Development and Application of Automated Quantum Chemical Workflows for the Computation of Non-Covalent Interactions and Mass Spectra

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

vorgelegt von
Johannes Gorges
aus
Andernach

Bonn, August 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:	31.10.2025
Erscheinungsjahr:	2025

Publications

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

1. J. Gorges, S. Grimme, and A. Hansen, *Reliable prediction of association (free) energies of supramolecular complexes with heavy main group elements - the HS13L benchmark set*, Phys. Chem. Chem. Phys. **24**.47 (2022) 28831, DOI: 10.1039/d2cp04049b.

2. J. Gorges, B. Bädorf, S. Grimme, and A. Hansen, *Efficient Computation of the Interaction Energies of Very Large Non-covalently Bound Complexes*, Synlett **34**.10 (2022) 1135, DOI: 10.1055/s-0042-1753141.

3. J. Gorges and S. Grimme, *QCxMS2 - a program for the calculation of electron ionization mass spectra via automated reaction network discovery*, Phys. Chem. Chem. Phys. **27**.14 (2025) 6899, DOI: 10.1039/d5cp00316d.

At the time of submission of this thesis, one article is under peer review and publicly available on a preprint server:

4. J. Gorges, M. Engeser, and S. Grimme, *Evaluation of the QCxMS2 method for the calculation of collision-induced-dissociation spectra via automated reaction network exploration*, ChemRxiv Prepr. (2025), DOI: 10.26434/chemrxiv-2025-gcws2.

This manuscript has since been accepted for publication in: J. Gorges, M. Engeser, and S. Grimme, *Evaluation of the QCxMS2 Method for the Calculation of Collision-Induced Dissociation Spectra via Automated Reaction Network Exploration*, Journal of the American Society for Mass Spectrometry **36**.10 (2025) 2276, DOI: 10.1021/jasms.5c00234.

Significant contributions have been made to the following article.

5. 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**.20 (2022) 12249, DOI: 10.1039/d1cp05805c.

6. 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**.11 (2024) 114110, DOI: 10.1063/5.0197592.

Presentations and posters on conferences are listed below.

1. Poster on "Reliable prediction of association (free) energies of supramolecular complexes with heavy main group elements – the HS13L benchmark set," International Conference on Noncovalent Interactions (ICNI), **July 2022**, Strasbourg, France.

2. Poster on "Reliable prediction of association (free) energies of supramolecular complexes with heavy main group elements – the HS13L benchmark set," Bunsen-Tagung, **September 2022**, Gießen, Germany.

3. Poster on "Reliable prediction of association (free) energies of supramolecular complexes with heavy main group elements – the HS13L benchmark set," Symposium on Theoretical Chemistry, **September 2022**, Heidelberg, Germany.

4. Conference talk on "Reliable prediction of association (free) energies of supramolecular complexes with heavy main group elements – the HS13L benchmark set," ORCA User Meeting, **December 2022**, Online conference.

5. Poster on "Quantum chemical calculation of mass spectra via automated transition state search – the development of QCxMS2," International Congress of Quantum Chemistry (ICQC), **June 2023**, Bratislava, Slovakia.

6. Poster on "Automated and efficient computation of (free) energies in solution and simulation of mass spectra with quantum mechanical methods," International Summer Course BASF, **August 2023**, Ludwigshafen, Germany.

7. Poster on "Quantum chemical calculation of mass spectra via automated transition state search – the development of QCxMS2," GDCh-Wissenschaftsforum Chemie, **September 2023**, Leipzig, Germany.

8. Poster on "Quantum chemical calculation of mass spectra via automated transition state search," 60th Symposium on Theoretical Chemistry, **September 2024**, Braunschweig, Germany.

"Numerous stainless steel flanges and electronics cabinets were tempting to be explored and – whoops – infected me with CMSD (chronic mass spectrometry disease)."

– Jürgen H. Gross –

Abstract

The automated and efficient quantum chemical description of chemical systems is a central task in computational chemistry, as it enables the routine use of computational methods to interpret experimental results and to guide experiments through predictions. In particular, non-covalent interactions (NCIs) and mass spectrometry (MS) represent active areas of research where significant improvements are still required. Supramolecular complexes bound by NCIs play crucial roles in areas such as molecular recognition and catalysis, while MS is a powerful technique for structure elucidation in diverse fields, including metabolomics or proteomics. Given their importance, computational modeling of these systems, allowing for both the interpretation and prediction of experimental results, is essential but challenging due to their complexity. To address both aspects, this thesis is divided into two parts: the first focuses on benchmarking computational methods and workflows for the calculation of NCIs, while the second describes the development of a new program for the computation of MS. In the following Chapter 1, NCIs and MS are introduced, and the associated challenges and opportunities in their computation are outlined. An overview of the relevant quantum chemical methods and theoretical concepts to accurately describe these systems is given in Chapter 2. The typically large system size of NCI complexes requires the use of efficient approximate low-cost methods, such as density functional theory (DFT), semiempirical quantum mechanical (SQM) methods, or force field (FF) methods, whose accuracy has to be benchmarked against more accurate and robust reference methods. Chapter 3 presents such a benchmark study, which assesses the performance of low-cost DFT, SQM, and FF methods for calculating gas-phase interaction energies of 16 very large, non-covalently bound complexes. With system sizes of up to approximately 2000 atoms, this benchmark provides a unique testing ground for evaluating the robustness of computational methods applied to large molecular systems. To describe the binding behavior of supramolecular complexes, their lowest-energy geometry must be determined, taking into account their conformational flexibility, as well as thermal and solvation effects. These factors are explored through a dedicated benchmark study employing different methods in a multilevel workflow for direct comparison to experimental binding constants, which is summarized in Chapter 4. Herein, excellent agreement with the experimental reference values was achieved.

In the second part of this thesis, the computation of MS is investigated. The challenges in computing mass spectra arise from the high energies involved in the experiments, which result in a large number of possible fragmentation reactions that must be computed efficiently and in an automated manner. Since a generally applicable and sufficiently accurate quantum chemical approach for this task is still lacking, a new program, QCxMS2, was developed. Whereas existing quantum chemical approaches, such as QCxMS, are based on molecular dynamics simulations, QCxMS2 follows a novel strategy based on automated reaction discovery. Chapter 5 presents the development of QCxMS2 and its superior

agreement with experimental electron ionization (EI) mass spectra compared to its predecessor and main competitor, QCxMS. The following Chapter 6 describes the extension of QCxMS2 to enable the calculation of collision-induced dissociation (CID) mass spectra. As for EI-MS, a significant improvement over QCxMS was demonstrated through comparison with experimental spectra. Finally, in Chapter 7, the findings of this work are summarized, and their implications for the computational modeling of supramolecular binding and MS prediction are evaluated. In conclusion, the compiled benchmark sets in this work provide useful information on which method to employ for the efficient modeling of supramolecular complexes. Furthermore, the newly developed open-source software QCxMS2 provides a valuable tool that can be integrated into automated structure elucidation workflows for the identification of unknown compounds.

