers of more than 0.6 kcal mol^{-1}. The ωB97M(2)/def2-QZVPP double hybrid functional shows slightly larger errors (MAD of 0.3 kcal mol^{-1}), similar to the best tested range-separated hybrid ωB97M-D4/def2-QZVPP and its counterpart ωB97M-V/def2-QZVPP. The global hybrid method r^{2}SCAN0-D4/def2-QZVPP performs similarly well, while B3LYP-D4/def2-QZVPP yields more and larger outliers. M06-2X/def2-QZVPP shows generally slightly increased errors with the largest outlier of the tested hybrid methods (1.8 kcal mol^{-1}). Among the tested (meta-)GGAs, r^{2}SCAN-D4/def2-QZVPP performs best with a similar MAD to r^{2}SCAN0-D4/def2-QZVPP. Other methods like BP86-D4/def2-QZVPP and M06-L/def2-QZVPP show larger MADs of 0.6 kcal mol^{-1}. Evaluating functionals with other basis sets revealed that the ωB97M-D3(BJ)/def2-TZVPPD, used as the reference level for the SPICE dataset, also yields accurate conformational energies similar to ωB97M-D4/def2-QZVPP with an MAD of 0.3 kcal mol^{-1}. When using functionals in combination with significantly smaller basis sets, larger errors are observed. For example, the BP86-D3(BJ)/DGauss-DZVP method used as reference for the original PeptideCs set yields large errors for several cases (largest error of 3.5 kcal mol^{-1}) despite the tuned D3 parameters[226]. The MAD of 0.8 kcal mol^{-1} is, in view of the rather low mean conformational energy of DipCONFS ($\overline{\Delta E}$ of 5.3 kcal mol^{-1}), rather large.

Table F.1: Mean deviation (MD), mean absolute deviation (MAD), standard deviation (SD), and maximum absolute error (AMAX) in kcal mol^{-1} of the tested WFT/DFT methods on the DipCONFS sets. The reference level is PNO-LCCSD(T)-F12b/AVQZ'. Additionally, the Spearman correlation coefficients are given. Except for the composite "3c" methods, ωB97M-D3(BJ)* (* = def2-TZVPPD), and BP86-D3(BJ)** (** = DGauss-DZVP), all DFT and WFT methods were applied with the def2-QZVPP basis set. The mean conformational energy of the DipCONFS is 5.3 kcal mol^{-1} .

Method	MD	MAD	SD	AMAX	Spearman
SCS-MP2[408]	−0.17	0.39	0.46	1.69	0.913
MP2[116]	0.18	0.63	0.82	4.07	0.866
revDSD-PBEP86-D4[125,199]	−0.11	0.18	0.19	0.63	0.962
ωB97M(2)[398]	0.05	0.26	0.32	1.32	0.946
PWPB95-D4[399]	−0.07	0.28	0.36	1.27	0.941
ωB97M-D4[400]	−0.11	0.25	0.30	1.09	0.948
ωB97X-V[401]	−0.01	0.25	0.31	1.23	0.959
ωB97X-D4[130,400]	−0.14	0.26	0.30	1.23	0.951
ωB97M-V[402]	0.01	0.30	0.38	1.35	0.944
r^{2}SCAN0-D4[403]	0.17	0.30	0.35	1.25	0.966
PBE0-D4[203]	0.13	0.31	0.38	1.41	0.959
B3LYP-D4	−0.04	0.32	0.43	1.53	0.938
PW6B95-D4[202]	−0.06	0.36	0.46	1.6	0.937
M06-2X[204]	−0.05	0.36	0.46	1.83	0.918
r^{2}SCAN-D4[222]	0.05	0.26	0.34	1.4	0.948
B97M-V[205,206]	−0.12	0.34	0.42	1.68	0.943
PBE-D4[407]	0.00	0.43	0.56	2.01	0.93
BP86-D4[207–209]	0.10	0.61	0.78	2.62	0.897
M06-L[223]	−0.24	0.62	0.78	3.19	0.902
ωB97M-D3(BJ)*	0.02	0.28	0.36	1.27	0.944
BP86-D3(BJ)**	0.45	0.78	0.91	3.47	0.879
r^{2}SCAN-3c[129]	0.00	0.30	0.40	1.39	0.936
ωB97X-3c[130]	0.18	0.44	0.60	3.11	0.917
B97-3c[409]	0.12	0.48	0.61	2.44	0.907
PBEh-3c[210]	−0.31	0.89	1.09	4.26	0.825

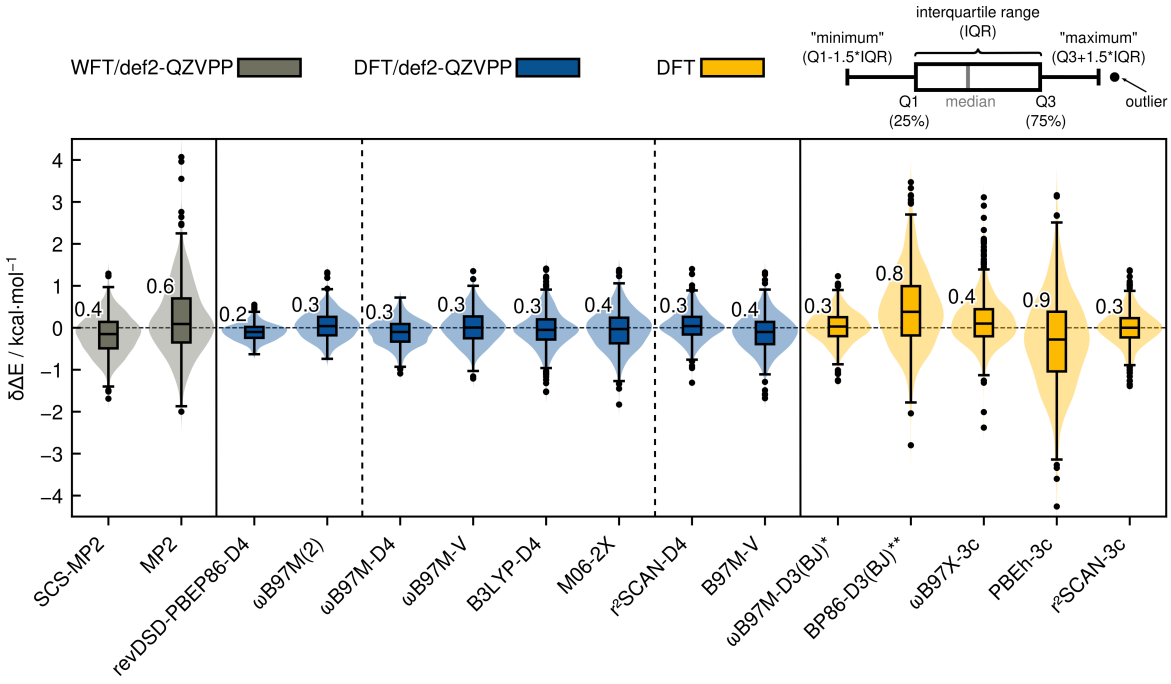

Figure F.5: Deviations of the conformational energies relative to the PNO-LCCSD(T)-F12b/AVQZ' reference depicted as box-plots and violin-plots including the MADs in kcal mol^{-1} . Only a representative selection of tested methods in Table F.1 is shown. Except for the "3c" methods, ωB97M-D3(BJ)* (* = def2-TZVPPD), and BP86-D3(BJ)** (**=DGauss-DZVP), the def2-QZVPP basis set was employed. The dashed lines separate double hybrids, hybrids, and (meta-)GGAs classes of functionals evaluated with this large basis set.

Considering this together with a Spearman coefficient below 0.9, this method is considered unsuitable as a reference method for a training set. Similarly, PBEh-3c with a modified def2-SVP basis set yields a large MAD of 0.9 kcal mol^{-1} with the largest maximum error of 4.3 kcal mol^{-1} . ωB97X-3c with its special valence double ζ (vDZP) basis set yields small basis set superposition errors (BSSEs) of about triple-ζ quality[130] leading to a lower MAD of 0.4 kcal mol^{-1} but still rather many and large outliers (maximum error of 3.1 kcal mol^{-1}). The composite method r^{2}SCAN-3c with a modified def2-TZVP basis set yields remarkably good results with an MAD of 0.3 kcal mol^{-1} similar to the best tested hybrid functionals in a larger basis set. Additionally, its vanishing MD on the DipCONFS makes it particularly suitable as an unbiased reference method. In principle, many of the tested DFT methods yield low MADs of 0.2-0.4 kcal mol^{-1} , MDs close to zero, and reasonable Spearman coefficients and are thus accurate enough for generating the reference energies for the DipCONFL set. However, not only the error but also the computational costs are crucial as almost 30,000 structures are included. To illustrate this, the wall times for a single *Q-W* conformer computed with the best performer of each class on four cores are shown in Table F.2. The WFT and double-hybrid methods are computationally the most expensive. They require about an hour of computation time for a single *Q-W* conformer. With ~45 min, also the ωB97M-D4 hybrid is practically impossible as the reference method for the DipCONFL set. With a significantly less computation time of ~7 min, r^{2}SCAN-D4/def2-QZVPP is more appropriate, but still about four times slower than r^{2}SCAN-3c, which shows similar performance. Thus, we choose r^{2}SCAN-3c as the reference method for the DipCONFL.

Table F.2: Absolute (in minutes) and relative wall times with respect to r^{2}SCAN-3c for computing the electronic energy of a *Q-W* conformer on four cores of an Intel® Xeon® CPU E3-1270 v5 @ 3.60GHz. Except for r^{2}SCAN-3c, the def2-QZVPP basis set was employed.

Method	Wall time	Relative wall time
SCS-MP2	57.2	35.8
revDSD-PBEP86-D4	61.8	38.6
wB97M-D4	44.5	27.8
r^{2}SCAN-D4	6.9	4.3
r^{2}SCAN-3c	1.6	1.0

Table F.3: Mean deviation (MD), mean absolute error (MAD), standard deviation (SD), and maximum absolute error (AMAX) in kcal mol^{-1} of the tested SQM, FF, and MLIP methods on the DipCONFL set. Additionally, the Spearman coefficients are shown. Values for the DipCONFS set are given in parentheses. The mean conformational energy of the DipCONFL is 9.8 kcal mol^{-1} (5.3 kcal mol^{-1} for the DipCONFS).

Method	MD	MAD	SD	AMAX	Spearman
PM7[134]	−0.03 (0.36)	1.90 (1.64)	2.38 (2.07)	11.27 (6.62)	0.406 (0.796)
GFN1-xTB[48]	−1.22 (−0.31)	2.16 (1.71)	2.41 (2.19)	12.16 (11.37)	0.337 (0.757)
GFN2-xTB[49]	−1.59 (−0.61)	2.25 (1.78)	2.30 (2.15)	10.72 (6.43)	0.318 (0.729)
PM6-D3H4X[133,410]	−0.88 (−0.08)	2.43 (2.06)	2.96 (2.65)	12.46 (10.02)	0.299 (0.732)
DFTB3[135]	−2.93 (−1.32)	3.11 (1.98)	2.27 (2.22)	13.96 (9.00)	0.120 (0.753)
MMFF94[211–215]	0.91 (0.72)	2.99 (2.48)	3.76 (3.12)	22.36 (13.34)	0.303 (0.663)
GFN-FF[50]	−2.42 (−1.35)	3.16 (2.42)	3.20 (2.97)	18.99 (11.77)	0.185 (0.675)
GAFF[411]	−4.14 (−2.43)	4.98 (3.74)	4.78 (4.47)	29.05 (19.72)	0.083 (0.503)
UFF[137]	−8.16 (−7.01)	19.55 (16.38)	26.04 (22.62)	170.54 (112.74)	0.039 (−0.065)
AIMNet2	−0.88 (−0.41)	1.55 (1.28)	1.72 (1.58)	8.47 (6.26)	0.473 (0.778)
ANI-2x	−0.88 (−0.16)	2.59 (2.17)	3.12 (2.76)	14.39 (9.23)	0.269 (0.708)

F.4.2 SQM, FF, and MLIP Performance

The performance of different SQM, FF, and MLIP methods is illustrated for the DipCONFS set in Figure F.6 A and for the DipCONFL set in Figure F.6 B. Respective statistical metrics are shown in Table F.3. An alternative evaluation of the DipCONFS set with respect to r^{2}SCAN-3c instead of PNO-LCCSD(T)-F12b/AVQZ' conformational energies is given in the supporting information, revealing a negligible influence on the error statistics of the tested SQM, FF, and MLIP methods.

When comparing the performance of different methods for the DipCONFS and DipCONFL sets, it is important to note that DipCONFL not only includes 30 times more conformers than DipCONFS but also has a higher mean conformational reference energy of 9.8 kcal mol^{-1} compared to 5.3 kcal mol^{-1} for DipCONFS. Higher MADs and SDs are observed on the DipCONFL for every tested method, and more and larger outliers occur as the additional conformers in the DipCONFL increase the probability of finding them. Moreover, the Spearman coefficients are lower for the DipCONFL compared to the DipCONS as more energetically close conformers are included, making it clearly more difficult to

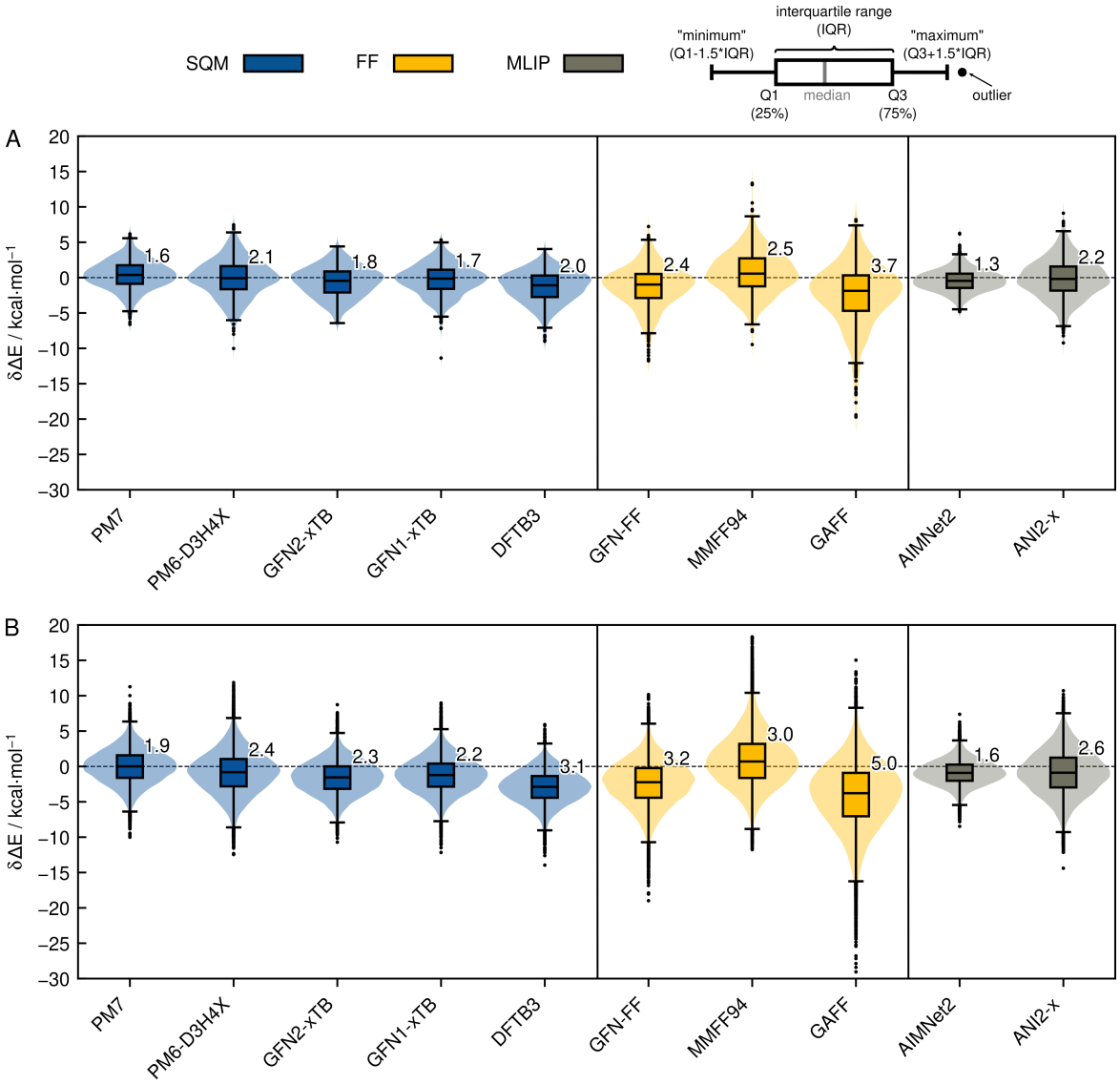

Figure F.6: Deviations of the DipCONFS conformational energies relative to the PNO-LCCSD(T)-F12b/AVQZ' (A) and the DipCONFL conformational energies relative to the r^{2}SCAN-3c reference depicted as boxplots and violinplots including the MADs in kcal mol^{-1} .

predict the correct ordering.