Contents

CHAPTER 1

Introduction

Understanding chemical systems at the atomistic and electronic scale is essential for advancing technologies of major societal relevance, such as the rational design of pharmaceutical drugs[8] and advanced materials.[9] While experimental methods have traditionally been the primary means of gaining such insights, they often reach their limits – particularly when it comes to unraveling detailed mechanisms of action or studying compounds that are difficult to synthesize or isolate. Computational chemistry fills this gap by enabling the exploration of molecular properties with high accuracy for essentially any thinkable structure. Thanks to sophisticated quantum chemical methods and modern computational infrastructure, computational chemistry now plays an integral role not only in academic research[10] but also in industrial applications.[11] This development is driven by both the emergence of new methods for electronic energy prediction and by automated workflows that combine these methods to compute composite properties that are not directly accessible.[12] In this thesis, such workflows are developed and applied to expand the scope of computational chemistry. Specifically, two important and broadly relevant areas are investigated: Non-covalent interactions (NCIs), which are fundamental to molecular recognition and self-assembly,[13] and mass spectrometry (MS), a key technique for the identification and structural elucidation of (unknown) compounds.[14]

NCIs play a central role in both chemistry and biology, governing life itself, as they determine the structure and function of supramolecular assemblies. A particularly prominent example is hydrogen bonding in the Watson-Crick base pairing of deoxyribonucleic acid, which underpins the very basis of life.[15] In practical applications, NCIs are widely exploited, e.g., for molecular sensing,[16] drug delivery,[17] molecular imaging,[18] or metal extraction.[13,19] From a theoretical perspective, a broad range of quantum chemical and classical methods is available to describe NCIs, each differing in computational cost and degree of empiricism. A schematic overview of the different classes of methods is shown in Figure 1.1. Wave function theory (WFT) methods are based on solving the Schrödinger equation and computing the many-electron wave function directly. They are often referred to as *ab initio* methods and are generally highly accurate but computationally expensive. Kohn-Sham density functional theory (DFT) employs the electron density instead of the wave function and is typically less computationally demanding than WFT methods, though it involves a higher degree of empiricism.[20] Semiempirical quantum mechanical (SQM) methods introduce further approximations, lowering the computational demand to enable calculations on larger systems.[21] Even faster are force field (FF) methods, which replace the electronic structure with classical atom-based interaction models.[22] In recent years, machine-learned interatomic potentials (MLIPs) have attracted growing attention.

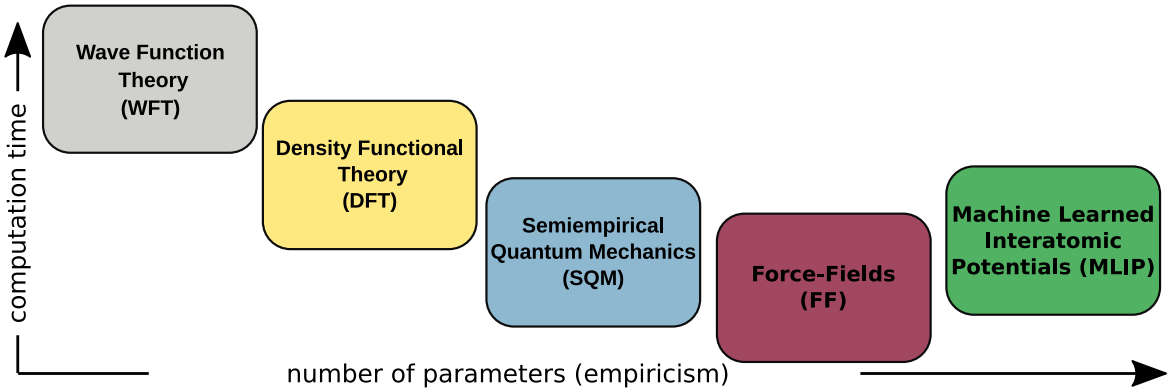


Figure 1.1: Schematic illustration of different classes of computational chemistry methods and their general trends in terms of computational cost and degree of empiricism.

These approaches aim to learn physical interactions from large datasets using neural networks or other regression models. Very recently, MLIPs such as AIMNet2[23] and the UMA models[24] have demonstrated remarkable agreement with their reference methods (typically DFT), while operating at a computational cost comparable to SQM methods. Given the broad range of available methods and the fact that their accuracy can often only be estimated in advance, systematic benchmarking is essential to assess their reliability and identify their strengths and limitations. This becomes increasingly important with higher degrees of empiricism. Benchmarking is typically performed by comparing low-cost methods to a more accurate and well-established reference method or to experimental data. Herein, the so-called "gold standard" WFT method, CCSD(T), from coupled-cluster theory is frequently used as a theoretical reference.[25] Since supramolecular systems of interest are often very large, low-cost methods are required for their routine study. However, NCIs pose a particular challenge for computational methods and often require specialized corrections or theoretical treatments. One example is the treatment of London dispersion (LD), which is discussed in Section 2.1.4. Therefore, benchmarking is essential to ensure that these methods can be reliably applied to large systems. This is especially useful in the screening process of extensive compound libraries. In this context, one of the central objectives of this thesis is the systematic benchmarking of methods and workflows for the accurate description of systems governed by NCIs. This includes the investigation of electronic energies as well as thermal and solvation contributions to the Gibbs free energy, as described in Section 2.2.

The other key topic of this thesis is MS, which plays a central role in many areas of analytical chemistry.[14] Based on a straightforward principle – ionizing a molecule to transfer it into the gas phase, inducing fragmentation, and directing the resulting charged fragments through magnetic and electric fields toward a mass detector – MS measures the mass-to-charge ratio (m/z) and relative intensities of the fragments to yield a characteristic fragmentation pattern for a given molecule. This makes MS applicable to a broad range of substances, offering high sensitivity and compatibility with high-throughput workflows. When MS is coupled to gas chromatography or liquid chromatography, it enables the analysis of complex mixtures with high chemical specificity and resolution. Electron ionization (EI), which uses an electron beam to ionize molecules, was the first ionization technique developed and is still widely used.[26] Since then, many alternative, "softer", ionization methods have been introduced that allow more controlled ionization, such as electrospray ionization (ESI).[27] ESI is often combined with a collision cell containing an inert gas to induce fragmentation, thereby

generating a characteristic fragmentation pattern. This technique, called collision-induced dissociation (CID), is applicable to a broad range of different compounds and is nowadays even more commonly used than EI-MS.[28] However, due to the complexity of the possible fragmentation pathways, structure elucidation from MS data remains highly challenging. In practice, compound identification is typically performed by comparing experimentally measured spectra to those in reference libraries.[29] While effective, this approach is limited by the incompleteness of existing spectral databases, which naturally cannot cover unknown or novel compounds. This limitation is particularly relevant in metabolomics, where the majority of detected compounds remain unidentified.[30,31] To overcome these limitations, computer-assisted structure elucidation (CASE) for mass spectrometry is of great interest. Such approaches aim to computationally predict fragmentation patterns from candidate structures – or even generate possible structures from spectra – thereby facilitating the identification of unknowns. Attempts to develop CASE workflows for mass spectrometry date back to the 1960s, beginning with the DENDRAL project, which pioneered the use of algorithms to determine chemical structures from EI-MS.[32,33] This early system was knowledge-based, relying on heuristic rules derived from expert knowledge, and already demonstrated performance comparable to that of trained chemists.[32] This so-called *de novo* spectrum-to-structure generation has significantly advanced in recent years, largely due to rapid progress in machine learning (ML) techniques. However, even state-of-the-art approaches still struggle with the inherent complexity of the task. For instance, the recently published DiffMS model,[34] evaluated on the large-scale MassSpecGym dataset,[35] achieved a correct structure identification rate of only 2.3 %. This illustrates that *de novo* structure elucidation from mass spectra remains an unsolved and highly challenging problem even for highly sophisticated ML models trained on large databases. In this context, quantum chemical (QC) approaches offer a more general and physically grounded alternative that do not rely on training data. Furthermore, in contrast to ML models, physics-based simulations can provide detailed mechanistic insight into the fragmentation process itself. A schematic depiction of a potential CASE workflow for MS is shown in Figure 1.2. The basic principle is to enhance experimental libraries with theoretically predicted spectra of compounds, that are not contained in the library. While ML models are used to propose candidate structures, QC approaches are employed to reliably predict spectra for structures that are not contained in spectral libraries. To this end, the QCEIMS program was developed by Grimme in 2013 for the simulation of EI-MS spectra.[36] This was later extended to CID-MS in the QCxMS program.[37] At the beginning of this thesis, QCxMS was the only fully automated program available for the simulation of both EI and CID mass spectra. More recently, another MD-based approach called CIDMD for CID-MS simulation has been introduced.[38] However, these approaches are fundamentally limited because MD simulations are computationally demanding, restricting the level of theory for typical system sizes to SQM methods and the feasible simulation time, which both affect the quality of spectra prediction. To overcome these limitations, a completely new methodology is developed in this thesis based on automated reaction network discovery. Although the theoretical foundations of this approach, namely the Rice-Ramsperger-Kassel-Marcus (RRKM)[39–41] and quasi-equilibrium theory (QET),[42] have been known for decades and successfully applied in numerous computational studies,[43] a fully automated program for the routine prediction of EI mass spectra is still lacking. Therefore, one goal of this thesis is to develop a fully automated computational workflow that takes a single input molecule and outputs a simulated mass spectrum, while employing various suitable QC methods (*vide supra*, Figure 1.1) and algorithms from modern computational chemistry. This workflow is developed as an open-source software package, available on GitHub under the name QCxMS2.[44]

Addressing both of the topics outlined above, NCIs and MS, this thesis is structured as follows.

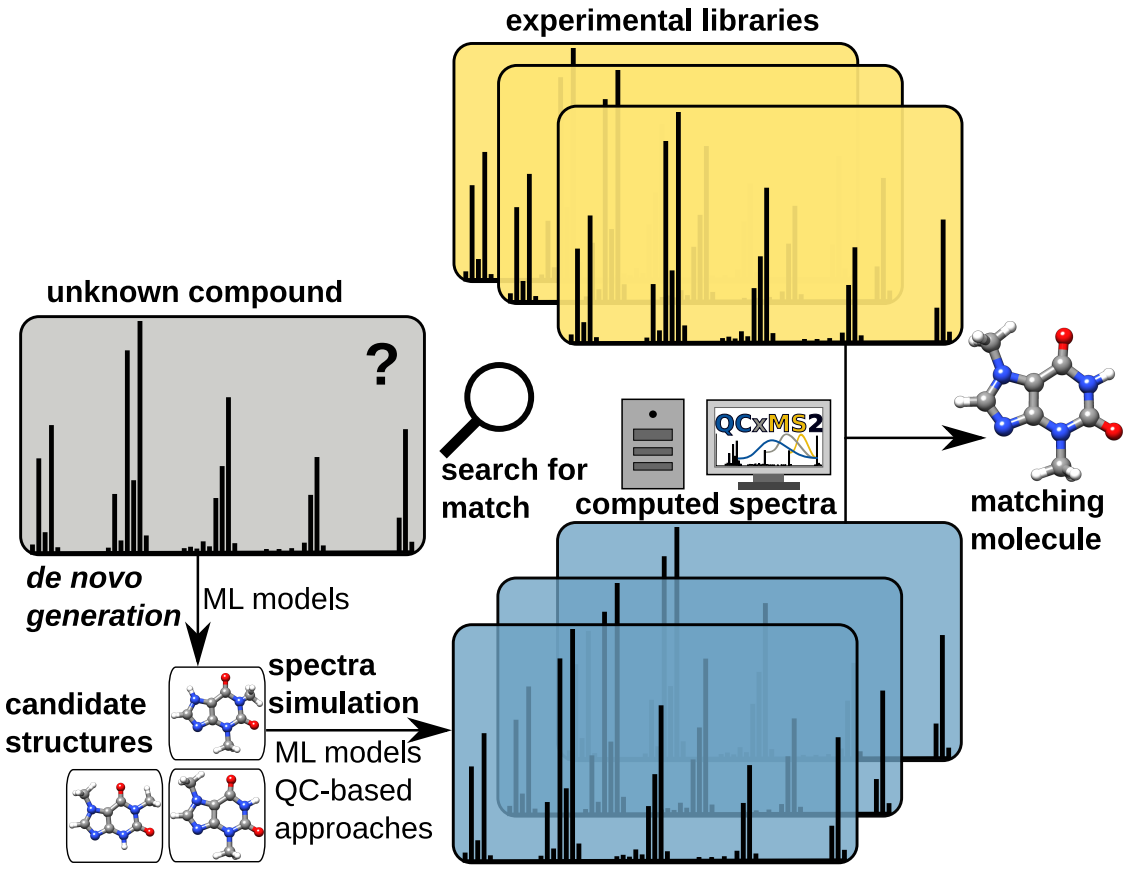


Figure 1.2: Schematic illustration of compound identification via matching of a measured mass spectrum with experimental and theoretical reference spectra.

First, in Chapter 2, an overview of the theoretical basis of the QC methods and the theoretical models for the calculation of free energies and reaction rates is given. To evaluate suitable QC methods for workflows, benchmark studies are performed to assess their accuracy and computational performance. This is performed for gas-phase interaction energies of very large supramolecular complexes in a theoretical comparison in Chapter 3. In the following Chapter 4, the comparison is extended by also considering thermal and solvation effects for the calculation of association Gibbs free energies of supramolecular complexes with heavy main group elements. In this benchmark study, the workflow employed is directly compared for realistic systems to experimental values to assess its real-world application. In the second half of this thesis, the development and assessment of QCxMS2 for the calculation of mass spectra is described. First, Chapter 5 covers the development of the QCxMS2 program and its performance for EI-MS. The following Chapter 6 outlines the extension of QCxMS2 for the calculation of CID-MS and evaluates its performance for this type of spectra. Finally, the findings, particularly with respect to the fields of NCIs and MS, as well as the impact of the newly developed program are summarized in Chapter 7.

CHAPTER 2

Theoretical Background

This chapter presents an overview of the theoretical foundation for the calculation of chemical properties in this thesis. In Section 2.1, the quantum mechanical concepts for the computation of the electronic energy are described. To this end, the electronic Hamiltonian is introduced (Section 2.1.1), the Hartree-Fock Theory (Section 2.1.2) and Kohn-Sham Density Functional Theory (Section 2.1.3) and their dispersion corrections (Section 2.1.4) are outlined and how they are solved with basis sets (Section 2.1.5), as well as the extended tight-binding methods (Section 2.1.6).