Overall, the tested SQM methods perform better than the FFs on both sets. PM7 is the best-performing SQM method with an MAD of 1.9 kcal mol^{-1} on the DipCONFL set (1.6 kcal mol^{-1} for DipCONFS) with an MD close to zero. The GFN*n*-xTB methods perform a little worse with MADs of about 2.2 kcal mol^{-1} for DipCONFL (1.7 kcal mol^{-1} and 1.8 kcal mol^{-1} , respectively, for DipCONFS). Surprisingly, this trend is different from the findings of other studies that deal with comparable conformers like MPCONF196[76] and 37Conf8,[74] where the GFN*n*-xTB methods clearly outperform PM7. Notably, the slightly negative MDs of both GFN*n*-xTB methods on the DipCONFS are about 1 kcal mol^{-1} more negative for the DipCONFL, which is particularly caused by

the dipeptides containing aromatic systems. The DFTB3 method is, especially for DipCONFL, the worst-tested SQM method with an MAD of 3.1 kcal mol^{-1} , which is much larger than for DipCONFS (2.0 kcal mol^{-1}). Thereby, the error ranges of DipCONFS and DipCONFL (9.0 kcal mol^{-1} and 14.0 kcal mol^{-1}) are comparable to PM7 (6.6 kcal mol^{-1} and 11.3 kcal mol^{-1}). However, the MD of DFTB3 is much lower for DipCONFL (-2.9 kcal mol^{-1}) than for the DipCONFS (-1.3 kcal mol^{-1}). In the case of the FFs, MMFF94 with an MAD of 3.0 kcal mol^{-1} for DipCONFL and GFN-FF (MAD of 3.2 kcal mol^{-1}) performs better than GAFF (MAD of 5.0 kcal mol^{-1}). However, all of these FFs show rather large outliers with maximum errors of about 20 kcal mol^{-1} and more, which is about half of the highest conformational energy in the set (37 kcal mol^{-1}). Regarding the MLIPs, AIMNet2 yields the lowest MAD of 1.6 kcal mol^{-1} on the DipCONFL set of all tested non-DFT/WFT methods (1.3 kcal mol^{-1} for DipCONFS). Together with the smallest maximum errors (8.5 kcal mol^{-1} for DipCONFL and 6.3 kcal mol^{-1} on the DipCONFS), it can be regarded as the best performer in this class, working robustly for both sets. The ANI-2x potential shows almost twice as large errors and cannot compete with the best of the tested SQM methods.

F.5 Conclusion

We introduced the DipCONFS benchmark and the DipCONFL dataset, which cover a broad range of the conformational space of 17 amino acids and their 289 possible dipeptides. The structures were selected from the exhaustive PeptideCs set and refined using r^{2}SCAN-3c(CPCM, water). A key feature of both sets is the broad variety of interaction motifs leading to substantial conformational energies of up to 20-30 kcal/mol. The DipCONFS set, comprising 918 conformers, is the largest benchmark set of this type so far featuring highly accurate PNO-LCCSD(T)-F12b/AVQZ' conformational energies thus allowing to assess even the best DFT methods meaningfully. We evaluated various DFT and WFT methods on this set and determined their suitability as reliable reference methods for training and testing more approximate methods like SQMs, FFs, and MLIPs. Except for the very accurate but rather expensive revDSDPBEP86-D4/def2-QZVPP double hybrid (MAD of 0.2 kcal mol^{-1}), the efficient r^{2}SCAN-3c composite method (MAD of 0.3 kcal mol^{-1}) competes with the best tested DFT methods while being more than an order of magnitude computationally faster. Other methods like ωB97M-D3(BJ)/def2-TZVPPD used to generate reference properties in datasets like SPICE also proved to be very accurate, while BP86-D3(BJ)/DGauss-DZVP previously used to generate the huge amount of conformational energies composed in the PeptideCs dataset showed larger errors of up to 3.5 kcal mol^{-1} with an MAD of 0.8 kcal mol^{-1} making them less suitable for generating such datasets. The second dataset, termed DipCONFL, includes 29,128 DFT-optimized conformers with reliable r^{2}SCAN-3c gas-phase conformational energies, gradients, and other properties relevant for training more approximate methods. These conformers, realistic in solution, are partly similar to gas-phase energy minima but also significantly different, resulting in a diverse set of properties. When comparing the performance of common methods of these classes on both sets, we generally observe larger errors for the DipCONFL set compared to the DipCONFS set, with more and larger outliers. The best-tested methods, such as PM7 and AIMNet2, yield reasonable results on both sets, while others like DFTB3, GAFF, and ANI-2x show much larger MADs for the DipCONFL compared to the DipCONFS set. Further, the set of unique structures composed in DipCONFL can be used to extend existing datasets like SPICE to facilitate a more robust training and validation of SQM, FF, and MLIP methods. All structures and properties are available as SI or from GitHub.[483]

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.4c00801.

- Definition of statistical error measures, evaluation of the DipCONFS structures relative to r^2SCAN-3c conformational energies, reference energy distribution of the DipCONFS set, gradient evaluation of the DipCONFL set, and information about the data storage (PDF)
- The DipCONFL set including r2SCAN-3c electronic properties in the HDF5 format (ZIP)
- The DipCONFS set including PNO-LCCSD(T)-F12b/AVQZ' conformational energies in the HDF5 format (ZIP)
- Conformational energies of the reference and all tested methods on the DipCONFL (XLSX)
- Conformational energies of the reference and all tested methods on the DipCONFS (XLSX)
- Structures of the DipCONFL set in xyz format (ZIP)
- Structures of the DipCONFS set in xyz format (ZIP)

Conflicts of Interest

There are no conflicts to declare.

Acknowledgement

This work was financially supported by Merck KGaA. The authors thank T. Gasevic and C. Hölzer for helpful discussions.

Bibliography

[1] M. Cavalleri, *Quantum chemistry reloaded*, Int. J. Quantum Chem. **113** (2013) 1, DOI: 10.1002/QUA.24364.

[2] S. Grimme and P. R. Schreiner, *Computational Chemistry: The Fate of Current Methods and Future Challenges*, Angew. Chem. Int. Ed. **57** (2018) 4170, DOI: 10.1002/ANIE.201709943.

[3] T. J. Marrone, J. M. Briggs, and J. A. McCammon, *Structure-based drug design: Computational advances*, Annu. Rev. Pharmacol. Toxicol. **37** (1997) 71, DOI: 10.1146/annurev.pharmtox.37.1.71.

[4] V. Vaissier Welborn and T. Head-Gordon, *Computational Design of Synthetic Enzymes*, Chem. Rev. **119** (2019) 6613, DOI: 10.1021/acs.chemrev.8b00399.

[5] G. Sliwoski, S. Kothiwale, J. Meiler, and E. W. Lowe, *Computational methods in drug discovery*, Pharmacol. Rev. **66** (2014) 334, DOI: 10.1124/pr.112.007336.

[6] A. Hillisch, N. Heinrich, and H. Wild, *Computational Chemistry in the Pharmaceutical Industry: From Childhood to Adolescence*, ChemMedChem **10** (2015) 1958, DOI: 10.1002/CMDC.201500346.

[7] S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, *The high-throughput highway to computational materials design*, Nat. Mater. **12** (2013) 191, DOI: 10.1038/nmat3568.

[8] J. Hafner, C. Wolverton, and G. Ceder, *Toward computational materials design: The impact of density functional theory on materials research*, MRS Bull. **31** (2006) 659, DOI: 10.1557/mrs2006.174.

[9] C. Poree and F. Schoenebeck, *A holy grail in chemistry: Computational catalyst design: Feasible or fiction?*, Acc. Chem. Res. **50** (2017) 605, DOI: 10.1021/acs.accounts.6b00606.

[10] S. Ahn, M. Hong, M. Sundararajan, D. H. Ess, and M. H. Baik, *Design and Optimization of Catalysts Based on Mechanistic Insights Derived from Quantum Chemical Reaction Modeling*, Chem. Rev. **119** (2019) 6509, DOI: 10.1021/acs.chemrev.9b00073.

[11] B. W. Chen, L. Xu, and M. Mavrikakis, *Computational Methods in Heterogeneous Catalysis*, Chem. Rev. **121** (2021) 1007, DOI: 10.1021/acs.chemrev.0c01060.

[12] J. Moult, *The current state of the art in protein structure prediction*, Curr. Opin. Biotechnol. **7** (1996) 422, DOI: 10.1016/S0958-1669(96)80118-2.

[13] D. H. Bowskill, I. J. Sugden, S. Konstantinopoulos, C. S. Adjiman, and C. C. Pantelides, *Crystal Structure Prediction Methods for Organic Molecules: State of the Art*, Annu. Rev. Chem. Biomol. Eng. **12** (2021) 593, DOI: 10.1146/ANNUREV-CHEMBIOENG-060718-030256.

[14] C. A. Bergström and P. Larsson, *Computational prediction of drug solubility in water-based systems: Qualitative and quantitative approaches used in the current drug discovery and development setting*, Int. J. Pharm. **540** (2018) 185, DOI: 10.1016/J.IJPHARM.2018.01.044.

[15] R. Fujiki, T. Matsui, Y. Shigeta, H. Nakano, and N. Yoshida, *Recent Developments of Computational Methods for pKa Prediction Based on Electronic Structure Theory with Solvation Models*, J **4** (2021) 849, DOI: 10.3390/J4040058.

[16] Y. P. Chin, N. W. See, I. D. Jenkins, and E. H. Krenske, *Computational discoveries of reaction mechanisms: recent highlights and emerging challenges*, Org. Biomol. Chem. **20** (2022) 2028, DOI: 10.1039/D1OB02139G.

[17] P. Gatt, R. Stranger, and R. J. Pace, *Application of computational chemistry to understanding the structure and mechanism of the Mn catalytic site in photosystem II – A review*, J. Photochem. Photobiol. B: Biol. **104** (2011) 80, DOI: 10.1016/J.JPHOTOBIOL.2011.02.008.

[18] X. Liu, R. Wang, X. Wang, and D. Xu, *High-Throughput Computational Screening and Machine Learning Model for Accelerated Metal-Organic Frameworks Discovery in Toluene Vapor Adsorption*, J. Phys. Chem. C **127** (2023) 11268, DOI: 10.1021/acs.jpcc.3c01749.

[19] X. Lin, X. Li, and X. Lin, *A Review on Applications of Computational Methods in Drug Screening and Design*, Molecules **25** (2020) 1375, DOI: 10.3390/MOLECULES25061375.

[20] D. J. Wales and T. V. Bogdan, *Potential energy and free energy landscapes*, J. Phys. Chem. B **110** (2006) 20765, DOI: 10.1021/JP0680544.

[21] G. P. Moss, *Basic terminology of stereochemistry (IUPAC Recommendations 1996)*, Pure Appl. Chem. **68** (1996) 2193, DOI: 10.1351/PAC199668122193.

[22] H. B. Schlegel, *Exploring potential energy surfaces for chemical reactions: An overview of some practical methods*, J. Comput. Chem. **24** (2003) 1514, DOI: 10.1002/JCC.10231.

[23] P. L. Gentili, *The Conformational Contribution to Molecular Complexity and Its Implications for Information Processing in Living Beings and Chemical Artificial Intelligence*, Biomimetics **9** (2024) 121, DOI: 10.3390/BIOMIMETICS9020121.

[24] S. P. Jarvis, S. Taylor, J. D. Baran, N. R. Champness, J. A. Larsson, and P. Moriarty, *Measuring the mechanical properties of molecular conformers*, Nat. Commun. **6** (2015) 1, DOI: 10.1038/ncomms9338.

[25] H. Hegyi and M. Gerstein, *The relationship between protein structure and function: a comprehensive survey with application to the yeast genome*, J. Mol. Biol. **288** (1999) 147, DOI: 10.1006/JMBI.1999.2661.

[26] Y. Zhou, I. Limbu, M. J. Garson, and E. H. Krenske, *Conformational Sampling in Computational Studies of Natural Products: Why Is It Important?*, J. Nat. Prod. **87** (2024) 2543, DOI: 10.1021/ACS.JNATPROD.4C00852.

[27] E. S. Ameh, *A review of basic crystallography and x-ray diffraction applications*, J. Adv. Manuf. Technol. **105** (2019) 3289, DOI: 10.1007/S00170-019-04508-1.

[28] M. Alberts, T. Laino, and A. C. Vaucher, *Leveraging infrared spectroscopy for automated structure elucidation*, Commun. Chem. **7** (2024) 1, DOI: 10.1038/s42004-024-01341-w.

[29] D. C. Burns and W. F. Reynolds, *Review of Optimizing NMR Methods for Structure Elucidation: Characterizing Natural Products and Other Organic Compounds (New Developments in NMR)*, J. Nat. Prod. **83** (2020) 3764, DOI: 10.1021/ACS.JNATPROD.0C01046.

[30] P. C. Hawkins, *Conformation Generation: The State of the Art*, J. Chem. Inf. Model. **57** (2017) 1747, DOI: 10.1021/ACS.JCIM.7B00221.

[31] A. T. McNutt, F. Bisiriyu, S. Song, A. Vyas, G. R. Hutchison, and D. R. Koes, *Conformer Generation for Structure-Based Drug Design: How Many and How Good?*, J. Chem. Inf. Model. **63** (2023) 6598, DOI: 10.1021/acs.jcim.3c01245.

[32] S. Taherivardanjani, J. Blasius, M. Brehm, R. Dötzer, and B. Kirchner, *Conformer Weighting and Differently Sized Cluster Weighting for Nicotine and Its Phosphorus Derivatives*, J. Phys. Chem. A **126** (2022) 7070, DOI: 10.1021/ACS.JPCA.2C03133.

[33] B. de Souza, *GOAT: A Global Optimization Algorithm for Molecules and Atomic Clusters*, Angew. Chem. Int. Ed. (2025) e202500393, DOI: 10.1002/ANIE.202500393.

[34] S. J. Cyvin et al., *Staggered conformers of alkanes: complete solution of the enumeration problem*, J. Mol. Struct. **413-414** (1997) 227, DOI: 10.1016/S0022-2860(97)00025-2.

[35] S. Grimme, F. Bohle, A. Hansen, P. Pracht, S. Spicher, and M. Stahn, *Efficient Quantum Chemical Calculation of Structure Ensembles and Free Energies for Nonrigid Molecules*, J. Phys. Chem. A **125** (2021) 4039, DOI: 10.1021/acs.jpca.1c00971.

[36] M. Bursch, J.-M. Mewes, A. Hansen, and S. Grimme, *Best-Practice DFT Protocols for Basic Molecular Computational Chemistry*, Angew. Chem. **134** (2022) e202205735, DOI: 10.1002/ANGE.202205735.

[37] R. Laplaza, M. D. Wodrich, and C. Corminboeuf, *Overcoming the Pitfalls of Computing Reaction Selectivity from Ensembles of Transition States*, J. Phys. Chem. Lett. **15** (2024) 7363, DOI: 10.1021/ACS.JPCLETT.4C01657.

[38] J. C. Zapata Trujillo, M. M. Pettyjohn, and L. K. Mckemmish, *High-throughput quantum chemistry: empowering the search for molecular candidates behind unknown spectral signatures in exoplanetary atmospheres*, Mon. Not. R. Astron. Soc. **524** (2023) 361, DOI: 10.1093/MNRAS/STAD1717.

[39] P. Zhou, J. Huang, and F. Tian, *Specific Noncovalent Interactions at Protein-Ligand Interface: Implications for Rational Drug Design*, Curr. Med. Chem. **19** (2012) 226, DOI: 10.2174/092986712803414150.

[40] X. Du et al., *Insights into Protein–Ligand Interactions: Mechanisms, Models, and Methods*, Int. J. Mol. Sci. **17** (2016) 144, DOI: 10.3390/IJMS17020144.

[41] A. Haque, K. M. Alenezi, M. S. Khan, W. Y. Wong, and P. R. Raithby, *Non-covalent interactions (NCIs) in π-conjugated functional materials: advances and perspectives*, Chem. Soc. Rev. **52** (2023) 454, DOI: 10.1039/D2CS00262K.

[42] K. T. Mahmudov, M. N. Kopylovich, M. F. C. Guedes da Silva, and A. J. Pombeiro, *Non-covalent interactions in the synthesis of coordination compounds: Recent advances*, Coord. Chem. Rev. **345** (2017) 54, DOI: 10.1016/J.CCR.2016.09.002.

[43] K. Müller-Dethlefs and P. Hobza, *Noncovalent Interactions: A Challenge for Experiment and Theory*, Chem. Rev. **100** (2000) 143, DOI: 10.1021/CR9900331.

[44] P. Pracht, F. Bohle, and S. Grimme, *Automated exploration of the low-energy chemical space with fast quantum chemical methods*, Phys. Chem. Chem. Phys. **22** (2020) 7169, DOI: 10.1039/c9cp06869d.

[45] N. J. King, I. D. LeBlanc, and A. Brown, *A variant on the CREST iMTD algorithm for noncovalent clusters of flexible molecules*, J. Comput. Chem. **45** (2024) 2431, DOI: 10.1002/JCC.27458.

[46] P. Hobza and K. Müller-Dethlefs, *Non-covalent Interactions: Theory and Experiment*, Theoretical and Computational Chemistry Series, Cambridge: The Royal Society of Chemistry, 2009, ISBN: 978-1-84755-853-4, DOI: 10.1039/9781847559906.

[47] S. Grimme, C. Bannwarth, E. Caldeweyher, J. Pisarek, and A. Hansen, *A general intermolecular force field based on tight-binding quantum chemical calculations*, J. Chem. Phys. **147** (2017) 161708, DOI: 10.1063/1.4991798.

[48] S. Grimme, C. Bannwarth, and P. Shushkov, *A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1–86)*, J. Chem. Theory Comput. **13** (2017) 1989, DOI: 10.1021/ACS.JCTC.7B00118.

[49] C. Bannwarth, S. Ehlert, and S. Grimme, *GFN2-xTB - An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions*, J. Chem. Theory Comput. **15** (2019) 1652, DOI: 10.1021/ACS.JCTC.8B01176.

[50] S. Spicher and S. Grimme, *Robust Atomistic Modeling of Materials, Organometallic, and Biochemical Systems*, Angew. Chem. Int. Ed. **59** (2020) 15665, DOI: 10.1002/ANIE.202004239.

[51] X.-Y. Meng, H.-X. Zhang, M. Mezei, and M. Cui, *Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery*, Curr. Comput.-Aided Drug Des. **7** (2012) 146, DOI: 10.2174/157340911795677602.

[52] N. S. Pagadala, K. Syed, and J. Tuszynski, *Software for molecular docking: a review*, Biophys. Rev. **9** (2017) 91, DOI: 10.1007/S12551-016-0247-1.

[53] B. J. Bender et al., *A practical guide to large-scale docking*, Nat. Protoc. **16** (2021) 4799, DOI: 10.1038/s41596-021-00597-z.

[54] C. Sepali, S. Gómez, E. Grifoni, T. Giovannini, and C. Cappelli, *Computational Spectroscopy of Aqueous Solutions: The Underlying Role of Conformational Sampling*, J. Phys. Chem. B **128** (2024) 5083, DOI: 10.1021/ACS.JPCB.4C01443.

[55] J. N. Dahanayake and K. R. Mitchell-Koch, *How does solvation layer mobility affect protein structural dynamics?*, Front. Mol. Biosci. **5** (2018) 370518, DOI: 10.3389/FMOLB.2018.00065.

[56] R. Narayan and R. Dominko, *Fluorinated solvents for better batteries*, Nat. Rev. Chem. **6** (2022) 449, DOI: 10.1038/s41570-022-00387-5.

[57] A. N. Paparella et al., *Use of Deep Eutectic Solvents in Plastic Depolymerization*, Catalysts **13** (2023) 1035, DOI: 10.3390/CATAL13071035.

[58] A. Bose Majumdar, I. J. Kim, and H. Na, *Effect of solvent on protein structure and dynamics*, Phys. Biol. **17** (2020) 036006, DOI: 10.1088/1478-3975.

[59] H. Svith et al., *On the nature of solvent effects on redox properties*, J. Phys. Chem. A **108** (2004) 4805, DOI: 10.1021/JP031268Q.

[60] A. Allerhand and P. R. Von Schleyer, *Solvent Effects in Infrared Spectroscopic Studies of Hydrogen Bonding*, J. Am. Chem. Soc. **85** (1963) 371, DOI: 10.1021/JA00887A001.

[61] C. J. Burrows, J. B. Harper, W. Sander, and D. J. Tantillo, *Solvation Effects in Organic Chemistry*, J. Org. Chem. Res. **87** (2022) 1599, DOI: 10.1021/ACS.JOC.1C03148.

[62] J. J. Varghese and S. H. Mushrif, *Origins of complex solvent effects on chemical reactivity and computational tools to investigate them: a review*, React. Chem. Eng. **4** (2019) 165, DOI: 10.1039/C8RE00226F.

[63] M. Bensberg, P. L. Türtscher, J. P. Unsleber, M. Reiher, and J. Neugebauer, *Solvation Free Energies in Subsystem Density Functional Theory*, J. Chem. Theory Comput. **18** (2022) 723, DOI: 10.1021/ACS.JCTC.1C00864.

[64] J. R. Pliego and J. M. Riveros, *The cluster-continuum model for the calculation of the solvation free energy of ionic species*, J. Phys. Chem. A **105** (2001) 7241, DOI: 10.1021/JP004192W.

[65] W. Wu and J. Kieffer, *New Hybrid Method for the Calculation of the Solvation Free Energy of Small Molecules in Aqueous Solutions*, J. Chem. Theory Comput. **15** (2019) 371, DOI: 10.1021/ACS.JCTC.8B00615.

[66] C. Plett, *Quantum Cluster Growth: An Automated Description of Explicit Solvation by Force-Field and Tight-Binding Methods*, MA thesis: Rheinischen Friedrich Wilhelms Universität Bonn, 2021.

[67] D. van der Spoel, J. Zhang, and H. Zhang, *Quantitative predictions from molecular simulations using explicit or implicit interactions*, Wiley Interdiscip. Rev. Comput. Mol. Sci. **12** (2022) e1560, DOI: 10.1002/WCMS.1560.

[68] J. Zhang, H. Zhang, T. Wu, Q. Wang, and D. Van Der Spoel, *Comparison of Implicit and Explicit Solvent Models for the Calculation of Solvation Free Energy in Organic Solvents*, J. Chem. Theory Comput. **13** (2017) 1034, DOI: 10.1021/ACS.JCTC.7B00169.

[69] C. J. Fennell and K. A. Dill, *Physical Modeling of Aqueous Solvation*, J. Stat. Phys. **145** (2011) 209, DOI: 10.1007/S10955-011-0232-9.

[70] N. Mardirossian and M. Head-Gordon, *Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals*, Mol. Phys. **115** (2017) 2315, DOI: 10.1080/00268976.2017.1333644.

[71] L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi, and S. Grimme, *A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions*, Phys. Chem. Chem. Phys. **19** (2017) 32184, DOI: 10.1039/C7CP04913G.

[72] M. Marianski, A. Supady, T. Ingram, M. Schneider, and C. Baldauf, *Assessing the Accuracy of Across-the-Scale Methods for Predicting Carbohydrate Conformational Energies for the Examples of Glucose and α-Maltose*, J. Chem. Theory Comput. **12** (2016) 6157, DOI: 10.1021/ACS.JCTC.6B00876.

[73] M. Bursch, A. Hansen, P. Pracht, J. T. Kohn, and S. Grimme, *Theoretical study on conformational energies of transition metal complexes*, Phys. Chem. Chem. Phys. **23** (2021) 287, DOI: 10.1039/D0CP04696E.

[74] D. I. Sharapa, A. Genaev, L. Cavallo, and Y. Minenkov, *A Robust and Cost-Efficient Scheme for Accurate Conformational Energies of Organic Molecules*, ChemPhysChem **20** (2019) 92, DOI: 10.1002/CPHC.201801063.

[75] S. Ehlert, S. Grimme, and A. Hansen, *Conformational Energy Benchmark for Longer n-Alkane Chains*, J. Phys. Chem. A **126** (2022) 3521, DOI: 10.1021/ACS.JPCA.2C02439.

[76] J. Řezáč, D. Bím, O. Gutten, and L. Rulíšek, *Toward Accurate Conformational Energies of Smaller Peptides and Medium-Sized Macrocycles: MPCONF196 Benchmark Energy Data Set*, J. Chem. Theory Comput. **14** (2018) 1254, DOI: 10.1021/ACS.JCTC.7B01074.

[77] J. Řezáč, *Non-Covalent Interactions Atlas Benchmark Data Sets: Hydrogen Bonding*, J. Chem. Theory Comput. **16** (2020) 2355, DOI: 10.1021/ACS.JCTC.9B01265.

[78] J. Řezáč, *Non-Covalent Interactions Atlas Benchmark Data Sets 2: Hydrogen Bonding in an Extended Chemical Space*, J. Chem. Theory Comput. **16** (2020) 6305, DOI: 10.1021/ACS.JCTC.0C00715.

[79] K. Kříž, M. Nováček, and J. Řezáč, *Non-Covalent Interactions Atlas Benchmark Data Sets 3: Repulsive Contacts*, J. Chem. Theory Comput. **17** (2021) 1548, DOI: 10.1021/ACS.JCTC.0C01341.

[80] K. Kříž and J. Řezáč, *Non-covalent interactions atlas benchmark data sets 4: σ-hole interactions*, Phys. Chem. Chem. Phys. **24** (2022) 14794, DOI: 10.1039/D2CP01600A.

[81] V. S. Bryantsev, M. S. Diallo, A. C. Van Duin, and W. A. Goddard, *Evaluation of B3LYP, X3LYP, and M06-Class density functionals for predicting the binding energies of neutral, protonated, and deprotonated water clusters*, J. Chem. Theory Comput. **5** (2009) 1016, DOI: 10.1021/CT800549F.

[82] B. Temelso, K. A. Archer, and G. C. Shields, *Benchmark structures and binding energies of small water clusters with anharmonicity corrections*, J. Phys. Chem. A **115** (2011) 12034, DOI: 10.1021/JP2069489.

[83] L. W. Chung et al., *The ONIOM Method and Its Applications*, Chem. Rev. **115** (2015) 5678, DOI: 10.1021/CR5004419.

[84] C. Bannwarth et al., *Extended tight-binding quantum chemistry methods*, Wiley Interdiscip. Rev. Comput. Mol. Sci. **11** (2021) e1493, DOI: 10.1002/WCMS.1493.

[85] A. Katbashev et al., *Overview on Building Blocks and Applications of Efficient and Robust Extended Tight Binding*, J. Phys. Chem. A **129** (2025), DOI: 10.1021/acs.jpca.4c08263.

[86] J. Jumper et al., *Highly accurate protein structure prediction with AlphaFold*, Nature **596** (2021) 583, DOI: 10.1038/s41586-021-03819-2.

[87] B. G. del Rio, B. Phan, and R. Ramprasad, *A deep learning framework to emulate density functional theory*, Npj Comput. Mater. **9** (2023) 1, DOI: 10.1038/s41524-023-01115-3.

[88] K. Ryczko, D. A. Strubbe, and I. Tamblyn, *Deep learning and density-functional theory*, Phys. Rev. A **100** (2019) 022512, DOI: 10.1103/PHYSREVA.100.022512.

[89] D. M. Anstine, R. Zubatyuk, and O. Isayev, *AIMNet2: a neural network potential to meet your neutral, charged, organic, and elemental-organic needs*, Chem. Sci. **16** (2025) 10228, DOI: 10.1039/D4SC08572H.

[90] B. M. Wood et al., *UMA: A Family of Universal Models for Atoms*, ChemRxiv (2025), DOI: 10.48550/arXiv.2506.23971.

[91] G. Wang, C. Wang, X. Zhang, Z. Li, J. Zhou, and Z. Sun, *Machine learning interatomic potential: Bridge the gap between small-scale models and realistic device-scale simulations*, iScience **27** (2024) 109673, DOI: 10.1016/J.ISCI.2024.109673.

[92] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, *Physics-informed machine learning*, Nat. Rev. Phys. **3** (2021) 422, DOI: 10.1038/s42254-021-00314-5.

[93] D. Montes de Oca Zapiain, M. A. Wood, N. Lubbers, C. Z. Pereyra, A. P. Thompson, and D. Perez, *Training data selection for accuracy and transferability of interatomic potentials*, Npj Comput. Mater. **8** (2022) 1, DOI: 10.1038/s41524-022-00872-x.

[94] J. T. Margraf, *Science-Driven Atomistic Machine Learning*, Angew. Chem. Int. Ed. **62** (2023) e202219170, DOI: 10.1002/anie.202219170.

[95] M. Kulichenko et al., *Data Generation for Machine Learning Interatomic Potentials and Beyond*, Chem. Rev. **124** (2024) 13681, DOI: 10.1021/ACS.CHEMREV.4C00572.

[96] P. Atkins, J. de Paula, and J. Keeler, *Physikalische Chemie*, 5th ed., Weinheim: Wiley-VCH, 2013, ISBN: 978-3-527-33247-2.

[97] R. E. Skyner, J. L. McDonagh, C. R. Groom, T. Van Mourik, and J. B. Mitchell, *A review of methods for the calculation of solution free energies and the modelling of systems in solution*, Phys. Chem. Chem. Phys. **17** (2015) 6174, DOI: 10.1039/C5CP00288E.

[98] R. Sure and S. Grimme, *Comprehensive Benchmark of Association (Free) Energies of Realistic Host-Guest Complexes*, J. Chem. Theory Comput. **11** (2015) 3785, DOI: 10.1021/acs.jctc.5b00296.

[99] F. Jensen, *Introduction to Computational Chemistry*, 2nd ed., West Sussex: John Wiley & Sons, Ltd, 2007, ISBN: 978-0-470-01186-7.

[100] J. Reinhold, *Quantentheorie der Moleküle*, 5th ed., Wiesbaden: Springer Fachmedien Wiesbaden, 2015, ISBN: 978-3-658-09409-6, DOI: 10.1007/978-3-658-09410-2.

[101] M. Born and R. Oppenheimer, *Zur Quantentheorie der Molekeln*, Ann Phys. **389** (1927) 457, DOI: 10.1002/ANDP.19273892002.

[102] O. Vahtras, J. Almlöf, and M. W. Feyereisen, *Integral approximations for LCAO-SCF calculations*, Chem. Phys. Lett. **213** (1993) 514, DOI: 10.1016/0009-2614(93)89151-7.

[103] F. Neese, F. Wennmohs, A. Hansen, and U. Becker, *Efficient, approximate and parallel Hartree–Fock and hybrid DFT calculations. A 'chain-of-spheres' algorithm for the Hartree–Fock exchange*, Chem. Phys. **356** (2009) 98, DOI: 10.1016/J.CHEMPHYS.2008.10.036.

[104] T. H. Dunning, *Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen*, J. Chem. Phys. **90** (1989) 1007, DOI: 10.1063/1.456153.

[105] F. Weigend and R. Ahlrichs, *Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy*, Phys. Chem. Chem. Phys. **7** (2005) 3297, DOI: 10.1039/b508541a.

[106] D. Rappoport and F. Furche, *Property-optimized Gaussian basis sets for molecular response calculations*, J. Chem. Phys. **133** (2010) 134105, DOI: 10.1063/1.3484283.

[107] F. Jensen, *The magnitude of intramolecular basis set superposition error*, Chem. Phys. Lett. **261** (1996) 633, DOI: 10.1016/0009-2614(96)01033-0.

[108] M. Gutowski, J. H. Van Lenthe, J. Verbeek, F. B. Van Duijneveldt, and G. Chałasinski, *The basis set superposition error in correlated electronic structure calculations*, Chem. Phys. Lett. **124** (1986) 370, DOI: 10.1016/0009-2614(86)85036-9.

[109] S. F. Boys and F. Bernardi, *The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors*, Mol. Phys. **19** (1970) 553, DOI: 10.1080/00268977000101561.

[110] H. Kruse and S. Grimme, *A geometrical correction for the inter- and intra-molecular basis set superposition error in Hartree-Fock and density functional theory calculations for large systems*, J. Chem. Phys. **136** (2012) 154101, DOI: 10.1063/1.3700154.

[111] R. J. Bartlett and I. Shavitt, *Determination of the size-consistency error in the single and double excitation configuration interaction model*, Int. J. Quantum Chem. **12** (1977) 165, DOI: 10.1002/QUA.560120821.

[112] K. Raghavachari, G. W. Trucks, J. A. Pople, and M. Head-Gordon, *A fifth-order perturbation comparison of electron correlation theories*, Chem. Phys. Lett. **157** (1989) 479, DOI: 10.1016/S0009-2614(89)87395-6.

[113] R. J. Bartlett and M. Musiał, *Coupled-cluster theory in quantum chemistry*, Rev. Mod. Phys. **79** (2007) 291, DOI: 10.1103/REVMODPHYS.79.291.

[114] R. Ahlrichs et al., *PNO–CI (pair natural orbital configuration interaction) and CEPA–PNO (coupled electron pair approximation with pair natural orbitals) calculations of molecular systems. I. Outline of the method for closed-shell states*, J. Chem. Phys. **62** (1975) 1225, DOI: 10.1063/1.430637.

[115] W. Kutzelnigg, *$r_1$2-Dependent terms in the wave function as closed sums of partial wave amplitudes for large l*, Theor. Chim. Acta **68** (1985) 445, DOI: 10.1007/BF00527669.

[116] C. Møller and M. S. Plesset, *Note on an Approximation Treatment for Many-Electron Systems*, Phys. Rev. **46** (1934) 618, DOI: 10.1103/PhysRev.46.618.

[117] P. Hohenberg and W. Kohn, *Inhomogeneous electron gas*, Phys. Rev. **136** (1964) B864, DOI: 10.1103/PhysRev.136.B864.

[118] N. H. March, *Theory of the Inhomogeneous Electron Gas*, Boston, MA: Springer US, 1983 1, ISBN: 978-1-4899-0415-7, DOI: 10.1007/978-1-4899-0415-7.

[119] J. P. Perdew and K. Schmidt, *Jacob's ladder of density functional approximations for the exchange-correlation energy*, AIP Conf. Proc. **577** (2001) 1, DOI: 10.1063/1.1390175.

[120] S. Kristyán and P. Pulay, *Can (semi)local density functional theory account for the London dispersion forces?*, Chem. Phys. Lett. **229** (1994) 175, DOI: 10.1016/0009-2614(94)01027-7.

[121] J. M. Pérez-Jordá and A. D. Becke, *A density-functional study of van der Waals forces: rare gas diatomics*, Chem. Phys. Lett. **233** (1995) 134, DOI: 10.1016/0009-2614(94)01402-H.

[122] O. A. Vydrov and T. Van Voorhis, *Nonlocal van der Waals density functional: The simpler the better*, J. Chem. Phys. **133** (2010) 244103, DOI: 10.1063/1.3521275.

[123] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, *A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu*, J. Chem. Phys. **132** (2010) 154104, DOI: 10.1063/1.3382344.

[124] S. Grimme, S. Ehrlich, and L. Goerigk, *Effect of the damping function in dispersion corrected density functional theory*, J. Comput. Chem. **32** (2011) 1456, DOI: 10.1002/jcc.21759.

[125] E. Caldeweyher et al., *A generally applicable atomic-charge dependent London dispersion correction*, J. Chem. Phys. **150** (2019) 154122, DOI: 10.1063/1.5090222.

[126] S. A. Ghasemi, A. Hofstetter, S. Saha, and S. Goedecker, *Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network*, Phys. Rev. B Condens. Matter **92** (2015) 045131, DOI: 10.1103/PHYSREVB.92.045131.

[127] W. J. Mortier, S. K. Ghosh, and S. Shankar, *Electronegativity Equalization Method for the Calculation of Atomic Charges in Molecules*, J. Am. Chem. Soc. **108** (1986) 4315, DOI: 10.1021/JA00275A013.