Furthermore, approaches to compute free energies using the Rigid-Rotor-Harmonic-Oscillator approximation (Section 2.2.1) and implicit solvation models (Section 2.2.2) are presented. For the calculation of MS, the modeling of experimental techniques such as EI and ESI, including CID-MS, is described in Section 2.3. This also includes the important theoretical concepts for computing reaction rates in MS, namely RRKM/QET (Section 2.3.2) and conventional transition state theory (Section 2.3.3).

2.1 Electronic Structure Methods

The target quantity of the electronic structure methods described here is the total electronic energy, which corresponds to the energy released when assembling a molecule from nuclei and electrons initially separated by infinite distances. Throughout this section, atomic units are used in all equations.[45]

2.1.1 The Electronic Hamiltonian

The foundation of most electronic structure methods is the time-independent non-relativistic Schrödinger equation[46]

$$\hat{H}\Psi = E\Psi, \tag{2.1}$$

that yields the total electronic energy E of a system with the wave function Ψ. The Hamiltonian operator $\hat{H}$ consists of kinetic and potential energy terms. For an atom or molecule with N electrons

and M nuclei, it is given by

$$\hat{H} = \underbrace{-\sum_{i=1}^{N} \frac{1}{2}\nabla_i^2}_{\hat{T}_\mathrm{e}} \underbrace{-\sum_{A=1}^{M} \frac{1}{2M_A}\nabla_A^2}_{\hat{T}_\mathrm{n}} \underbrace{-\sum_{i=1}^{N}\sum_{A=1}^{M} \frac{Z_A}{r_{iA}}}_{\hat{V}_\mathrm{ne}} + \underbrace{\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{1}{r_{ij}}}_{\hat{V}_\mathrm{ee}} + \underbrace{\sum_{A=1}^{M}\sum_{B>A}^{M} \frac{Z_A Z_B}{R_{AB}}}_{\hat{V}_\mathrm{nn}}. \tag{2.2}$$

Here, $\hat{T}_n$ and $\hat{T}_e$ denote the kinetic energy operators for nuclei and electrons, respectively. The potential energy consists of Coulomb terms: the nucleus-nucleus repulsion $\hat{V}_{nn}$, electron-electron repulsion $\hat{V}_{ee}$, and the electron-nucleus attraction $\hat{V}_{ne}$. The distance between the ith and jth electron is r_{ij}, r_{iA} denotes the distance between the ith electron and the Ath nucleus, and R_{AB} is the distance between nuclei A and B.[45] Because nuclei are much heavier and thus move more slowly than electrons, the nuclear and electronic motion can be separated following the Born-Oppenheimer approximation,[47] which yields the electronic Hamiltonian

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

where $\hat{V}_{nn}$ is added as a constant to the energy. $\hat{T}_n$ is computed separately via thermostatistical mechanics as described in Section 2.2. Since the error introduced by this approximation is typically negligible, all calculations in this thesis employ the Born-Oppenheimer approximation. Since the Schrödinger equation can only be solved exactly for one-electron systems, approximations to obtain solutions for many-electron systems are employed, which are introduced in the following.

2.1.2 Hartree-Fock Theory

In Hartree-Fock (HF) theory, the many-electron wave function is constructed from one-electron spin orbitals using a Slater determinant (SD), formed from N orthonormal molecular orbitals (MOs) $\phi_i(k)$ occupied by the k^{th} electron, as given by

$$\Psi^{\mathrm{SD}}(1, 2, \cdots, N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(1) & \phi_2(1) & \cdots & \phi_N(1) \\ \phi_1(2) & \phi_2(2) & \cdots & \phi_N(2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(N) & \phi_2(N) & \cdots & \phi_N(N) \end{vmatrix}. \tag{2.4}$$

This construction ensures that the SD satisfies two fundamental quantum mechanical requirements: the Pauli exclusion principle and the indistinguishability of electrons. Assuming optimal molecular orbitals that minimize the energy, the HF energy is obtained as the expectation value of the electronic Hamiltonian applied to the SD, expressed in Dirac notation[48] as

$$\begin{aligned} E^{\mathrm{HF}} &= \langle \Psi^{\mathrm{SD}} | \hat{H}_e | \Psi^{\mathrm{SD}} \rangle \\ &= \sum_{i=1}^{N} \underbrace{\langle \phi_i | -\frac{1}{2}\nabla_i^2 - \sum_{A=1}^{M} \frac{Z_A}{r_{iA}} | \phi_i \rangle}_{h_i} + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} (\underbrace{\langle \phi_i \phi_j | \frac{1}{r_{ij}} | \phi_i \phi_j \rangle}_{J_{ij}} - \underbrace{\langle \phi_i \phi_j | \frac{1}{r_{ij}} | \phi_j \phi_i \rangle}_{K_{ij}}), \end{aligned} \tag{2.5}$$

where the terms are grouped into the one-electron energies h_i, and the two-electron Coulomb J_{ij} and exchange K_{ij} energies. According to the variational principle, the set of MOs that minimizes the energy is determined under the constraint that the MOs remain orthogonal and normalized. This is achieved by introducing Lagrange multipliers λ_{ij}, leading to the HF equations

$$\underbrace{((h_i + \sum_{j=1}^{N}(J_j - K_j))}_{F_i}\phi_i = \sum_{j=1}^{N}\lambda_{ij}\phi_i \quad . \tag{2.6}$$

Here, F_i is the Fock operator, which is effectively a one-electron energy operator that yields the energy of one electron in the mean field of all other electrons.[20] Thus, explicit electron-electron correlation arising from Coulomb interactions between opposite-spin electrons is not captured. However, correlation between same-spin electrons – also referred to as Pauli or Fermi correlation – is included through the exchange integrals.

Many different sophisticated WFT methods build on HF and recover the missing electron (Coulomb) correlation energy by including additional determinants obtained through excitations of electrons from occupied to virtual orbitals. Notable examples include Møller-Plesset perturbation theory (MP2),[49] and coupled cluster theory with singles, doubles, and perturbative triples (CCSD(T)),[50] which is often used as reference method for benchmark studies.