[128] R. Sure and S. Grimme, *Corrected small basis set Hartree-Fock method for large systems*, J. Comput. Chem. **34** (2013) 1672, DOI: 10.1002/JCC.23317.

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

[130] M. Müller, A. Hansen, and S. Grimme, *ωB97X-3c: A composite range-separated hybrid DFT method with a molecule-optimized polarized valence double-ζ basis set*, J. Chem. Phys. **158** (2023) 14103, DOI: 10.1063/5.0133026.

[131] J. A. Pople, *Electron interaction in unsaturated hydrocarbons*, Trans. Faraday Soc. **49** (1953) 1375, DOI: 10.1039/TF9534901375.

[132] R. Pariser and R. G. Parr, *A Semi-Empirical Theory of the Electronic Spectra and Electronic Structure of Complex Unsaturated Molecules. I.*, J. Chem. Phys. **21** (1953) 466, DOI: 10.1063/1.1698929.

[133] J. J. P. Stewart, *Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements*, J. Mol. Model. **13** (2007) 1173, DOI: 10.1007/S00894-007-0233-4.

[134] J. J. Stewart, *Optimization of parameters for semiempirical methods VI: More modifications to the NDDO approximations and re-optimization of parameters*, J. Mol. Model. **19** (2013) 1, DOI: 10.1007/S00894-012-1667-X.

[135] M. Gaus, Q. Cui, and M. Elstner, *DFTB3: Extension of the self-consistent-charge density-functional tight-binding method (SCC-DFTB)*, J. Chem. Theory Comput. **7** (2011) 931, DOI: 10.1021/CT100684S.

[136] S. Riniker, *Fixed-Charge Atomistic Force Fields for Molecular Dynamics Simulations in the Condensed Phase: An Overview*, J. Chem. Inf. Model. **58** (2018) 565, DOI: 10.1021/ACS.JCIM.8B00042.

[137] A. K. Rappé, C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff, *UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations*, J. Am. Chem. Soc. **114** (1992) 10024, DOI: 10.1021/JA00051A040.

[138] S. Grimme, *Supramolecular Binding Thermodynamics by Dispersion-Corrected Density Functional Theory*, Chem. Eur. J. **18** (2012) 9955, DOI: 10.1002/chem.201200497.

[139] J. M. Herbert, *Dielectric continuum methods for quantum chemistry*, Wiley Interdiscip. Rev. Comput. Mol. Sci. **11** (2021) e1519, DOI: 10.1002/WCMS.1519.

[140] D. Sitkoff, K. A. Sharp, and B. Honig, *Accurate calculation of hydration free energies using macroscopic solvent models*, J. Phys. Chem. **98** (1994) 1978, DOI: 10.1021/J100058A043.

[141] B. Lee and F. M. Richards, *The interpretation of protein structures: Estimation of static accessibility*, J. Mol. Biol. **55** (1971) 379, DOI: 10.1016/0022-2836(71)90324-X.

[142] R. C. Harris and B. M. Pettitt, *Examining the Assumptions Underlying Continuum-Solvent Models*, J. Chem. Theory Comput. **11** (2015) 4593, DOI: 10.1021/acs.jctc.5b00684.

[143] A. Klamt, *Conductor-like screening model for real solvents: A new approach to the quantitative calculation of solvation phenomena*, J. Phys. Chem. **99** (1995) 2224, DOI: 10.1021/j100007a062.

[144] F. Eckert and A. Klamt, *Fast Solvent Screening via Quantum Chemistry: COSMO-RS Approach*, AICHE J. **48** (2002) 369, DOI: 10.1002/aic.690480220.

[145] A. V. Marenich, C. J. Cramer, and D. G. Truhlar, *Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions*, J. Phys. Chem. B **113** (2009) 6378, DOI: 10.1021/jp810292n.

[146] J. G. Kirkwood, *Theory of solutions of molecules containing widely separated charges with special application to zwitterions*, J. Chem. Phys. **2** (1934) 351, DOI: 10.1063/1.1749489.

[147] W. Clark Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, *Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics*, J. Am. Chem. Soc. **112** (1990) 6127, DOI: 10.1021/ja00172a038.

[148] M. Born, *Volumen und Hydratationswärme der Ionen*, Z. Phys. **1** (1920) 45, DOI: 10.1007/BF01881023.

[149] A. V. Onufriev and D. A. Case, *Generalized Born Implicit Solvent Models for Biomolecules*, Annu. Rev. Biophys. **48** (2019) 275, DOI: 10.1146/annurev-biophys-052118-115325.

[150] G. Sigalov, P. Scheffel, and A. Onufriev, *Incorporating variable dielectric environments into the generalized Born model*, J. Chem. Phys. **122** (2005) 94511, DOI: 10.1063/1.1857811.

[151] G. Sigalov, A. Fenley, and A. Onufriev, *Analytical electrostatics for biomolecules: Beyond the generalized Born approximation*, J. Chem. Phys. **124** (2006) 124902, DOI: 10.1063/1.2177251.

[152] J. Tomasi and M. Persico, *Molecular Interactions in Solution: An Overview of Methods Based on Continuous Distributions of the Solvent*, Chem. Rev. **94** (1994) 2027, DOI: 10.1021/cr00031a013.

[153] J. Tomasi, B. Mennucci, and R. Cammi, *Quantum mechanical continuum solvation models*, Chem. Rev. **105** (2005) 2999, DOI: 10.1021/cr9904009.

[154] V. Barone and M. Cossi, *Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model*, J. Phys. Chem. A **102** (1998) 1995, DOI: 10.1021/jp9716997.

[155] A. Klamt and G. Schüürmann, *COSMO: A new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient*, J. Chem. Soc. Perkin Trans. 2 (1993) 799, DOI: 10.1039/P29930000799.

[156] A. Klamt, V. Jonas, T. Bürger, and J. C. Lohrenz, *Refinement and parametrization of COSMO-RS*, J. Phys. Chem. A **102** (1998) 5074, DOI: 10.1021/jp980017s.

[157] R. Anandakrishnan, A. Drozdetski, R. C. Walker, and A. V. Onufriev, *Speed of Conformational Change: Comparing Explicit and Implicit Solvent Molecular Dynamics Simulations*, Biophys. J. **108** (2015) 1153, DOI: 10.1016/J.BPJ.2014.12.047.

[158] R. W. Zwanzig, *High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases*, J. Chem. Phys. **22** (1954) 1420, DOI: 10.1063/1.1740409.

[159] D. Shivakumar, J. Williams, Y. Wu, W. Damm, J. Shelley, and W. Sherman, *Prediction of absolute solvation free energies using molecular dynamics free energy perturbation and the opls force field*, J. Chem. Theory Comput. **6** (2010) 1509, DOI: 10.1021/ct900587b.

[160] D. Wu and D. A. Kofke, *Phase-space overlap measures. II. Design and implementation of staging methods for free-energy calculations*, J. Chem. Phys. **123** (2005) 84109, DOI: 10.1063/1.2011391.

[161] J. G. Kirkwood, *Statistical mechanics of fluid mixtures*, J. Chem. Phys. **3** (1935) 300, DOI: 10.1063/1.1749657.

[162] M. R. Shirts and V. S. Pande, *Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration*, J. Chem. Phys. **122** (2005) 144107, DOI: 10.1063/1.1873592.

[163] C. H. Bennett, *Efficient estimation of free energy differences from Monte Carlo data*, J. Comput. Phys. **22** (1976) 245, DOI: 10.1016/0021-9991(76)90078-4.

[164] D. Suárez and N. Díaz, *Direct methods for computing single-molecule entropies from molecular simulations*, Wiley Interdiscip. Rev. Comput. Mol. Sci. **5** (2015) 1, DOI: 10.1002/WCMS.1195.

[165] J. W. Gibbs, *Elementary principles in statistical mechanics: Developed with especial reference to the rational foundation of thermodynamics*, Cambridge: Cambridge University Press, 2010 1, ISBN: 9780511686948, DOI: 10.1017/CBO9780511686948.

[166] C. E. Shannon, *A Mathematical Theory of Communication*, Bell Labs Tech. J. **27** (1948) 379, DOI: 10.1002/J.1538-7305.1948.TB01338.X.

[167] H. B. Schlegel, *Geometry optimization*, Wiley Interdiscip. Rev. Comput. Mol. Sci. **1** (2011) 790, DOI: 10.1002/WCMS.34.

[168] R. Fletcher, *Practical Methods of Optimization*, John Wiley & Sons, Ltd, 2000, ISBN: 9781118723203, DOI: 10.1002/9781118723203.

[169] D. C. Liu and J. Nocedal, *On the limited memory BFGS method for large scale optimization*, Math. Program. **45** (1989) 503, DOI: 10.1007/BF01589116.

[170] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Society for Industrial and Applied Mathematics, 1996, DOI: 10.1137/1.9781611971200.

[171] W. Huang and B. Leimkuhler, *The Adaptive Verlet Method*, SIAM J. Sci. Comput. **18** (2006) 239, DOI: 10.1137/S1064827595284658.

[172] A. Quarteroni and A. Valli, *Domain Decomposition for Partial Differential Equations*, Oxford University Press, 1999, ISBN: 9780198501787.

[173] J. P. Ryckaert, G. Ciccotti, and H. J. Berendsen, *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes*, J. Comput. Phys. **23** (1977) 327, DOI: 10.1016/0021-9991(77)90098-5.

[174] H. J. Berendsen, J. P. Postma, W. F. Van Gunsteren, A. Dinola, and J. R. Haak, *Molecular dynamics with coupling to an external bath*, J. Chem. Phys. **81** (1984) 3684, DOI: 10.1063/1.448118.

[175] W. G. Hoover, *Canonical dynamics: Equilibrium phase-space distributions*, Phys. Rev. A **31** (1985) 1695, DOI: 10.1103/PhysRevA.31.1695.

[176] S. Nosé, *A unified formulation of the constant temperature molecular dynamics methods*, J. Chem. Phys. **81** (1984) 511, DOI: 10.1063/1.447334.

[177] S. Grimme, *Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations*, J. Chem. Theory Comput. **15** (2019) 2847, DOI: 10.1021/acs.jctc.9b00143.

[178] S. Riniker and G. A. Landrum, *Better Informed Distance Geometry: Using What We Know to Improve Conformation Generation*, J. Chem. Inf. Model. **55** (2015) 2562, DOI: 10.1021/ACS.JCIM.5B00654.

[179] P. Pracht et al., *CREST–A program for the exploration of low-energy molecular chemical space*, J. Chem. Phys. **160** (2024) 114110, DOI: 10.1063/5.0197592.

[180] *Semiempirical Extended Tight-Binding Program Package*, **2022**, https://github.com/grimme-lab/xtb.

[181] S. Ehlert, M. Stahn, S. Spicher, and S. Grimme, *Robust and efficient implicit solvation model for fast semiempirical methods*, J. Chem. Theory Comput. **17** (2021) 4250, DOI: 10.1021/ACS.JCTC.1C00471.

[182] Z. Shpilt et al., *Homoleptic Ti[ONO]$_2$ type complexes of amino-acid-tethered phenolato Schiff-base ligands: Synthesis, characterization, time-resolved fluorescence spectroscopy, and cytotoxicity against ovarian and colon cancer cells*, Appl. Organomet. Chem. **34** (2020) e5309, DOI: 10.1002/aoc.5309.

[183] P. Liebing, E. Pietrasiak, E. Otth, J. Kalim, D. Bornemann, and A. Togni, *Supramolecular Aggregation of Perfluoroorganyl Iodane Reagents in the Solid State and in Solution*, European J. Org. Chem. **2018** (2018) 3771, DOI: 10.1002/ejoc.201800358.

[184] R. A. Howie, E. R. Tiekink, J. L. Wardell, and S. M. Wardell, *Complementary supramolecular aggregation via O-H· · ·O Hydrogen-bonding and Hg· · ·S Interactions in Bis[N,N'-di(2-hydroxyethyl)-dithiocarbamato-S,S']mercury(II): Hg[$S_2CN(CH_2CH_2OH)_2$]$_2$*, J. Chem. Crystallogr. **39** (2009) 293, DOI: 10.1007/s10870-008-9473-0.

[185] G. Soavi et al., *Encapsulation of Trimethine Cyanine in Cucurbit[8]uril: Solution versus Solid-State Inclusion Behavior*, Chem. Eur. J. **28** (2022) e202200185, DOI: 10.1002/chem.202200185.

[186] S. Furukawa et al., *Rhodium-Organic Cuboctahedra as Porous Solids with Strong Binding Sites*, Inorg. Chem. **55** (2016) 10843, DOI: 10.1021/acs.inorgchem.6b02091.

[187] D. Fujita, Y. Ueda, S. Sato, N. Mizuno, T. Kumasaka, and M. Fujita, *Self-assembly of tetravalent Goldberg polyhedra from 144 small components*, Nature **540** (2016) 563, DOI: 10.1038/nature20771.

[188] J. Gu et al., *pH-triggered reversible "stealth" polycationic micelles*, Biomacromolecules **9** (2008) 255, DOI: 10.1021/bm701084w.

[189] J. B. Sperry et al., *Kiloscale buchwald-hartwig amination: Optimized coupling of base-sensitive 6-bromoisoquinoline-1-carbonitrile with (s)-3-amino-2-methylpropan-1-ol*, Org. Process Res. Dev. **18** (2014) 1752, DOI: 10.1021/op5002319.

[190] R. Y. Rohling, E. J. M. Hensen, and E. A. Pidko, *Multi-site Cooperativity in Alkali-Metal-Exchanged Faujasites for the Production of Biomass-Derived Aromatics*, ChemPhysChem **19** (2018) 446, DOI: 10.1002/cphc.201701058.

[191] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, *Comparison of simple potential functions for simulating liquid water*, J. Chem. Phys. **79** (1983) 926, DOI: 10.1063/1.445869.

[192] A. Klamt, *The COSMO and COSMO-RS solvation models*, Wiley Interdiscip. Rev. Comput. Mol. Sci. **8** (2018) e1338, DOI: 10.1002/wcms.1338.

[193] M. Steiner, T. Holzknecht, M. Schauperl, and M. Podewitz, *Quantum Chemical Microsolvation by Automated Water Placement*, Molecules **26** (2021) 1793, DOI: 10.3390/molecules26061793.

[194] P. S. Patel et al., *Bacillaene, a novel inhibitor of procaryotic protein synthesis produced by Bacillus subtilis: production, taxonomy, isolation, physico-chemical characterization and biological activity*, J. Antibiot. **48** (1995) 997, DOI: 10.7164/antibiotics.48.997.

[195] Q. Ma, M. Schwilk, C. Köppl, and H. J. Werner, *Scalable Electron Correlation Methods. 4. Parallel Explicitly Correlated Local Coupled Cluster with Pair Natural Orbitals (PNO-LCCSD-F12)*, J. Chem. Theory Comput. **13** (2017) 4871, DOI: 10.1021/ACS.JCTC.7B00799.

[196] Q. Ma and H. J. Werner, *Explicitly correlated local coupled-cluster methods using pair natural orbitals*, Wiley Interdiscip. Rev. Comput. Mol. Sci. **8** (2018) e1371, DOI: 10.1002/WCMS.1371.

[197] Q. Ma and H. J. Werner, *Scalable Electron Correlation Methods. 8. Explicitly Correlated Open-Shell Coupled-Cluster with Pair Natural Orbitals PNO-RCCSD(T)-F12 and PNO-UCCSD(T)-F12*, J. Chem. Theory Comput. **17** (2021) 902, DOI: 10.1021/ACS.JCTC.0C01129.

[198] H.-J. Werner and A. Hansen, *Accurate Calculation of Isomerization and Conformational Energies of Larger Molecules Using Explicitly Correlated Local Coupled Cluster Methods in Molpro and ORCA*, J. Chem. Theory Comput. **19** (2023) 7007, DOI: 10.1021/acs.jctc.3c00270.

[199] G. Santra, N. Sylvetsky, and J. M. Martin, *Minimally Empirical Double-Hybrid Functionals Trained against the GMTKN55 Database: RevDSD-PBEP86-D4, revDOD-PBE-D4, and DOD-SCAN-D4*, J. Phys. Chem. A (2019), DOI: 10.1021/ACS.JPCA.9B03157.

[200] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch, *Ab Initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields*, J. Phys. Chem. **98** (1994) 11623, DOI: 10.1021/J100096A001.

[201] A. D. Becke, *Density-functional thermochemistry. III. The role of exact exchange*, J. Chem. Phys. **98** (1993) 5648, DOI: 10.1063/1.464913.

[202] Y. Zhao and D. G. Truhlar, *Design of density functionals that are broadly accurate for thermochemistry, thermochemical kinetics, and nonbonded interactions*, J. Phys. Chem. A **109** (2005) 5656, DOI: 10.1021/jp050536c.

[203] C. Adamo and V. Barone, *Toward reliable density functional methods without adjustable parameters: The PBE0 model*, J. Chem. Phys. **110** (1999) 6158, DOI: 10.1063/1.478522.

[204] Y. Zhao and D. G. Truhlar, *The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: Two new functionals and systematic testing of four M06-class functionals and 12 other function*, Theor. Chem. Acc. **120** (2008) 215, DOI: 10.1007/S00214-007-0310-X.

[205] A. Najibi and L. Goerigk, *The Nonlocal Kernel in van der Waals Density Functionals as an Additive Correction: An Extensive Analysis with Special Emphasis on the B97M-V and ωb97M-V Approaches*, J. Chem. Theory Comput. **14** (2018) 5725, DOI: 10.1021/ACS.JCTC.8B00842.

[206] N. Mardirossian and M. Head-Gordon, *Mapping the genome of meta-generalized gradient approximation density functionals: The search for B97M-V*, J. Chem. Phys. **142** (2015) 74111, DOI: 10.1063/1.4907719.

[207] A. Selloni, P. Carnevali, E. Tosatti, C. D. Chen, M. M. Mohan, and A. Griffin, *Erratum: Density-functional approximation for the correlation energy of the inhomogeneous electron gas*, Phys. Rev. B **34** (1986) 7406, DOI: 10.1103/PhysRevB.34.7406.

[208] J. P. Perdew, *Density-functional approximation for the correlation energy of the inhomogeneous electron gas*, Phys. Rev. B **33** (1986) 8822, DOI: 10.1103/PhysRevB.33.8822.

[209] A. D. Becke, *Density-functional exchange-energy approximation with correct asymptotic behavior*, Phys. Rev. A **38** (1988) 3098, DOI: 10.1103/PhysRevA.38.3098.

[210] S. Grimme, J. G. Brandenburg, C. Bannwarth, and A. Hansen, *Consistent structures and interactions by density functional theory with small atomic orbital basis sets*, J. Chem. Phys. **143** (2015) 54107, DOI: 10.1063/1.4927476.

[211] T. A. Halgren, *Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94*, J. Comput. Chem. **17** (1996) 490, DOI: 10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P.

[212] T. A. Halgren, *Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions*, J. Comput. Chem. **17** (1996) 520, DOI: 10.1002/(SICI)1096-987X(199604)17:5/6<520::AID-JCC2>3.0.CO;2-W.

[213] T. A. Halgren, *Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94*, J. Comput. Chem. **17** (1996) 553, DOI: 10.1002/(SICI)1096-987X(199604)17:5/6<553::AID-JCC3>3.0.CO;2-T.

[214] T. A. Halgren and R. B. Nachbar, *Merck Molecular Force Field. IV. Conformational Energies and Geometries for MMFF94*, J. Comput. Chem. **17** (1996) 587, DOI: 10.1002/(SICI)1096-987X(199604)17:5/6<587::AID-JCC4>3.0.CO;2-Q.

[215] T. A. Halgren, *Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules*, J. Comput. Chem. **17** (1996) 616, DOI: 10.1002/(SICI)1096-987X(199604)17:5/6<616::AID-JCC5>3.0.CO;2-X.

[216] H. Lin and D. G. Truhlar, *QM/MM: What have we learned, where are we, and where do we go from here?*, Theor. Chem. Acc. **117** (2007) 185, DOI: 10.1007/S00214-006-0143-Z.

[217] K. Y. Dorogov, A. V. Churakov, L. G. Kuzmina, J. A. Howard, and G. I. Nikonov, *Direct Hydride/Silyl Exchange — Synthesis and X-ray Study of the Bis(silyl) Complex [$Cp_2NbH(SiCl_3)_2$]*, Eur. J. Inorg. Chem. **2004** (2004) 771, DOI: 10.1002/EJIC.200300134.

[218] J. H. Cavka et al., *A new zirconium inorganic building brick forming metal organic frameworks with exceptional stability*, J. Am. Chem. Soc. **130** (2008) 13850, DOI: 10.1021/JA8057953.

[219] Z. Zhang, J. Chen, Z. Bao, G. Chang, H. Xing, and Q. Ren, *Insight into the catalytic properties and applications of metal–organic frameworks in the cyanosilylation of aldehydes*, RSC Adv. **5** (2015) 79355, DOI: 10.1039/C5RA13102B.

[220] M. A. Majewski et al., *Covalent Template-Directed Synthesis of a Spoked 18-Porphyrin Nanoring*, Angew. Chem. Int. Ed. **62** (2023) e202302114, DOI: 10.1002/ANIE.202302114.

[221] T. Kalvoda, M. Culka, L. Rulíšek, and E. Andris, *Exhaustive Mapping of the Conformational Space of Natural Dipeptides by the DFT-D3//COSMO-RS Method*, J. Phys. Chem. B **126** (2022) 5949, DOI: 10.1021/acs.jpcb.2c02861.

[222] J. W. Furness, A. D. Kaplan, J. Ning, J. P. Perdew, and J. Sun, *Accurate and Numerically Efficient r^2SCAN Meta-Generalized Gradient Approximation*, J. Phys. Chem. Lett. **11** (2020) 8208, DOI: 10.1021/ACS.JPCLETT.0C02405.

[223] Y. Zhao and D. G. Truhlar, *A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions*, J. Chem. Phys. **125** (2006), DOI: 10.1063/1.2370993.

[224] P. Eastman et al., *SPICE, A Dataset of Drug-like Molecules and Peptides for Training Machine Learning Potentials*, Sci. Data **10** (2023) 1, DOI: 10.1038/s41597-022-01882-6.

[225] N. Godbout, D. R. Salahub, J. Andzelm, and E. Wimmer, *Optimization of Gaussian-type basis sets for local spin density functional calculations. Part I. Boron through neon, optimization technique and validation*, Can. J. Chem. **70** (1992) 560, DOI: 10.1139/v92-079.

[226] J. Hostaš and J. Řezáč, *Accurate DFT-D3 Calculations in a Small Basis Set*, J. Chem. Theory Comput. **13** (2017) 3575, DOI: 10.1021/acs.jctc.7b00365.

[227] D. M. Anstine, R. Zubatyuk, and O. Isayev, *AIMNet2: A Neural Network Potential to Meet your Neutral, Charged, Organic, and Elemental-Organic Needs*, ChemRxiv (2023), DOI: 10.26434/CHEMRXIV-2023-296CH.

[228] S. Y. Sheu, D. Y. Yang, H. L. Selzle, and E. W. Schlag, *Energetics of hydrogen bonds in peptides*, Proc. Natl. Acad. Sci. U.S.A. **100** (2003) 12683, DOI: 10.1073/PNAS.2133366100.

[229] C. Aguiar, N. Dattani, and I. Camps, *Möbius carbon nanobelts interacting with heavy metal nanoclusters*, J. Mol. Model. **29** (2023) 1, DOI: 10.1007/S00894-023-05669-3.

[230] L. Chen et al., *MORE-Q, a dataset for molecular olfactorial receptor engineering by quantum mechanics*, Sci. Data **12** (2025) 1, DOI: 10.1038/s41597-025-04616-6.

[231] P. Skurski and J. Brzeski, *Carbonless DNA*, Phys. Chem. Chem. Phys. **27** (2025) 2343, DOI: 10.1039/D4CP04410J.

[232] M. Tarek Ibrahim, E. Wait, and P. Ren, *Quantum Mechanics Characterization of Non-Covalent Interaction in Nucleotide Fragments*, Molecules **29** (2024) 3258, DOI: 10.3390/molecules29143258.

[233] M. Díaz-Abellás, I. Neira, A. Blanco-Gómez, C. Peinador, and M. D. García, *Synergy-Promoted Specific Alkyltriphenylphosphonium Binding to CB[8]*, J. Org. Chem. Res. **90** (2025) 4149, DOI: 10.1021/ACS.JOC.4C02546.

[234] J. T. Kohn, S. Grimme, and A. Hansen, *A semi-automated quantum-mechanical workflow for the generation of molecular monolayers and aggregates*, J. Chem. Phys. **161** (2024) 124707, DOI: 10.1063/5.0230341.

[235] A. R. Puente and P. L. Polavarapu, *Influence of microsolvation on vibrational circular dichroism spectra in dimethyl sulfoxide solvent: A Bottom-Up approach using Quantum cluster growth*, Spectrochim. Acta – A: Mol. Biomol. Spectrosc. **303** (2023) 123231, DOI: 10.1016/J.SAA.2023.123231.

[236] A. S. Perera, C. D. Carlson, J. Cheramy, and Y. Xu, *Infrared and vibrational circular dichroism spectra of methyl β-D-glucopyranose in water: The application of the quantum cluster growth and clusters-in-a-liquid solvation models*, Chirality **35** (2023) 718, DOI: 10.1002/CHIR.23576.

[237] S. A. Katsyuba and T. I. Burganov, *Computational analysis of the vibrational spectra and structure of aqueous cytosine*, Phys. Chem. Chem. Phys. **25** (2023) 24121, DOI: 10.1039/D3CP03059H.

[238] S. Pantaleone, C. I. Gho, R. Ferrero, V. Brunella, and M. Corno, *Exploration of the Conformational Scenario for α-, β-, and γ-Cyclodextrins in Dry and Wet Conditions, from Monomers to Crystal Structures: A Quantum-Mechanical Study*, Int. J. Mol. Sci. **24** (2023) 16826, DOI: 10.3390/IJMS242316826.

[239] L. Jiang and K. Zheng, *Electronic structures of zwitterionic and protonated forms of glycine betaine in water: Insights into solvent effects from ab initio simulations*, J. Mol. Liq. **369** (2023) 120871, DOI: 10.1016/J.MOLLIQ.2022.120871.

[240] G. A. Lara-Cruz, T. Rose, S. Grimme, and A. Jaramillo-Botero, *Reaction-Free Energies for Complexation of Carbohydrates by Tweezer Diboronic Acids*, J. Phys. Chem. B **128** (2024) 2025, DOI: 10.1021/ACS.JPCB.4C04846.

[241] C. Spino, M. Latil, R. Lessard, Q. Fevre-Renault, and C. Y. Legault, *N-Oxides as Control Element for the Direction of a Sigmatropic Rearrangement: Application as a Switch for Fluorescence*, Chem. Eur. J. **29** (2023) e202301356, DOI: 10.1002/CHEM.202301356.

[242] R. S. Kingsbury et al., *Kinetic barrier networks reveal rate limitations in ion-selective membranes*, Matter **7** (2024) 2161, DOI: 10.1016/j.matt.2024.03.021.

[243] N. Savale et al., *Structural and thermal properties of cellulose regenerated from superbase ionic liquid: effect of green co-solvents*, Cellulose **32** (2025) 2919, DOI: 10.1007/S10570-025-06452-8.

[244] D. A. Jelski and T. F. George, *Clusters: Link between molecules and solids*, J. Chem. Educ. **65** (1988) 879, DOI: 10.1021/ED065P879.

[245] J. Elm et al., *Modeling the formation and growth of atmospheric molecular clusters: A review*, J. Aerosol Sci. **149** (2020) 105621, DOI: 10.1016/J.JAEROSCI.2020.105621.

[246] C. A. Bergström and P. Larsson, *Computational prediction of drug solubility in water-based systems: Qualitative and quantitative approaches used in the current drug discovery and development setting*, International Journal of Pharmaceutics **540** (1-2 2018) 185, DOI: 10.1016/j.ijpharm.2018.01.044.

[247] T. Froitzheim, M. Müller, A. Hansen, and S. Grimme, *g-xTB: A General-Purpose Extended Tight-Binding Electronic Structure Method For the Elements H to Lr (Z=1–103)*, ChemRxiv (2025), DOI: 10.26434/CHEMRXIV-2025-BJXVT.

[248] R. Sure, M. el Mahdali, A. Plajer, and P. Deglmann, *Towards a converged strategy for including microsolvation in reaction mechanism calculations*, J. Comput.-Aided Mol. Des. **35** (2021) 473, DOI: 10.1007/S10822-020-00366-2.

[249] G. Kaur, H. Kumar, and M. Singla, *Diverse applications of ionic liquids: A comprehensive review*, J. Mol. Liq. **351** (2022) 118556, DOI: 10.1016/J.MOLLIQ.2022.118556.

[250] H. B. Casimir and D. Polder, *The Influence of Retardation on the London-van der Waals Forces*, Phys. Rev. **73** (1948) 360, DOI: 10.1103/PhysRev.73.360.

[251] Y. Muto, *Force between nonpolar molecules*, Proc. Phys.-Math. Soc. Jpn. **17** (1943) 629.

[252] B. M. Axilrod and E. Teller, *Interaction of the van der Waals Type Between Three Atoms*, J. Chem. Phys. **11** (1943) 299, DOI: 10.1063/1.1723844.

[253] J. F. Dobson, *Beyond pairwise additivity in London dispersion interactions*, Int. J. Quantum Chem. **114** (2014) 1157, DOI: 10.1002/QUA.24635.

[254] E. Caldeweyher, C. Bannwarth, and S. Grimme, *Extension of the D3 dispersion coefficient model*, J. Chem. Phys. **147** (2017), DOI: 10.1063/1.4993215.

[255] J. Cao and B. J. Berne, *Many-body dispersion forces of polarizable clusters and liquids*, J. Chem. Phys. **97** (1992) 8628, DOI: 10.1063/1.463381.

[256] J. M. Foster and S. F. Boys, *Canonical Configurational Interaction Procedure*, Rev. Mod. Phys. **32** (1960) 300, DOI: 10.1103/RevModPhys.32.300.

[257] V. M. Anisimov, G. Lamoureux, I. V. Vorobyov, N. Huang, B. Roux, and A. D. MacKerell, *Determination of electrostatic parameters for a polarizable force field based on the classical Drude oscillator*, J. Chem. Theory Comput. **1** (2005) 153, DOI: 10.1021/ct049930p.

[258] S. Grimme and C. Bannwarth, *Ultra-fast computation of electronic spectra for large systems by tight-binding based simplified Tamm-Dancoff approximation (sTDA-xTB)*, J. Chem. Phys. **145** (2016) 54103, DOI: 10.1063/1.4959605.

[259] Z. Zhou, X. Yan, T. R. Cook, M. L. Saha, and P. J. Stang, *Engineering Functionalization in a Supramolecular Polymer: Hierarchical Self-Organization of Triply Orthogonal Non-covalent Interactions on a Supramolecular Coordination Complex Platform*, J. Am. Chem. Soc. **138** (2016) 806, DOI: 10.1021/jacs.5b12986.

[260] C. Rest, R. Kandanelli, and G. Fernández, *Strategies to create hierarchical self-assembled structures via cooperative non-covalent interactions*, Chem. Soc. Rev. **44** (2015) 2543, DOI: 10.1039/c4cs00497c.

[261] A. Karshikoff, *Non-Covalent Interactions in Proteins*, 2nd, Singapore: World Scientific, 2021, DOI: 10.1142/12035.

[262] C. P. A. Anconi, *Relative Position and Relative Rotation in Supramolecular Systems through the Analysis of the Principal Axes of Inertia: Ferrocene/Cucurbit[7]uril and Ferrocenyl Azide/beta-Cyclodextrin Case Studies*, ACS Omega **5** (2020) 5013, DOI: 10.1021/acsomega.9b03914.

[263] K. N. Houk, F. Liu, Z. Yang, and J. I. Seeman, *Evolution of the Diels–Alder Reaction Mechanism since the 1930s: Woodward, Houk with Woodward, and the Influence of Computational Chemistry on Understanding Cycloadditions*, Angew. Chem., Int. Ed. **60** (2021) 12660, DOI: 10.1002/anie.202001654.

[264] A. S. Christensen, T. Kubař, Q. Cui, and M. Elstner, *Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications*, Chem. Rev. **116** (2016) 5301, DOI: 10.1021/acs.chemrev.5b00584.

[265] J. A. Harrison, J. D. Schall, S. Maskey, P. T. Mikulski, M. T. Knippenberg, and B. H. Morrow, *Review of force fields and intermolecular potentials used in atomistic computational materials research*, AAppl. Phys. Rev. **5** (2018) 031104, DOI: 10.1063/1.5020808.

[266] M. Bursch, H. Neugebauer, and S. Grimme, *Structure Optimisation of Large Transition-Metal Complexes with Extended Tight-Binding Methods*, Angew. Chem., Int. Ed. **58** (2019) 11078, DOI: 10.1002/anie.201904021.

[267] S. Spicher, M. Bursch, and S. Grimme, *Efficient calculation of small molecule binding in metal-organic frameworks and porous organic cages*, J. Phys. Chem. C **124** (2020) 27529, DOI: 10.1021/acs.jpcc.0c08617.

[268] S. Spicher and S. Grimme, *Efficient Computation of Free Energy Contributions for Association Reactions of Large Molecules*, J. Phys. Chem. Lett. **11** (2020) 6606, DOI: 10.1021/acs.jpclett.0c01930.

[269] J. Černý and P. Hobza, *Non-covalent interactions in biomacromolecules*, Phys. Chem. Chem. Phys. **9** (2007) 5291, DOI: 10.1039/b704781a.

[270] P. Hobza, *Calculations on Noncovalent Interactions and Databases of Benchmark Interaction Energies*, Acc. Chem. Res. **45** (2012) 663, DOI: 10.1021/ar200255p.

[271] J. Contreras-García et al., *NCIPLOT: A program for plotting noncovalent interaction regions*, J. Chem. Theory Comput. **7** (2011) 625, DOI: 10.1021/ct100641a.

[272] Z. Wang et al., *Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: The prediction accuracy of sampling power and scoring power*, Phys. Chem. Chem. Phys. **18** (2016) 12964, DOI: 10.1039/c6cp01555g.

[273] R. Wang, Y. Lu, and S. Wang, *Comparative evaluation of 11 scoring functions for molecular docking*, J. Med. Chem. **46** (2003) 2287, DOI: 10.1021/jm0203783.

[274] E. Yuriev, J. Holien, and P. A. Ramsland, *Improvements, trends, and new ideas in molecular docking: 2012-2013 in review*, J. Mol. Recognit. **28** (2015) 581, DOI: 10.1002/jmr.2471.

[275] R. Chen, L. Li, and Z. Weng, *ZDOCK: An initial-stage protein-docking algorithm*, Proteins Struct. Funct. Genet. **52** (2003) 80, DOI: 10.1002/prot.10389.

[276] L. Li, R. Chen, and Z. Weng, *RDOCK: Refinement of Rigid-body Protein Docking Predictions*, Proteins Struct. Funct. Genet. **53** (2003) 693, DOI: 10.1002/prot.10460.

[277] M. Rarey, B. Kramer, and T. Lengauer, *Multiple automatic base selection: Protein-ligand docking based on incremental construction without manual intervention*, J. Comput. Aided. Mol. Des. **11** (1997) 369, DOI: 10.1023/A:1007913026166.