2.1.3 Kohn-Sham Density Functional Theory

A different route to obtaining the total electronic energy of a system, circumventing the determination of the electron wave function, was introduced by Hohenberg and Kohn. They proved the existence of a one-to-one mapping between the electron density ρ and the ground-state energy of a system.[51] In pure or so-called orbital-free DFT, this is expressed directly through a functional of the density. However, the exact form of the electronic kinetic energy functional is unknown and most likely too complex to be solved for practical applications.[52] To overcome this, Kohn and Sham replaced the kinetic energy functional with the known expression for the kinetic energy from HF theory, computed for an assumed reference system of non-interacting electrons.[53] Thus, the ground-state energy in Kohn-Sham DFT (KS-DFT) is given by

$$E^{\text{KS-DFT}}[\rho] = T_{\text{S}}^{\text{KS}}[\rho] + V_{ne}[\rho] + J[\rho] + E_{XC}[\rho], \tag{2.7}$$

where $T_{\text{S}}[\rho]$ denotes the kinetic energy computed from a SD of the non-interacting reference system, given by

$$T_s^{KS}[\rho] = \sum_{i}^{N_{\text{MO}}} n_i \left\langle \phi_i \right| -\frac{1}{2}\nabla^2 \left| \phi_i \right\rangle , \tag{2.8}$$

with n_i as the occupation numbers of the one-electron KS orbitals ϕ_i. The electron density of this reference system is calculated via

$$\rho(\mathbf{r}) = \sum_{i=1}^{N_{\text{MO}}} n_i |\phi_i(\mathbf{r})|^2. \tag{2.9}$$

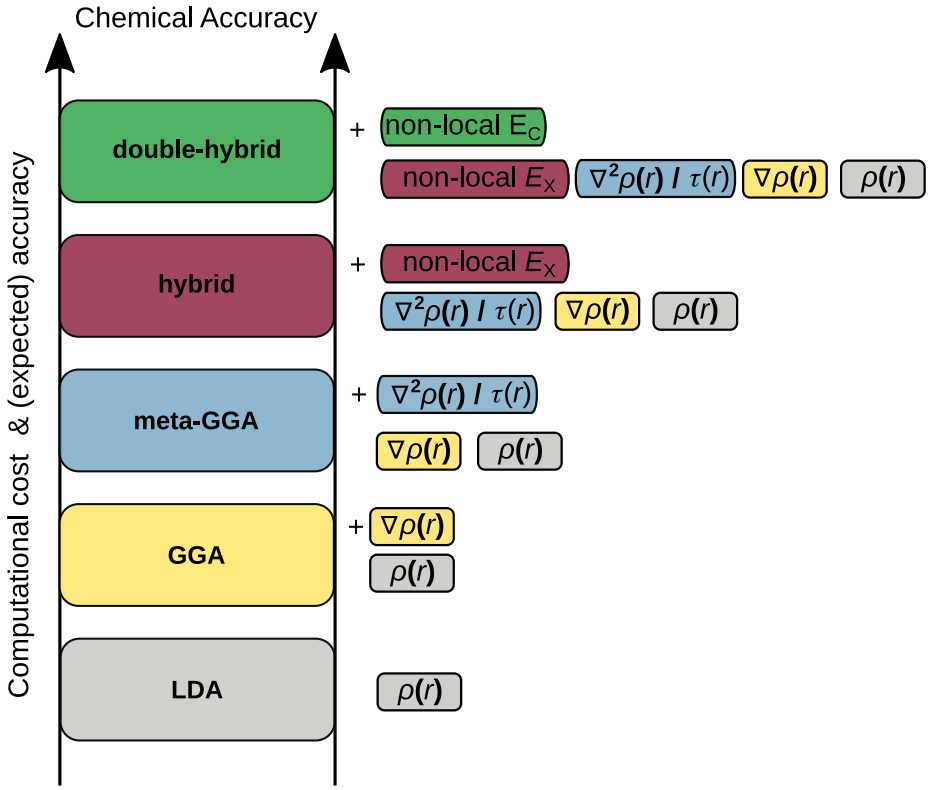


Figure 2.1: Allegorical depiction, in the form of Jacob's ladder, of different classes of KS-DFT methods and the information they incorporate.

The other terms in Eq. 2.7 represent the potential energy between nuclei and electrons $V_{ne}[\rho]$, the classical electron-electron repulsion $J[\rho]$, and the exchange-correlation (XC) functional $E_{XC}[\rho]$, which also accounts for the difference in kinetic electronic energy between the non-interacting reference system and the real, interacting system. For E_{XC}, the exact functional form is unknown, and thus many different density functional approximations (DFAs) have been developed over the years, incorporating various types of information to improve accuracy.

These DFAs can be categorized according to the famous allegory of Jacob's ladder, introduced by Perdew,[54] which is illustrated schematically in Figure 2.1. Different classes of DFAs are arranged according to the type of information they utilize. Higher rungs incorporate more information, leading to greater computational complexity and generally higher accuracy, although this can be system-dependent.

On the lowest rung are the local-density approximations (LDAs), which depend only on the electron density $\rho(\mathbf{r})$ and are primarily valid for metallic systems. Next are the generalized gradient approximation (GGA) functionals, which also include the gradient of the density $\nabla\rho(\mathbf{r})$. A well-known example is the PBE functional.[55]

By additionally including the second derivative of $\rho(\mathbf{r})$ or the orbital kinetic energy density τ, meta-GGAs are obtained. The r^{2}SCAN functional[56] is a notable member of this class.

A notorious deficiency of DFT is the self-interaction or electron-delocalization error, stemming from the fact that electrons incorrectly interact with themselves in the DFT formalism. To mitigate this,

hybrid DFAs replace a fraction of the exchange energy with non-local Fock exchange from HF theory, which is, by construction, self-interaction free. A prominent example is the B3LYP functional.[57–60] Because the optimal amount of exchange varies with interelectronic distance, range-separated hybrids (RSH) such as the ωB97X functional[61] were introduced.

On the highest rung are the double-hybrid DFAs, which also incorporate non-local correlation energy obtained via inclusion of virtual orbitals, for example through the MP2 correlation energy. B2PLYP[62] is one of the most prominent and widely used representatives of this class.

2.1.4 Dispersion-Corrected Density Functional Theory

An important deficiency of KS-DFT is its inability to describe long-range electron correlation effects, in particular, London dispersion (LD). LD is an attractive energy arising from instantaneous fluctuations in the electron density that induce dipole-dipole and higher-order multipole interactions, requiring excitations to virtual orbitals.[63] This missing contribution can be effectively addressed by applying dispersion correction schemes, such as the DFT-D approaches with the semi-classical D3[64,65] or D4[66,67] corrections, or by using non-local correlation functionals like VV10.[68] In these approaches, the dispersion energy is simply added to the total electronic energy of the system, as given by

$$E_{\text{total}} = E^{\text{KS-DFT}} + E_{\text{disp}}. \tag{2.10}$$

Currently, the most widely used scheme is the D3(BJ) correction. It consists of pairwise dipole-dipole ($E^{(6)}_{\text{disp}}$) and dipole-quadrupole ($E^{(8)}_{\text{disp}}$) interaction terms, given by

$$E^{D3}_{\text{disp}} = -\sum_{\text{AB}} \sum_{\text{n=6,8}} s_n \frac{C^{\text{AB}}_n}{R^n_{\text{AB}}} f^{(\text{n})}_{\text{damp}}(R), \tag{2.11}$$

where R_{AB} denotes the interatomic distance between atoms A and B. The pairwise dispersion coefficients C^{AB}_n are computed on-the-fly during a D3(BJ) calculation for a given molecular geometry, based on precomputed values for each atom pair and coordination number determined for reference systems.[69] Since DFT functionals already include short-range correlation to some extent, depending on the specific DFA, a damping function is introduced to smoothly interpolate in the short-range regime. This function is individually fitted to each DFA.[65] The Becke-Johnson (BJ) damping function[70] is used as the default in D3(BJ):

$$f^{(\text{n})}_{\text{damp,BJ}} = \frac{R^n}{R^n + (a_1 R_0 + a_2)^n}. \tag{2.12}$$