[278] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe, *A fast flexible docking method using an incremental construction algorithm*, J. Mol. Biol. **261** (1996) 470, DOI: 10.1006/jmbi.1996.0477.

[279] G. M. Morris et al., *Software news and updates AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility*, J. Comput. Chem. **30** (2009) 2785, DOI: 10.1002/jcc.21256.

[280] O. Trott and A. J. Olson, *AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading*, J. Comput. Chem. **31** (2009) 455, DOI: 10.1002/jcc.21334.

[281] M. L. Verdonk, J. C. Cole, M. J. Hartshorn, C. W. Murray, and R. D. Taylor, *Improved protein-ligand docking using GOLD*, Proteins Struct. Funct. Genet. **52** (2003) 609, DOI: 10.1002/prot.10465.

[282] T. A. Halgren et al., *Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening*, J. Med. Chem. **47** (2004) 1750, DOI: 10.1021/jm030644s.

[283] H. Zhao and A. Caflisch, *Discovery of ZAP70 inhibitors by high-throughput docking into a conformation of its kinase domain generated by molecular dynamics*, Bioorganic Med. Chem. Lett. **23** (2013) 5721, DOI: 10.1016/j.bmcl.2013.08.009.

[284] R. Leardi, *Genetic algorithms in chemistry*, J. Chromatogr. A **1158** (2007) 226, DOI: 10.1016/j.chroma.2007.04.025.

[285] *Documentation for xtb and related software*, **2022**, https://xtb-docs.readthedocs.io/.

[286] T. Harimoto and Y. Ishigaki, *Redox-Active Hydrocarbons: Isolation and Structural Determination of Cationic States toward Advanced Response Systems*, Chempluschem (2022) e202200013, DOI: 10.1002/cplu.202200013.

[287] H. Maeda, T. Nishimura, A. Tsujii, K. Takaishi, M. Uchiyama, and A. Muranaka, *Helical π-systems of bidipyrrin-metal complexes*, Chem. Lett. **43** (2014) 1078, DOI: 10.1246/cl.140260.

[288] *Conformer-Rotamer Ensemble Sampling Tool based on the xtb Semiempirical Extended Tight-Binding Program Package*, **2022**, https://github.com/grimme-lab/crest.

[289] S. Spicher, C. Plett, P. Pracht, A. Hansen, and S. Grimme, *Automated Molecular Cluster Growing for Explicit Solvation by Efficient Force Field and Tight Binding Methods*, J. Chem. Theory Comput. **18** (2022) 3174, DOI: 10.1021/acs.jctc.2c00239.

[290] R. W. Hartley, *Barnase and barstar: two small proteins to fold and fit together*, Trends Biochem. Sci. **14** (1989) 450, DOI: 10.1016/0968-0004(89)90104-7.

[291] J. S. Butler, D. M. Mitrea, G. Mitrousis, G. Cingolani, and S. N. Loh, *Structural and Thermodynamic Analysis of a Conformationally Strained Circular Permutant of Barnase*, Biochemistry **48** (2009) 3497, DOI: 10.1021/bi900039e.

[292] U. M. Lindström, *Stereoselective organic reactions in water*, Chem. Rev. **102** (2002) 2751, DOI: 10.1021/CR010122P.

[293] M. Toupin, T. Brousse, and D. Bélanger, *Charge storage mechanism of MnO_2 electrode used in aqueous electrochemical capacitor*, Chem. Mater. **16** (2004) 3184, DOI: 10.1021/cm049649j.

[294] A. Nicholls, K. A. Sharp, and B. Honig, *Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons*, Proteins: Struct. Funct. Genet. **11** (1991) 281, DOI: 10.1002/PROT.340110407.

[295] C. M. Dobson, *Protein folding and misfolding*, Nature **426** (2003) 884, DOI: 10.1038/nature02261.

[296] D. van der Spoel, J. Zhang, and H. Zhang, *Quantitative predictions from molecular simulations using explicit or implicit interactions*, WIREs Comput. Mol. Sci. (2021) e1560, DOI: 10.1002/wcms.1560.

[297] C. Caleman, P. J. van Maaren, M. Hong, J. S. Hub, L. T. Costa, and D. van der Spoel, *Force Field Benchmark of Organic Liquids: Density, Enthalpy of Vaporization, Heat Capacities, Surface Tension, Isothermal Compressibility, Volumetric Expansion Coefficient, and Dielectric Constant*, J. Chem. Theory Comput. **8** (2012) 61, DOI: 10.1021/ct200731v.

[298] J. S. Hub, C. Caleman, and D. van der Spoel, *Organic molecules on the surface of water droplets–an energetic perspective*, Phys. Chem. Chem. Phys. **14** (2012) 9537, DOI: 10.1039/C2CP40483D.

[299] H. Daver, A. G. Algarra, J. Rebek, J. N. Harvey, and F. Himo, *Mixed Explicit–Implicit Solvation Approach for Modeling of Alkane Complexation in Water-Soluble Self-Assembled Capsules*, J. Am. Chem. Soc. **140** (2018) 12527, DOI: 10.1021/jacs.8b06984.

[300] P. E. Smith and B. M. Pettitt, *Modeling solvent in biomolecular systems*, J. Phys. Chem. **98** (1994) 9700, DOI: 10.1021/j100090a002.

[301] R. M. Levy and E. Gallicchio, *Computer simulations with explicit solvent: recent progress in the thermodynamic decomposition of free energies and in modeling electrostatic effects*, Annu. Rev. Phys. Chem. **49** (1998) 531, DOI: 10.1146/annurev.physchem.49.1.531.

[302] J. Zhang, B. Tuguldur, and D. van der Spoel, *Force Field Benchmark of Organic Liquids. 2. Gibbs Energy of Solvation*, J. Chem. Inf. Model **55** (2015) 1192, DOI: 10.1021/acs.jcim.5b00106.

[303] C. J. Cramer and D. G. Truhlar, *Implicit solvation models: equilibria, structure, spectra, and dynamics*, Chem. Rev. **99** (1999) 2161, DOI: 10.1021/cr960149m.

[304] H. M. Senn and W. Thiel, *QM/MM methods for biomolecular systems*, Angew. Chem. Int. Ed. **48** (2009) 1198, DOI: 10.1002/anie.200802019.

[305] G. König, F. C. Pickard, Y. Mei, and B. R. Brooks, *Predicting hydration free energies with a hybrid QM/MM approach: an evaluation of implicit and explicit solvation models in SAMPL4*, J. Comput. Aided Mol. Des. **28** (2014) 245, DOI: 10.1007/s10822-014-9708-4.

[306] M. Schwörer, C. Wichmann, and P. Tavan, *A polarizable QM/MM approach to the molecular dynamics of amide groups solvated in water*, J. Chem. Phys. **144** (2016) 114504, DOI: 10.1063/1.4943972.

[307] L.-P. Wang and T. Van Voorhis, *A Polarizable QM/MM Explicit Solvent Model for Computational Electrochemistry in Water*, J. Chem. Theory Comput. **8** (2012) 610, DOI: 10.1021/ct200340x.

[308] M. Bondanza, M. Nottoli, L. Cupellini, F. Lipparini, and B. Mennucci, *Polarizable embedding QM/MM: the future gold standard for complex (bio)systems?*, Phys. Chem. Chem. Phys. **22** (2020) 14433, DOI: 10.1039/D0CP02119A.

[309] S. Miertuš, E. Scrocco, and J. Tomasi, *Electrostatic interaction of a solute with a continuum. A direct utilizaion of AB initio molecular potentials for the prevision of solvent effects*, Chem. Phys. **55** (1981) 117, DOI: 10.1016/0301-0104(81)85090-2.

[310] R. Cammi and J. Tomasi, *Remarks on the use of the apparent surface charges (ASC) methods in solvation problems: Iterative versus matrix-inversion procedures and the renormalization of the apparent charges*, J. Comput. Chem. **16** (1995) 1449, DOI: 10.1002/jcc.540161202.

[311] M. E. Davis and J. A. McCammon, *Electrostatics in biomolecular structure and dynamics*, Chem. Rev. **90** (1990) 509, DOI: 10.1021/cr00101a005.

[312] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon, *Electrostatics of nanosystems: application to microtubules and the ribosome*, Proc. Natl. Acad. Sci. U.S.A. **98** (2001) 10037, DOI: 10.1073/pnas.181342398.

[313] D. Qiu, P. S. Shenkin, F. P. Hollinger, and W. C. Still, *The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii*, J. Phys. Chem. A **101** (1997) 3005, DOI: 10.1021/jp961992r.

[314] G. D. Hawkins, C. J. Cramer, and D. G. Truhlar, *Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium*, J.Phys. Chem. **100** (1996) 19824, DOI: 10.1021/jp961710n.

[315] M. Schaefer and M. Karplus, *A comprehensive analytical treatment of continuum electrostatics*, J. Phys. Chem. **100** (1996) 1578, DOI: 10.1021/jp9521621.

[316] B. Mennucci, R. Cammi, and J. Tomasi, *Excited states and solvatochromic shifts within a nonequilibrium solvation approach: A new formulation of the integral equation formalism method at the self-consistent field, configuration interaction, and multiconfiguration self-consistent field level*, J. Chem. Phys. **109** (1998) 2798, DOI: 10.1063/1.476878.

[317] A. Klamt, *The COSMO and COSMO-RS solvation models*, Wiley Interdiscip. Rev. Comput. Mol. Sci. **1** (2011) 699, DOI: 10.1002/WCMS.56.

[318] F. Eckert, I. Leito, I. Kaljurand, A. Kütt, A. Klamt, and M. Diedenhofen, *Prediction of acidity in acetonitrile solution with COSMO-RS*, J. Comput. Chem. **30** (2009) 799, DOI: 10.1002/jcc.21103.

[319] R. F. Ribeiro, A. V. Marenich, C. J. Cramer, and D. G. Truhlar, *Prediction of SAMPL2 aqueous solvation free energies and tautomeric ratios using the SM8, SM8AD, and SMD solvation models*, J. Comput. Aided Mol. Des. **24** (2010) 317, DOI: 10.1007/s10822-010-9333-9.

[320] K. I. Assaf et al., *HYDROPHOBE Challenge: A Joint Experimental and Computational Study on the Host–Guest Binding of Hydrocarbons to Cucurbiturils, Allowing Explicit Evaluation of Guest Hydration Free-Energy Contributions*, J. Phys. Chem. B **121** (2017) 11144, DOI: 10.1021/acs.jpcb.7b09175.

[321] N. Fleck, C. Heubach, T. Hett, S. Spicher, S. Grimme, and O. Schiemann, *Ox-SLIM: Synthesis of and Site-Specific Labelling with a Highly Hydrophilic Trityl Spin Label*, Chem. Eur. J. **27** (2021) 5292, DOI: 10.1002/chem.202100013.

[322] D. S. D. Larsson and D. van der Spoel, *Screening for the Location of RNA using the Chloride Ion Distribution in Simulations of Virus Capsids*, J. Chem. Theory Comput. **8** (2012) 2474, DOI: 10.1021/ct3002128.

[323] P. Larsson and E. Lindahl, *A high-performance parallel-generalized born implementation enabled by tabulated interaction rescaling*, J. Chem. Theory Comput. **31** (2010) 2593, DOI: 10.1002/jcc.21552.

[324] H. Zhang, T. Tan, and D. van der Spoel, *Generalized Born and Explicit Solvent Models for Free Energy Calculations in Organic Solvents: Cyclodextrin Dimerization*, J. Chem. Theory Comput. **11** (2015) 5103, DOI: 10.1021/acs.jctc.5b00620.

[325] B. J. Alder and T. E. Wainwright, *Studies in molecular dynamics. I. General method*, J. Chem. Phys. **31** (1959) 459, DOI: 10.1063/1.1730376.

[326] A. Rahman, *Correlations in the motion of atoms in liquid argon*, Phys. Rev. **136** (1964) A405, DOI: 10.1103/PhysRev.136.A405.

[327] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *Equation of state calculations by fast computing machines*, J. Chem. Phys. **21** (1953) 1087, DOI: 10.1063/1.1699114.

[328] N. Metropolis and S. Ulam, *The monte carlo method*, J. Am. Stat. Assoc. **44** (1949) 335, DOI: 10.1080/01621459.1949.10483310.

[329] S. A. Martins, S. F. Sousa, M. J. Ramos, and P. A. Fernandes, *Prediction of solvation free energies with thermodynamic integration using the general amber force field*, J. Chem. Theory Comput. **10** (2014) 3570, DOI: 10.1021/ct500346y.

[330] G. Brancato, N. Rega, and V. Barone, *A hybrid explicit/implicit solvation method for first-principle molecular dynamics simulations*, J. Chem. Phys. **128** (2008) 144501, DOI: 10.1063/1.2897759.

[331] J. Aqvist, *Ion–water interaction potentials derived from free energy perturbation simulations*, J. Phys. Chem. **94** (1990) 8021, DOI: 10.1021/j100384a009.

[332] F. Weinhold, *Quantum cluster equilibrium theory of liquids: General theory and computer implementation*, J. Chem. Phys. **109** (1998) 367, DOI: 10.1063/1.476573.

[333] B. Kirchner et al., *What can clusters tell us about the bulk?: Peacemaker: Extended quantum cluster equilibrium calculations*, Comput. Phys. Commun. **182** (2011) 1428, DOI: 10.1016/j.cpc.2011.03.011.

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

[335] J. Yin et al., *Overview of the SAMPL5 host–guest challenge: Are we doing better?*, J. Comput. Aided Mol. Des. **31** (2017) 1, DOI: 10.1007/s10822-016-9974-4.

[336] G. Kumar, Z.-W. Qu, S. Ghosh, S. Grimme, and I. Chatterjee, *Boron Lewis Acid-Catalyzed Regioselective Hydrothiolation of Conjugated Dienes with Thiols*, ACS Catal. **9** (2019) 11627, DOI: 10.1021/acscatal.9b04647.

[337] J. R. Pliego Jr and J. M. Riveros, *Hybrid discrete-continuum solvation methods*, WIREs Comput. Mol. Sci. **10** (2020) e1440, DOI: 10.1002/wcms.1440.

[338] C. P. Kelly, C. J. Cramer, and D. G. Truhlar, *Aqueous solvation free energies of ions and ion-water clusters based on an accurate value for the absolute aqueous solvation free energy of the proton*, J. Phys. Chem. B **110** (2006) 16066, DOI: 10.1021/jp063552y.

[339] J. Thar, S. Zahn, and B. Kirchner, *When is a molecule properly solvated by a continuum model or in a cluster ansatz? a first-principles simulation of alanine hydration*, J. Phys. Chem. B **112** (2008) 1456, DOI: 10.1021/jp077341k.

[340] G. N. Simm, P. L. Türtscher, and M. Reiher, *Systematic microsolvation approach with a cluster-continuum scheme and conformational sampling*, J. Comput. Chem. **41** (2020) 1144, DOI: 10.1002/jcc.26161.

[341] E. Hruska, A. Gale, X. Huang, and F. Liu, *AutoSolvate: A toolkit for automating quantum chemistry design and discovery of solvated molecules*, J. Chem. Phys. **156** (2022) 124801, DOI: 10.1063/5.0084833.

[342] S. Spicher, D. Abdullin, S. Grimme, and O. Schiemann, *Modeling of spin–spin distance distributions for nitroxide labeled biomacromolecules*, Phys. Chem. Chem. Phys. **22** (2020) 24282, DOI: 10.1039/D0CP04920D.

[343] J. Buša et al., *ARVO: A Fortran package for computing the solvent accessible surface area and the excluded volume of overlapping spheres via analytic equations*, Comput. Phys. Commun. **165** (2005) 59, DOI: 10.1016/j.cpc.2004.08.002.

[344] M. D. Tissandier et al., *The proton's absolute aqueous enthalpy and Gibbs free energy of solvation from cluster-ion solvation data*, J. Phys. Chem. A **102** (1998) 7787, DOI: 10.1021/jp982638r.

[345] P. Kollman, *Free energy calculations: applications to chemical and biochemical phenomena*, Chem. Rev. **93** (1993) 2395, DOI: 10.1021/j100384a009.

[346] L. Tomaník, E. Muchová, and P. Slavíček, *Solvation energies of ions with ensemble cluster-continuum approach*, Phys. Chem. Chem. Phys. **22** (2020) 22357, DOI: 10.1039/D0CP02768E.

[347] P. Pracht and S. Grimme, *Calculation of absolute molecular entropies and heat capacities made simple*, Chem. Sci. **12** (2021) 6551, DOI: 10.1039/D1SC00621E.

[348] *Conformer-Rotamer Ensemble Sampling Tool based on the xtb Semiempirical Extended Tight-Binding Program Package crest*, **2021**, https://github.com/grimme-lab/crest.

[349] *Semiempirical Extended Tight-Binding Program Package xtb*, Version 6.4.0., **2020**, https://github.com/grimme-lab/xtb.

[350] A. Schäfer, H. Horn, and R. Ahlrichs, *Fully optimized contracted Gaussian basis sets for atoms Li to Kr*, J. Chem. Phys. **97** (1992) 2571, DOI: 10.1063/1.463096.

[351] COSMOtherm, C3.0, release 1601, COSMOlogic GmbH & Co KG, http://www.cosmologic.de.

[352] P. Pracht, D. F. Grant, and S. Grimme, *Comprehensive Assessment of GFN Tight-Binding and Composite Density Functional Theory Methods for Calculating Gas-Phase Infrared Spectra*, J. Chem. Theory Comput. **16** (2020) 7044, DOI: 10.1021/acs.jctc.0c00877.

[353] F. Furche, R. Ahlrichs, C. Hättig, W. Klopper, M. Sierka, and F. Weigend, *Turbomole*, Wiley Interdiscip. Rev. Comput. Mol. Sci. **4** (2014) 91, DOI: 10.1002/wcms.1162.

[354] R. Ahlrichs, M. Bär, M. Häser, H. Horn, and C. Kölmel, *Electronic Structure Calculations on Workstation Computers: The Program System Turbomole*, Chem. Phys. Lett. (1989) 165.

[355] *TURBOMOLE V7.5.1 **2020**, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007; available from http://www.turbomole.com.*

[356] E. F. Pettersen et al., *UCSF Chimera—A visualization system for exploratory research and analysis*, J. Comput. Chem. **25** (2004) 1605, DOI: 10.1002/jcc.20084.

[357] T. Williams and C. Kelley, *Gnuplot 5.0: an interactive plotting program*, http://gnuplot.sourceforge.net/, 2018.

[358] D.A. Caseet al., 2021, Amber 2021, University of California, San Francisco.

[359] C. N. Nguyen, T. Kurtzman Young, and M. K. Gilson, *Grid inhomogeneous solvation theory: Hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril*, J. Chem. Phys. **137** (2012) 044101, DOI: 10.1063/1.4733951.

[360] C. Lee, W. Yang, and R. G. Parr, *Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density*, Phys. Rev. B **37** (1988) 785, DOI: 10.1103/PhysRevB.37.785.

[361] F. Weigend and R. Ahlrichs, *Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy*, Phys. Chem. Chem. Phys. **7** (2005) 3297, DOI: 10.1039/B508541A.

[362] C. Kalai, M. Alikhani, and E. Zins, *The molecular electrostatic potential analysis of solutes and water clusters: a straightforward tool to predict the geometry of the most stable micro-hydrated complexes of β-propiolactone and formamide.*, Theor. Chem. Acc. **137** (2018) 144, DOI: 10.1007/s00214-018-2345-6.

[363] S. A. Katsyuba, S. Spicher, T. P. Gerasimova, and S. Grimme, *Revisiting conformations of methyl lactate in water and methanol*, J. Chem. Phys. **155** (2021) 024507, DOI: 10.1063/5.0057024.

[364] S. A. Katsyuba, S. Spicher, T. P. Gerasimova, and S. Grimme, *Fast and Accurate Quantum Chemical Modeling of Infrared Spectra of Condensed-Phase Systems*, J. Phys. Chem. B **124** (2020) 6664, DOI: 10.1021/acs.jpcb.0c05857.

[365] S. A. Katsyuba, T. P. Gerasimova, S. Spicher, F. Bohle, and S. Grimme, *Computer-aided simulation of infrared spectra of ethanol conformations in gas, liquid and in CCl4 solution*, J. Comput. Chem. **43** (2022) 279, DOI: 10.1002/jcc.26788.

[366] K. Bünnemann and C. Merten, *Solvation of a chiral carboxylic acid: effects of hydrogen bonding on the IR and VCD spectra of α-methoxyphenylacetic acid*, Phys. Chem. Chem. Phys. **19** (2017) 18948, DOI: 10.1039/C7CP02049J.

[367] J. Gorges, S. Grimme, A. Hansen, and P. Pracht, *Towards understanding solvation effects on the conformational entropy of non-rigid molecules*, Phys. Chem. Chem. Phys. **24** (2022) 12249, DOI: 10.1039/D1CP05805C.

[368] O. Gutten, P. Jurečka, Z. Aliakbar Tehrani, M. Buděšínský, J. Řezáč, and L. Rulíšek, *Conformational energies and equilibria of cyclic dinucleotides in vacuo and in solution: computational chemistry vs. NMR experiments*, Phys. Chem. Chem. Phys. **23** (2021) 7280, DOI: 10.1039/D0CP05993E.

[369] P. J. Dyson and P. G. Jessop, *Solvent effects in catalysis: rational improvements of catalysts via manipulation of solvent interactions*, Catal. Sci. Technol. **6** (2016) 3302, DOI: 10.1039/C5CY02197A.

[370] A. Muhammad, G. Di Carmine, L. Forster, and C. D'Agostino, *Solvent Effects in the Homogeneous Catalytic Reduction of Propionaldehyde with Aluminium Isopropoxide Catalyst: New Insights from PFG NMR and NMR Relaxation Studies*, ChemPhysChem **21** (2020) 1101, DOI: 10.1002/CPHC.202000267.

[371] S. Wernersson, S. Birgersson, and M. Akke, *Cosolvent Dimethyl Sulfoxide Influences Protein-Ligand Binding Kinetics via Solvent Viscosity Effects: Revealing the Success Rate of Complex Formation Following Diffusive Protein-Ligand Encounter*, Biochemistry **62** (2023) 44, DOI: 10.1021/ACS.BIOCHEM.2C00507.

[372] S. M. Gopal, F. Klumpers, C. Herrmann, and L. V. Schäfer, *Solvent effects on ligand binding to a serine protease*, Phys. Chem. Chem. Phys. **19** (2017) 10753, DOI: 10.1039/C6CP07899K.

[373] S. Schmid and T. Hugel, *Controlling protein function by fine-tuning conformational flexibility*, eLife **9** (2020) e57180, DOI: 10.7554/ELIFE.57180.

[374] J. H. Ha and S. N. Loh, *Protein Conformational Switches: From Nature to Design*, Chem. Eur. J. **18** (2012) 7984, DOI: 10.1002/CHEM.201200348.

[375] M. C. Bellissent-Funel et al., *Water Determines the Structure and Dynamics of Proteins*, Chem. Rev. **116** (2016) 7673, DOI: 10.1021/ACS.CHEMREV.5B00664.

[376] P. Radivojac et al., *A large-scale evaluation of computational protein function prediction*, Nat. Methods **10** (2013) 221, DOI: 10.1038/nmeth.2340.

[377] W. Nowak, "Applications of Computational Methods to Simulations of Protein Dynamics", *Handbook of Computational Chemistry*, Dordrecht: Springer Netherlands, 2016 1, ISBN: 978-94-007-6169-8, DOI: 10.1007/978-94-007-6169-8_31-2.

[378] Z. Su and Y. Wu, *Computational studies of protein–protein dissociation by statistical potential and coarse-grained simulations: a case study on interactions between colicin E9 endonuclease and immunity proteins*, Phys. Chem. Chem. Phys. **21** (2019) 2463, DOI: 10.1039/C8CP05644G.

[379] J. Řezáč, *Non-Covalent Interactions Atlas benchmark data sets 5: London dispersion in an extended chemical space*, Phys. Chem. Chem. Phys. **24** (2022) 14780, DOI: 10.1039/D2CP01602H.

[380] S. Dasgupta, E. Lambros, J. P. Perdew, and F. Paesani, *Elevating density functional theory to chemical accuracy for water simulations through a density-corrected many-body formalism*, Nat. Commun. **12** (2021) 1, DOI: 10.1038/s41467-021-26618-9.

[381] A. Cumberworth, J. M. Bui, and J. Gsponer, *Free energies of solvation in the context of protein folding: Implications for implicit and explicit solvent models*, J. Comput. Chem. **37** (2016) 629, DOI: 10.1002/JCC.24235.

[382] G. Norjmaa, G. Ujaque, and A. Lledós, *Beyond Continuum Solvent Models in Computational Homogeneous Catalysis*, Top. Catal. **65** (2022) 118, DOI: 10.1007/S11244-021-01520-2.

[383] M. Rahbar and C. J. Stein, *A Statistical Perspective on Microsolvation*, The J. Phys. Chem. A **127** (2023) 2176, DOI: 10.1021/acs.jpca.2c08763.

[384] H. Valdes, K. Pluháčková, M. Pitonák, J. Řezáč, and P. Hobza, *Benchmark database on isolated small peptides containing an aromatic side chain: comparison between wave function and density functional theory methods and empirical force field*, Phys. Chem. Chem. Phys. **10** (2008) 2747, DOI: 10.1039/B719294K.

[385] H.-J. Werner and A. Hansen, *Accurate Calculation of Isomerization and Conformational Energies of Larger Molecules Using Explicitly Correlated Local Coupled Cluster Methods in Molpro and ORCA*, J. Chem. Theory Comput. **19** (2023) 7007, DOI: 10.1021/ACS.JCTC.3C00270.

[386] C. Plett and S. Grimme, *Automated and Efficient Generation of General Molecular Aggregate Structures*, Angew. Chem. Int. Ed. **62** (2023) e202214477, DOI: 10.1002/anie.202214477.

[387] F. Neese and J. Wiley, *The ORCA program system*, Wiley Interdiscip. Rev. Comput. Mol. Sci. **2** (2012) 73, DOI: 10.1002/WCMS.81.

[388] M. Garcia-Ratés and F. Neese, *Effect of the Solute Cavity on the Solvation Energy and its Derivatives within the Framework of the Gaussian Charge Scheme*, J. Comput. Chem. **41** (2020) 922, DOI: 10.1002/JCC.26139.

[389] C. Riplinger, P. Pinski, U. Becker, E. F. Valeev, and F. Neese, *Sparse maps - A systematic infrastructure for reduced-scaling electronic structure methods. II. Linear scaling domain based pair natural orbital coupled cluster theory*, J. Chem. Phys. **144** (2016) 24109, DOI: 10.1063/1.4939030.

[390] C. Riplinger, B. Sandhoefer, A. Hansen, and F. Neese, *Natural triple excitations in local coupled cluster calculations with pair natural orbitals*, J. Chem. Phys. **139** (2013) 134101, DOI: 10.1063/1.4821834.

[391] K. Sorathia and D. P. Tew, *Basis set extrapolation in pair natural orbital theories*, J. Chem. Phys. **153** (2020) 174112, DOI: 10.1063/5.0022077.

[392] H. J. Werner et al., *The Molpro quantum chemistry package*, J. Chem. Phys. **152** (2020) 144107, DOI: 10.1063/5.0005081.

[393] H.-J. Werner et al., *MOLPRO, 2022.3 , a package of ab initio programs*.

[394] F. Weigend, *Accurate Coulomb-fitting basis sets for H to Rn*, Phys. Chem. Chem. Phys. **8** (2006) 1057, DOI: 10.1039/B515623H.

[395] B. Hourahine et al., *DFTB+, a software package for efficient approximate density functional theory based atomistic simulations*, J. Chem. Phys. **152** (2020) 124101, DOI: 10.1063/1.5143190.

[396] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, *Open Babel: An Open chemical toolbox*, J. Cheminform. **3** (2011) 1, DOI: 10.1186/1758-2946-3-33.

[397] R. Pollice, F. Fleckenstein, I. Shenderovich, and P. Chen, *Compensation of London Dispersion in the Gas Phase and in Aprotic Solvents*, Angew. Chem. Int. Ed. **58** (2019) 14281, DOI: 10.1002/ANIE.201905436.

[398] N. Mardirossian and M. Head-Gordon, *Survival of the most transferable at the top of Jacob's ladder: Defining and testing the ω B97M(2) double hybrid density functional*, J. Chem. Phys. **148** (2018) 241736, DOI: 10.1063/1.5025226.

[399] L. Goerigk and S. Grimme, *Efficient and accurate double-hybrid-meta-GGA density functionals- evaluation with the extended GMTKN30 database for general main group thermochemistry, kinetics, and noncovalent interactions*, J. Chem. Theory Comput. **7** (2011) 291, DOI: 10.1021/CT100466K.

[400] A. Najibi and L. Goerigk, *DFT-D4 counterparts of leading meta-generalized-gradient approximation and hybrid density functionals for energetics and geometries*, J. Comput. Chem. **41** (2020) 2562, DOI: 10.1002/JCC.26411.

[401] N. Mardirossian and M. Head-Gordon, *ωB97X-V: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy*, Phys. Chem. Chem. Phys. **16** (2014) 9904, DOI: 10.1039/C3CP54374A.

[402] N. Mardirossian and M. Head-Gordon, *ωB97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation*, J. Chem. Phys. **144** (2016) 214110, DOI: 10.1063/1.4952647.

[403] M. Bursch, H. Neugebauer, S. Ehlert, and S. Grimme, *Dispersion corrected r2SCAN based global hybrid functionals: r^2SCANh, r^2SCAN0, and r^2SCAN50*, J. Chem. Phys. **156** (2022) 134105, DOI: 10.1063/5.0086040.

[404] H. S. Yu, X. He, and D. G. Truhlar, *MN15-L: A New Local Exchange-Correlation Functional for Kohn-Sham Density Functional Theory with Broad Accuracy for Atoms, Molecules, and Solids*, J. Chem. Theory Comput. **12** (2016) 1280, DOI: 10.1021/ACS.JCTC.5B01082.

[405] J. Tao, J. P. Perdew, V. N. Staroverov, and G. E. Scuseria, *Climbing the density functional ladder: Nonempirical meta–generalized gradient approximation designed for molecules and solids*, Phys. Rev. Lett. **91** (2003) 146401, DOI: 10.1103/PHYSREVLETT.91.146401.

[406] S. Grimme, *Semiempirical GGA-type density functional constructed with a long-range dispersion correction*, J. Comput. Chem. **27** (2006) 1787, DOI: 10.1002/JCC.20495.

[407] J. P. Perdew, K. Burke, and M. Ernzerhof, *Generalized gradient approximation made simple*, Phys. Rev. Lett. **77** (1996) 3865, DOI: 10.1103/PhysRevLett.77.3865.

[408] S. Grimme, *Improved second-order Møller–Plesset perturbation theory by separate scaling of parallel- and antiparallel-spin pair correlation energies*, J. Chem. Phys. **118** (2003) 9095, DOI: 10.1063/1.1569242.

[409] J. G. Brandenburg, C. Bannwarth, A. Hansen, and S. Grimme, *B97-3c: A revised low-cost variant of the B97-D density functional method*, J. Chem. Phys. **148** (2018) 64104, DOI: 10.1063/1.5012601.

[410] J. Řezáč, K. E. Riley, and P. Hobza, *Benchmark calculations of noncovalent interactions of halogenated molecules*, J. Chem. Theory Comput. **8** (2012) 4285, DOI: 10.1021/CT300647K.

[411] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, *Development and testing of a general amber force field*, J. Comput. Chem. **25** (2004) 1157, DOI: 10.1002/JCC.20035.

[412] T. Hassinen and M. Peräkylä, *New energy terms for reduced protein models implemented in an off-lattice force field*, J. Comput. Chem. **22** (2001) 1229, DOI: 10.1002/JCC.1080.

[413] J. L. Bao, L. Gagliardi, and D. G. Truhlar, *Self-Interaction Error in Density Functional Theory: An Appraisal*, J. Phys. Chem. Lett. **9** (2018) 2353, DOI: 10.1021/ACS.JPCLETT.8B00242.

[414] I. V. Kolesnichenko and E. V. Anslyn, *Practical applications of supramolecular chemistry*, Chem. Soc. Rev. **46** (2017) 2385, DOI: 10.1039/C7CS00078B.

[415] M. S. Gordon, J. M. Mullin, S. R. Pruitt, L. B. Roskop, L. V. Slipchenko, and J. A. Boatz, *Accurate methods for large molecular systems*, J. Phys. Chem. B **113** (2009) 9646.

[416] S. Kobayashi and H. Uyama, *Biomacromolecules and Bio-Related Macromolecules*, Macromol. Chem. Phys. **204** (2003) 235, DOI: 10.1002/MACP.200290084.

[417] R. Morris, K. A. Black, and E. J. Stollar, *Uncovering protein function: from classification to complexes*, Essays Biochem. **66** (2022) 255, DOI: 10.1042/EBC20200108.

[418] A. C. Ghosh et al., *Rhodium-Based Metal-Organic Polyhedra Assemblies for Selective CO_2 Photoreduction*, J. Am. Chem. Soc. **144** (2022) 3626, DOI: 10.1021/JACS.1C12631.

[419] F. Lancia, A. Ryabchun, and N. Katsonis, *Life-like motion driven by artificial molecular machines*, Nat. Rev. Chem. **3** (2019) 536, DOI: 10.1038/s41570-019-0122-2.

[420] J. Kohn, S. Spicher, M. Bursch, and S. Grimme, *Quickstart guide to model structures and interactions of artificial molecular muscles with efficient computational methods*, Chem. Commun. **58** (2021) 258, DOI: 10.1039/D1CC05759F.

[421] S. Mallakpour, E. Nikkhoo, and C. M. Hussain, *Application of MOF materials as drug delivery systems for cancer therapy and dermal treatment*, Coord. Chem. Rev. **451** (2022) 214262, DOI: 10.1016/J.CCR.2021.214262.

[422] J. Cao, X. Li, and H. Tian, *Metal-Organic Framework (MOF)-Based Drug Delivery*, Curr. Med. Chem. **27** (2019) 5949, DOI: 10.2174/0929867326666190618152518.

[423] R. F. Mendes and F. A. A. Paz, *Transforming metal–organic frameworks into functional materials*, Inorg. Chem. Front. **2** (2015) 495, DOI: 10.1039/C4QI00222A.

[424] A. Bavykina, N. Kolobov, I. S. Khan, J. A. Bau, A. Ramirez, and J. Gascon, *Metal–organic frameworks in heterogeneous catalysis: recent progress, new trends, and future perspectives*, Chem. Rev. **120** (2020) 8468.

[425] M. L. Hu et al., *Taking organic reactions over metal-organic frameworks as heterogeneous catalysis*, Micropor. Mesopor. Mat. **256** (2018) 111, DOI: 10.1016/J.MICROMESO.2017.07.057.

[426] A. Lledós, *Computational Organometallic Catalysis: Where We Are, Where We Are Going*, Eur. J. Inorg. Chem. **2021** (2021) 2547, DOI: 10.1002/ejic.202100330.