It contains the functional-specific parameters a_1 and a_2, and a cutoff radius $R_0 = \sqrt{\frac{C^{\text{AB}}_8}{C^{\text{AB}}_6}}$. Beyond pairwise interactions, many-body effects also contribute to the dispersion energy. The most important of these is the three-body term, known as the Axilrod-Teller-Muto (ATM) term,[71,72] which is computed as

$$E^{\text{ATM}} = \frac{C^{\text{ABC}}_9 \left(3 \cos\theta_a \cos\theta_b \cos\theta_c + 1\right)}{\left(R_{\text{AB}} R_{\text{BC}} R_{\text{CA}}\right)^3} f^{(9)}_{\text{damp}}(R_{\text{ABC}}), \tag{2.13}$$

where θ_a, θ_b, and θ_c are the internal angles of an atom triplet. In the successor D4 scheme, the charge dependence of the atomic polarizabilities is additionally taken into account and the ATM term is included by default.[67]

2.1.5 Basis Set Approximation

The molecular orbitals (MOs) ϕ_i used in the Slater determinant (SD) are usually obtained using the linear combination of atomic orbitals (LCAO) approach

$$\phi_i = \sum_{\alpha}^{M} c_{\alpha i} \chi_\alpha. \tag{2.14}$$

Herein, M basis functions χ_α, often referred to as atomic orbitals and centered on the nuclei, are combined linearly via the expansion coefficients $c_{\alpha i}$.

This formulation allows the HF equations to be reformulated in matrix notation as the Roothaan-Hall equations:

$$\mathbf{FC} = \mathbf{SC}\epsilon, \tag{2.15}$$

where $\mathbf{F}$ is the Fock matrix, representing the Fock operator applied to the basis functions, $\mathbf{S}$ is the overlap matrix of the basis functions, and $\mathbf{C}$ contains the molecular orbital coefficients. The eigenvalues ϵ correspond to the orbital energies. The coefficients in $\mathbf{C}$ are optimized variationally in an iterative process known as the self-consistent field (SCF) procedure which continues until the total energy converges to a stationary value, typically corresponding to a local minimum.[45] Different types of functions can be employed as basis functions χ_α. In molecular quantum chemistry, Gaussian-type orbitals (GTOs) are the most widely used due to their favorable computational properties, particularly in evaluating integrals of products of Gaussians. Although GTOs do not exactly reproduce the correct cusp behavior and long-range asymptotics of atomic orbitals, they can be linearly combined to approximate Slater-type orbitals, which more accurately reflect the true shape of the electron density.[20]

The accuracy of energies computed by QC methods is strongly dependent on the quality of the basis set, which must offer sufficient flexibility to accurately represent the shape of the MOs. This flexibility can be achieved by adding more basis functions. However, this also increases the computational cost. Especially WFT methods need large basis sets, and extrapolation schemes to the complete basis set (CBS) limit are commonly employed.[73]

A more recent approach involves the use of charge-dependent linear coefficients for the GTOs, as implemented in the q-vSZP basis set. This adaptive scheme enables a more realistic representation of the electron density while preserving the computational efficiency of a minimal basis set.[74]

2.1.6 Extended Tight-Binding Methods

Due to the still considerable computational cost of DFT, SQM methods have re-emerged as an alternative approach to enable the calculation of larger systems within a reasonable amount of computational time. These methods introduce additional approximations, which are compensated by empirical parameters fitted to reference data.

A very popular class of SQM methods are density functional tight-binding (DFTB) approaches, in which the XC-energy is expressed as a Taylor expansion around a known reference electron density ρ_0, which corresponds to the superposition of atomic densities of the molecule. The molecular energy

is computed by considering charge fluctuations $\delta\rho$ with respect to the reference density to construct the molecular density, usually up to third order, according to

$$E_{\mathrm{XC}}[\rho] = E^{(0)}[\rho_0] + E^{(1)}[\rho_0, \delta\rho] + E^{(2)}[\rho_0, (\delta\rho)^2] + E^{(3)}[\rho_0, (\delta\rho)^3] + \cdots . \tag{2.16}$$

A prominent example of this class are the GFNn-xTB methods, developed with the goal of accurately describing geometries, vibrational frequencies, and non-covalent interactions.[75] The currently most employed GFN2-xTB method[76] comprises the following energy contributions:

$$E^{\text{GFN2-xTB}} = E^{\mathrm{rep}} + E^{\mathrm{EHT}} + E^{\mathrm{IES+IXC}} + E^{\mathrm{AES}} + E^{\mathrm{AXC}} + E^{\mathrm{D4}}_{\mathrm{disp}} + G_{\mathrm{Fermi}}. \tag{2.17}$$

Here, E^{rep} is a semi-classical pairwise Coulomb repulsion energy and E^{EHT} is an extended Hückel-type Hamiltonian that allows for covalent bond formation. The isotropic electrostatic/XC correction, $E_{\mathrm{IES+IXC}}$, originates from the second-order term in the tight-binding expansion, as well as from third-order terms, of which only onsite isotropic XC contributions are computed. GFN2-xTB also includes an anisotropic electrostatic term, E^{AES}, based on a multipole expansion up to monopole-quadrupole and dipole-dipole interactions, as well as an anisotropic exchange-correlation (XC) term, E^{AXC}. LD interactions are modeled using a modified D4 scheme,[67] which employs self-consistent atomic partial charges derived from the GFN2-xTB charge calculation. The basis set employed is a partially polarized, minimal valence basis of Gaussian-type orbitals (GTOs).[21,76] Furthermore, an electronic entropy contribution G_{Fermi} due to Fermi smearing at finite electronic temperature[77] is contained to allow for SCF solutions with fractional occupations. GFN2-xTB can be considered an approximation to a generalized gradient approximation (GGA) functional. As such, it inherits the common GGA limitation of self-interaction error (SIE), as discussed in Section 2.1.3, typically resulting in underestimated reaction barriers and orbital gaps.

A very recent tight-binding method addressing this issue is the g-xTB method, which approximates a range-separated hybrid (RSH) functional and was parametrized against ωB97M-V. The "g" stands for "general", emphasizing its broader applicability compared to the GFN methods, in particular for the computation of thermochemistry. Its energy expression is given by

$$\begin{aligned} E^{\text{g-xTB}} = {} & E^{\mathrm{incr}} + E^{\mathrm{rep}} + E^{\mathrm{EHT}} + E^{\mathrm{ACP}} + E^{(1)} + E^{(2)} + E^{\mathrm{AES}} + E^{\mathrm{spin}} \\ & + E^{(3)} + E^{(4)} + E^{\mathrm{lr,MFX}} + E^{\mathrm{OFX}} + E^{\mathrm{revD4}}_{\mathrm{disp}}. \end{aligned} \tag{2.18}$$