[427] T. S. Hofer, *From macromolecules to electrons-grand challenges in theoretical and computational chemistry*, Front. Chem. **1** (2013) 6, DOI: 10.3389/FCHEM.2013.00006.

[428] W. Thiel, *Semiempirical quantum–chemical methods*, Wiley Interdiscip. Rev. Comput. Mol. Sci. **4** (2014) 145, DOI: 10.1002/WCMS.1161.

[429] S. Piana, P. Robustelli, D. Tan, S. Chen, and D. E. Shaw, *Development of a Force Field for the Simulation of Single-Chain Proteins and Protein-Protein Complexes*, J. Chem. Theory Comput. **16** (2020) 2494, DOI: 10.1021/ACS.JCTC.9B00251.

[430] F. Spiegelman et al., *Density-functional tight-binding: basic concepts and applications to molecules and clusters*, Adv. Phys. X **5** (2020) 1710252, DOI: 10.1080/23746149.2019.1710252.

[431] S. Grimme, F. Bohle, A. Hansen, P. Pracht, S. Spicher, and M. Stahn, *Efficient Quantum Chemical Calculation of Structure Ensembles and Free Energies for Nonrigid Molecules*, J. Phys. Chem. A **125** (2021) 4039, DOI: 10.1021/ACS.JPCA.1C00971.

[432] Y. Q. Chen, Y. J. Sheng, Y. Q. Ma, and H. M. Ding, *Efficient calculation of protein–ligand binding free energy using GFN methods: the power of the cluster model*, Phys. Chem. Chem. Phys. **24** (2022) 14339, DOI: 10.1039/D2CP00161F.

[433] N. Koga, R. Koga, G. Liu, J. Castellanos, G. T. Montelione, and D. Baker, *Role of backbone strain in de novo design of complex α/β protein structures*, Nat. Commun. **12** (2021) 1, DOI: 10.1038/s41467-021-24050-7.

[434] A. Galan and P. Ballester, *Stabilization of reactive species by supramolecular encapsulation*, Chem. Soc. Rev. **45** (2016) 1720, DOI: 10.1039/C5CS00861A.

[435] G. Olivo, G. Capocasa, D. Del Giudice, O. Lanzalunga, and S. Di Stefano, *New horizons for catalysis disclosed by supramolecular chemistry*, Chem. Soc. Rev. **50** (2021) 7681.

[436] J. Yang, L. Zhao, Y. Sun, and H. Sun, *ONIOM Study of Isomerization Reactions of Aromatic Hydrocarbons in Acidic Mordenite Zeolite*, ChemPhysChem **10** (2009) 946, DOI: 10.1002/CPHC.200800785.

[437] M. Ruan et al., *Computational study on the hydrolysis of halomethanes*, Theor. Chem. Acc. **137** (2018) 1, DOI: 10.1007/S00214-018-2389-7.

[438] I. Geronimo and P. Paneth, *A DFT and ONIOM study of C–H hydroxylation catalyzed by nitrobenzene 1,2-dioxygenase*, Phys. Chem. Chem. Phys. **16** (2014) 13889, DOI: 10.1039/C4CP01030B.

[439] C. E. Tzeliou, M. A. Mermigki, and D. Tzeli, *Review on the QM/MM Methodologies and Their Application to Metalloproteins*, Molecules **27** (2022) 2660, DOI: 10.3390/MOLECULES27092660.

[440] F. Maseras and K. Morokuma, *IMOMM: A new integrated ab initio + molecular mechanics geometry optimization scheme of equilibrium structures and transition states*, J. Comput. Chem. **16** (1995) 1170, DOI: 10.1002/JCC.540160911.

[441] A. Fu, W. Meng, H. Li, J. Nie, and J. A. Ma, *A density functional study of chiral phosphoric acid-catalyzed direct arylation of trifluoromethyl ketone and diarylation of methyl ketone: Reaction mechanism and the important role of the CF_3 group*, Org. Biomol. Chem. **12** (2014) 1908, DOI: 10.1039/c3ob42157k.

[442] G. Mora, B. Deschamps, S. Van Zutphen, X. F. Le Goff, L. Ricard, and P. Le Floch, *Xanthene-phosphole ligands: Synthesis, coordination chemistry, and activity in the palladium-catalyzed amine allylation*, Organometallics **26** (2007) 1846, DOI: 10.1021/om061172t.

[443] Y. Li et al., *A DFT study on the distributions of Al and Brönsted acid sites in zeolite MCM-22*, J. Mol. Catal. A Chem. **338** (2011) 24, DOI: 10.1016/j.molcata.2011.01.018.

[444] J. Řezáč and J. J. Stewart, *How well do semiempirical QM methods describe the structure of proteins?*, J. Chem. Phys. **158** (2023) 44118, DOI: 10.1063/5.0135091.

[445] S. Schmitz, J. Seibert, K. Ostermeir, A. Hansen, A. H. Göller, and S. Grimme, *Quantum Chemical Calculation of Molecular and Periodic Peptide and Protein Structures*, J. Phys. Chem. B **124** (2020) 3636, DOI: 10.1021/acs.jpcb.0c00549.

[446] M. Bursch, A. Hansen, and S. Grimme, *Fast and Reasonable Geometry Optimization of Lanthanoid Complexes with an Extended Tight Binding Quantum Chemical Method*, Inorg. Chem. **56** (2017) 12485, DOI: 10.1021/acs.inorgchem.7b01950.

[447] F. Neese, *Software update: The ORCA program system—Version 5.0*, Wiley Interdiscip. Rev. Comput. Mol. Sci. **12** (2022) e1606, DOI: 10.1002/WCMS.1606.

[448] *TURBOMOLE V7.6 2021, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007; available from http://www.turbomole.com.*

[449] R. Ahlrichs, M. Bär, M. Häser, H. Horn, and C. Kölmel, *Electronic structure calculations on workstation computers: The program system turbomole*, Chem. Phys. Lett. **162** (1989) 165, DOI: 10.1016/0009-2614(89)85118-8.

[450] L. Böselt, M. Thürlemann, and S. Riniker, *Machine Learning in QM/MM Molecular Dynamics Simulations of Condensed-Phase Systems*, J. Chem. Theory Comput. **17** (2021) 2641, DOI: 10.1021/ACS.JCTC.0C01112.

[451] G. A. Bramley, O. T. Beynon, P. V. Stishenko, and A. J. Logsdail, *The application of QM/MM simulations in heterogeneous catalysis*, Phys. Chem. Chem. Phys. **25** (2023) 6562, DOI: 10.1039/D2CP04537K.

[452] M. J. Field, P. A. Bash, and M. Karplus, *A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations*, J. Comput. Chem. **11** (1990) 700, DOI: 10.1002/JCC.540110605.

[453] H. M. Senn and W. Thiel, *QM/MM Methods for Biological Systems*, Top Curr. Chem. **268** (2007) 173, DOI: 10.1007/128_2006_084.

[454] G. Tóth and G. Náray-Szabó, *Novel semiempirical method for quantum Monte Carlo simulation: Application to amorphous silicon*, J. Chem. Phys. **100** (1994) 3742, DOI: 10.1063/1.466361.

[455] Semiempirical Extended Tight-Binding Program Package, 2023.

[456] D. Andrae, U. Häußermann, M. Dolg, H. Stoll, and H. Preuß, *Energy-adjusted ab initio pseudopotentials for the second and third row transition elements*, Theor. Chim. Acta **77** (1990) 123, DOI: 10.1007/BF01114537.

[457] E. Caldeweyher, J. M. Mewes, S. Ehlert, and S. Grimme, *Extension and evaluation of the D4 London-dispersion model for periodic systems*, Phys. Chem. Chem. Phys. **22** (2020) 8499, DOI: 10.1039/d0cp00502a.

[458] S. Dohm, M. Bursch, A. Hansen, and S. Grimme, *Semiautomated Transition State Localization for Organometallic Complexes with Semiempirical Quantum Chemical Methods*, J. Chem. Theory Comput. **16** (2020) 8, DOI: 10.1021/acs.jctc.9b01266.

[459] P. M. Zimmerman, *Growing string method with interpolation and optimization in internal coordinates: Method and examples*, J. Chem. Phys. **138** (2013) 184102, DOI: 10.1063/1.4804162.

[460] P. M. Zimmerman, *Reliable Transition State Searches Integrated with the Growing String Method.*, J. Chem. Theory Comput. **9** (2013) 3043, DOI: 10.1021/CT400319W.

[461] M. Jafari and P. M. Zimmerman, *Reliable and efficient reaction path and transition state finding for surface reactions with the growing string method*, J. Comput. Chem. **38** (2017) 645, DOI: 10.1002/JCC.24720.

[462] J. P. Perdew, J. Tao, V. N. Staroverov, and G. E. Scuseria, *Meta-generalized gradient approximation: Explanation of a realistic nonempirical density functional*, J. Chem. Phys. **120** (2004) 6898, DOI: 10.1063/1.1665298.

[463] C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward, *The Cambridge structural database*, Acta. Crystallogr. B. Struct. Sci. Cryst. Eng. Mater. **72** (2016) 171, DOI: 10.1107/S2052520616003954.

[464] R. C. Mohs and N. H. Greig, *Drug discovery and development: Role of basic biological research*, Alzheimer's Dement.: Transl. Res. Clin. Interv. **3** (2017) 651, DOI: 10.1016/j.trci.2017.10.005.

[465] A. L. Cunningham et al., *Vaccine development: From concept to early clinical testing*, Vaccine **34** (2016) 6655, DOI: 10.1016/j.vaccine.2016.10.016.

[466] J. J. Sheng and J. P. Jin, *TNNI1, TNNI2 and TNNI3: Evolution, regulation, and protein structure-function relationships*, Gene **576** (2016) 385, DOI: 10.1016/j.gene.2015.10.052.

[467] J. C. Whisstock and A. M. Lesk, *Prediction of protein function from protein sequence and structure*, Q. Rev. Biophys. **36** (2003) 307, DOI: 10.1017/S0033583503003901.

[468] M. Dorn, M. B. E Silva, L. S. Buriol, and L. C. Lamb, *Three-dimensional protein structure prediction: Methods and computational strategies*, Comput. Biol. Chem. **53** (2014) 251, DOI: 10.1016/j.compbiolchem.2014.10.001.

[469] L. M. Bertoline, A. N. Lima, J. E. Krieger, and S. K. Teixeira, *Before and after AlphaFold2: An overview of protein structure prediction*, Front. bioinform. **3** (2023) 1120370, DOI: 10.3389/fbinf.2023.1120370.

[470] B. Huang et al., *Protein Structure Prediction: Challenges, Advances, and the Shift of Research Paradigms*, Genom. Proteom. Bioinform. **21** (2023) 913, DOI: 10.1016/J.GPB.2022.11.014.

[471] B. T. Kaynak et al., *Sampling of Protein Conformational Space Using Hybrid Simulations: A Critical Assessment of Recent Methods*, Front. Mol. Biosci. **9** (2022) 832847, DOI: 10.3389/FMOLB.2022.832847.

[472] Z. Hazarika, S. Rajkhowa, and A. Nath Jha, "Role of Force Fields in Protein Function Prediction", *Homology Molecular Modeling - Perspectives and Applications*, IntechOpen, 2021, DOI: 10.5772/intechopen.93901.

[473] B. Kuhlman and P. Bradley, *Advances in protein structure prediction and design*, Nat. Rev. Mol. Cell Biol. **20** (2019) 681, DOI: 10.1038/s41580-019-0163-x.

[474] M. Torrisi, G. Pollastri, and Q. Le, *Deep learning methods in protein structure prediction*, Comput. Struct. Biotechnol. J. **18** (2020) 1301, DOI: 10.1016/j.csbj.2019.12.011.

[475] L. Maveyraud and L. Mourey, *Protein X-ray Crystallography and Drug Discovery*, Molecules **25** (2020) 1030, DOI: 10.3390/MOLECULES25051030.

[476] M. Carroni and H. R. Saibil, *Cryo electron microscopy to determine the structure of macromolecular complexes*, Methods **95** (2016) 78, DOI: 10.1016/J.YMETH.2015.11.023.

[477] Y. Hu et al., *NMR-Based Methods for Protein Analysis*, Anal. Chem. **93** (2021) 1866, DOI: 10.1021/ACS.ANALCHEM.0C03830.

[478] Y. Yuan, M. J. Mills, P. L. Popelier, and F. Jensen, *Comprehensive analysis of energy minima of the 20 natural amino acids*, J. Phys. Chem. A **118** (2014) 7876, DOI: 10.1021/jp503460m.

[479] V. K. Prasad, A. Otero-De-la-Roza, and G. A. Dilabio, *Data descriptor: Pepconf, a diverse data set of peptide conformational energies*, Sci. Data **6** (2019) 1, DOI: 10.1038/sdata.2018.310.

[480] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, *ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB*, J. Chem. Theory Comput. **11** (2015) 3696, DOI: 10.1021/ACS.JCTC.5B00255.

[481] RDKit: Open-source cheminformatics. https://www.rdkit.org.

[482] C. Plett, S. Grimme, and A. Hansen, *Conformational energies of biomolecules in solution: Extending the MPCONF196 benchmark with explicit water molecules*, J. Comput. Chem. **45** (2024) 419, DOI: 10.1002/jcc.27248.

[483] *Grimme Lab on GitHub*, **2024**, https://github.com/grimme-lab.

[484] B. P. Pritchard, D. Altarawy, B. Didier, T. D. Gibson, and T. L. Windus, *New Basis Set Exchange: An Open, Up-to-Date Resource for the Molecular Sciences Community*, J. Chem. Inf. Model. **59** (2019) 4814, DOI: 10.1021/acs.jcim.9b00725.

[485] G. L. Stoychev, A. A. Auer, and F. Neese, *Automatic Generation of Auxiliary Basis Sets*, J. Chem. Theory Comput. **13** (2017) 554, DOI: 10.1021/ACS.JCTC.6B01041.

[486] *Adjusted D3(BJ) parameters*, **2017**, https://www.rezacovi.cz/science/dft-d3.html.

[487] G. Santra, M. Cho, and J. M. Martin, *Exploring Avenues beyond Revised DSD Functionals: I. Range Separation, with xDSD as a Special Case*, J. Phys. Chem. A **125** (2021) 4614, DOI: 10.1021/ACS.JPCA.1C01294.

[488] B. Hourahine et al., *Erratum: DFTB+, a software package for efficient approximate density functional theory based atomistic simulations (Journal of Chemical Physics (2020) 152 (124101) DOI: 10.1063/1.5143190)*, J. Chem. Phys. **157** (2022) 124101.

[489] C. Devereux et al., *Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens*, J. Chem. Theory Comput. **16** (2020) 4192, DOI: 10.1021/ACS.JCTC.0C00121.

[490] P. O. Dral, *MLatom: A program package for quantum chemical research assisted by machine learning*, J. Comput. Chem. **40** (2019) 2339, DOI: 10.1002/JCC.26004.

[491] A. Hjorth Larsen et al., *The atomic simulation environment—a Python library for working with atoms*, J. Condens. Matter Phys. **29** (2017) 273002, DOI: 10.1088/1361-648X/AA680E.

[492] L. Goerigk, A. Karton, J. M. L. Martin, and L. Radom, *Accurate quantum chemical energies for tetrapeptide conformations: why MP2 data with an insufficient basis set should be handled with caution*, Phys. Chem. Chem. Phys. **15** (2013) 7028, DOI: 10.1039/C3CP00057E.

List of Figures

List of Tables

Acknowledgements

Conducting scientific research and embarking on the journey to a PhD would not have been possible without the support of many people, and I would like to take this opportunity to acknowledge them.

First, I wish to thank my PhD supervisor, Prof. Dr. Stefan Grimme, who gave me the freedom to pursue my own projects and ideas while always offering advice when needed. He warmly welcomed me into his group and introduced me to the fascinating world of science, which enabled me to contribute to many projects and meet inspiring people from diverse fields.

I also thank Prof. Dr. Thomas Bredow for kindly agreeing to review this thesis.

I am deeply thankful to Dr. Andreas Hansen, with whom I had the pleasure of collaborating on various projects, and who also supported me constantly throughout my journey. Further, I want to thank Jun.-Prof. Dr. Ala Bunescu for collaborating with me in such a constructive and enjoyable atmosphere. I really enjoyed our meetings.

Special thanks go to my friend and long-term office mate, Thomas Gasevic, for the great time we shared, as well as to Abylay Katbashev, Christian Selzer, and Christopher Staudt, who made our office an enjoyable place to work.

I am grateful to Dr. Sebastian Spicher and Dr. Markus Bursch for teaching me many valuable scientific and practical lessons. I would like to thank Dr. Marcel Stahn, with whom I not only had a great time working together, but also had a lot of fun traveling through Chicago. Further, I am grateful to Linda Nelles-Ziegler for our fruitful collaboration.

For proofreading my thesis, I am especially thankful to Dr. Marcel Müller, Benedikt Bädorf, Thomas Gasevic, and Tim Schramm. I also thank Jens Mekelburger and Claudia Kronz for their steady technical and administrative support.

I thoroughly enjoyed my time at the Mulliken Center, not least because of the many colleagues who accompanied me along the way, especially those not mentioned yet: Benedikt Bädorf, Robin Dahl, Dr. Sebastian Ehlert, Marvin Friede, Thomas Froitzheim, Johannes Gorges, Dr. Julia Kohn, Dr. Jeroen Koopman, Dr. Julius Kleine Büning, Lukas Kunze, Dr. Jan Mewes, Dr. Hagen Neugebauer, Dr. Philipp Pracht, Dr. Thomas Rose, Jonathan Schöps, Leopold Seidler, and Lukas Wittmann.

Finally, I wish to thank my friends and family. Above all, I express my heartfelt gratitude to my mother and aunt for their unconditional support throughout this journey, and to my wife, who, in times of exhaustion, has always lifted me up and given me the strength to carry on.