It includes several notable differences to its predecessor GFN2-xTB to improve in accuracy, which are in the following described. E^{incr} is an atom-wise energy increment term that shifts the total energy to match the total energy of the reference method. The repulsion energy E^{rep} contains, in addition to Coulomb repulsion, a Pauli-type exchange repulsion. The extended Hückel-type term E^{EHT} uses a diatomic frame scaled overlap matrix that enables separate scaling of σ-, π-, and δ-bonding contributions.[78] Atomic correction potentials (ACPs)[79] are employed in E^{ACP}, which compensate for missing polarization functions in the minimal basis set. A first-order term in the density fluctuations, $E^{(1)}$, is also explicitly included. While the anisotropic XC term from GFN2-xTB is omitted, a multipole expansion up to fourth order is introduced for a more accurate description of anisotropic electrostatics. Furthermore, a spin-polarization term E^{spin} is contained, following the approach used for the spGFNn-xTB methods.[80] The third-order term $E^{(3)}$ has been extended beyond purely onsite contributions to include offsite, pairwise XC interactions. A key new component is

the inclusion of long-range ($E^{\mathrm{lr,MFX}}$) and onsite (E^{OFX}) Fock exchange terms, computed using the Mulliken approximation,[81] which enables accurate computation of reaction barriers and large orbital gaps. To address instabilities in highly charged anionic systems, mainly due to the $E^{(3)}$ term, a new fourth-order onsite term $E^{(4)}$ is introduced. Finally, g-xTB makes use of the adaptive q-vSZP basis set described in the previous section.

It should be noted that the g-xTB method is still under development at the time of writing.[78] In particular, its parametrization may continue to evolve, and therefore minor differences in computed results compared to those shown in Section 5 and Section 6 may occur, which reflect the development status at the time of those publications.

2.2 Contributions to the Gibbs Free Energy

For the description of realistic systems, for example, to predict binding constants or reaction rates, the Gibbs free energy G has to be computed, given by

$$G = E + G^{\mathrm{T}} + \delta G^{\mathrm{T}}_{\mathrm{solv}}(X). \tag{2.19}$$

Besides the electronic energy E, it includes also thermal effects at a given temperature T summarized in the term G^{T}, and, if applicable, solvation effects $\delta G^{\mathrm{T}}_{\mathrm{solv}}(X)$, which is the solvation free energy released by solvating the molecule in solvent X. Reaction free energies, e.g., for the formation of a supramolecular complex, are then calculated as the difference between the free energy of the products and the reactants.

$$\Delta G = \sum_{i=1}^{N_{\mathrm{products}}} G_i - \sum_{j=1}^{N_{\mathrm{reactands}}} G_j \tag{2.20}$$

2.2.1 The Rigid-Rotor-Harmonic-Oscillator Approximation

The sum of thermal contributions G^{T} in the gas phase, as given by the Gibbs equation

$$G^{\mathrm{T}} = H^{\mathrm{T}} - T \cdot S, \tag{2.21}$$

contains the thermal contributions to the enthalpy H^{T} and the entropy S at a given temperature T. Both quantities stem from the translational, rotational, and vibrational degrees of freedom of a molecule and can be computed separately using statistical thermodynamics and assuming an ideal gas state, i.e., isolated non-interacting molecules. Although this assumption is not exact, the approximations made here are often applied and lead to good results in most practical applications of computational chemistry. For molecules with several thermally accessible electronic states, the electronic entropy term S_{elec} has to be considered. However, for most molecules, this is not necessary due to the typically large energy gap between the ground state and the first excited state.

Derived from the "particle in a box model", enthalpy and entropy stemming from the translational degrees of freedom are obtained as

$$H_{\mathrm{trans}} = \frac{5}{2} RT \tag{2.22}$$

and

$$S_{\text{trans}} = \frac{5}{2}R + R\ln\left(\frac{V}{N_{\text{A}}}\left(\frac{2\pi M k_{\text{B}}T}{h^2}\right)^{3/2}\right). \tag{2.23}$$

Here, R is the ideal gas constant, M is the molecular mass, k_{B} is the Boltzmann constant, h is Planck's constant, N_{A} is Avogadro's constant, and V is the ideal gas volume. The rotation of a molecule is described in the *rigid-rotor* approximation, yielding the terms

$$H_{\text{rot}} = \frac{3}{2}RT \tag{2.24}$$

and

$$S_{\text{rot}} = R\left(\frac{3}{2} + \ln\left(\frac{\sqrt{\pi}}{\sigma}\left(\frac{8\pi^2 k_{\text{B}}T}{h^2}\right)^{3/2}\sqrt{I_1 I_2 I_3}\right)\right), \tag{2.25}$$

where $I_{1,2,3}$ are the three moments of inertia of the molecule and σ the symmetry number, i.e., the number of symmetry operations that map the molecule onto an indistinguishable configuration. The vibrational energies are derived from the harmonic oscillator model, yielding

$$H_{\text{vib}} = R\sum_{i=1}^{3N-6(7)}\left(\frac{h\nu_i}{2k_{\text{B}}} + \frac{h\nu_i}{k_{\text{B}}}\frac{1}{e^{h\nu_i/k_{\text{B}}T} - 1}\right) \tag{2.26}$$

and

$$S_{\text{vib}} = R\sum_{i=1}^{3N-6(7)}\left(\frac{h\nu_i}{k_{\text{B}}T}\frac{1}{e^{h\nu_i/k_{\text{B}}T} - 1} - \ln\left(1 - e^{-h\nu_i/k_{\text{B}}T}\right)\right), \tag{2.27}$$

with the harmonic frequencies ν_i of the molecule. For transition states, the sum over vibrational frequencies includes only $3N - 7$ modes, as one mode corresponds to the reaction coordinate along which the reaction proceeds, and appears formally as an imaginary frequency in quantum chemical calculations.[20] Furthermore, for linear molecules, the vibrational terms contain one additional vibrational contribution and the rotational terms are

$$H_{\text{rot}}(\text{linear}) = RT \tag{2.28}$$

and

$$S_{\text{rot}}(\text{linear}) = R\left(1 + \ln\left(\frac{8\pi^2 I k_{\text{B}}T}{\sigma h^2}\right)\right). \tag{2.29}$$

A distinctive feature of the vibrational energy is that it remains nonzero even at 0 K, due to the so-called zero-point vibrational energy (E_{ZPVE}), which arises from the quantum mechanical uncertainty in the position and momentum of the oscillator.[82] As visible in Eq. 2.27, for low-lying frequencies, S_{vib} strives towards infinity. This leads to notorious inaccuracies in the calculation of the entropy contribution of these modes, which typically occur in supramolecular complexes. To solve this problem, Grimme proposed a special treatment of these modes as hindered rotations.[83] For any normal mode, the moment of inertia μ for a free-rotor with the same frequency is calculated with the following

equation:

$$\mu = \frac{h}{8\pi^2 \nu}. \tag{2.30}$$

With the average molecular moment of inertia I_{av} an effective moment of inertia is defined by

$$\mu' = \frac{\mu I_{\mathrm{av}}}{\mu + I_{\mathrm{av}}}. \tag{2.31}$$

This ensures that for low values of ν the resulting high values of μ are restricted to a reasonable value. The entropy of a low-frequency mode is then given by

$$S_{\mathrm{R}} = R \left[1/2 + ln \left\{ \left(\frac{8\pi^3 \mu' k_{\mathrm{B}} T}{h^2} \right)^{1/2} \right\} \right]. \tag{2.32}$$

To determine what is considered to be a low-lying mode and to achieve a smooth transition between "small" and conventional frequencies, both entropies are combined in the following equation

$$S = w(\nu) S_{\mathrm{V}} + [1 - w(\nu)]\, S_{\mathrm{R}}, \tag{2.33}$$

whereby the weighting factor w is given by the Head-Gordon damping function:[84]

$$w(\nu) = \frac{1}{1 + (\nu_0/\nu)^{\alpha}} \tag{2.34}$$

with $\alpha = 4$. Originally, a cutoff value of $\nu_0 = 100\,\mathrm{cm}^{-1}$ was chosen, replacing vibrational entropies of frequencies below this value by the corresponding free rotor entropy. For the association of large bimolecular systems with about 100 atoms, the difference of the computed association entropies between the conventional approach and the here described approach is typically 1-2 kcal mol^{-1}.[83] In later studies, a value of $\nu_0 = 50\,\mathrm{cm}^{-1}$ was found to give more accurate results.[85]

2.2.2 Solvation Free Energy Computation with Implicit Solvation Models

Solvation free energies can be computed using either an explicit or an implicit solvation model. Explicit models treat solvent molecules individually by positioning them around the solute. This can also be done automatically, for example, using the QCG program.[86] While such approaches are valuable for specific applications, such as spectroscopic simulations[87] or reaction mechanism studies where solvent-solute interactions are critical,[88] they are generally less accurate for bulk solvation free energies and come with significantly higher computational cost.[86]

In contrast, computationally efficient implicit solvation models are frequently employed. These models approximate the surrounding solvent environment as a continuous dielectric medium that exerts an external potential on the solute. Popular models in this category include the Generalized Born Surface Area (GBSA)[89,90] and the Solvation Model based on Density (SMD)[91]. In such models, the solvation free energy ΔG_{solv} is typically decomposed into polar and nonpolar contributions, along with a constant shift term:

$$\delta G_{\mathrm{solv}} = \delta G_{\mathrm{polar}} + \delta G_{\mathrm{npolar}} + \delta G_{\mathrm{shift}}\,. \tag{2.35}$$

The polar term δG_{polar} accounts for electrostatic interactions between solute and solvent, while δG_{npolar} is determined by the shape and surface area of the solute cavity. The constant shift δG_{shift} adjusts for the reference state of the solvation process.

2.3 Calculation of Mass Spectra

This section outlines the theoretical foundations and assumptions underlying the calculation of mass spectra, with a focus on the QCxMS2 program developed in this thesis. The primary objective of QCxMS2 is to accurately model the processes occurring within a mass spectrometer, enabling automated and robust predictions of mass spectra. This includes the quantitative computation of relative fragment abundances as well as the molecular survival yield. Achieving this requires the inclusion of all physically relevant processes, while deliberately omitting those that are less critical and would compromise automation or introduce prohibitive computational costs. The theoretical analysis begins with electron ionization mass spectrometry (EI-MS), which offers a clearly defined and standardized experimental framework. It is then extended to collision-induced dissociation mass spectrometry (CID-MS), a technique of greater practical relevance but also greater variability in experimental conditions.

2.3.1 Time Scales in Electron Ionization Mass Spectrometry

A schematic representation of the key processes and their associated time scales in EI-MS, as resolved by time-resolved spectroscopy, is presented in Figure 2.2. These processes span an extensive time range,

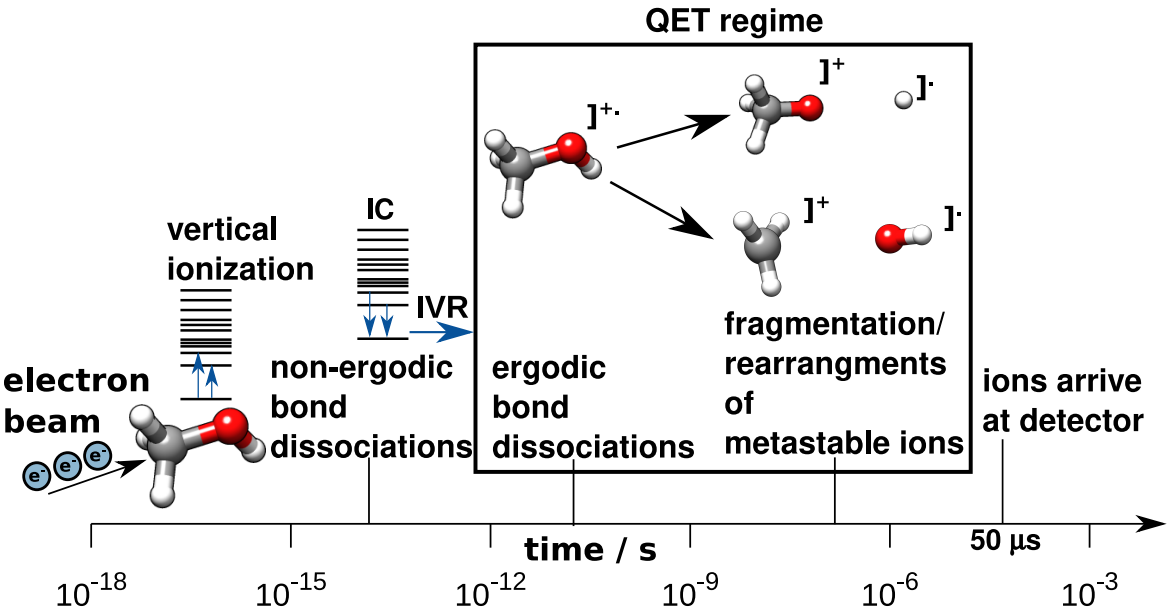


Figure 2.2: Overview of processes occurring in an EI mass spectrometer and their corresponding time scales.

covering up to 12 orders of magnitude – from attoseconds to microseconds. Ionization takes place on the attosecond time scale, as a 70 eV electron beam typically ejects one electron (though double or even triple ionization can also occur) from a neutral molecule. Following ionization, the molecule

relaxes from an electronically excited state to its ground state via internal conversion (IC), usually within 0.1 to 3 picoseconds. During IC, the electronic energy is redistributed into the vibrational modes of the molecule. Subsequently, intramolecular vibrational energy redistribution (IVR) occurs, spreading the energy equally throughout the molecule.[92] According to quasi-equilibrium theory (QET), the rate of IVR is typically several orders of magnitude faster than the rate of fragmentation or rearrangement reactions in a mass spectrometer.[42] Consequently, such reactions are assumed to proceed from thermally excited, yet quasi-equilibrated, ions. Reactions that occur faster than IVR are referred to as non-ergodic or non-statistical processes. Although these have been observed experimentally,[92] they are considered less relevant for typical applications discussed here. Modeling such ultrafast processes would require explicit treatment of electron dynamics, an approach that is currently impractical due to its high computational cost and the absence of a robust theoretical framework coupling many-electron dynamics with nuclear motion.[93] The relevant time scales that can be described within QET range from picoseconds (for fast dissociation) to nanoseconds and up to milliseconds, where fragmentation of metastable ions and rearrangement reactions occur. In EI-MS, a typical fragmentation time window is around 50 μs,[26] whereas in modern ion trap CID instruments, cycle times can range from milliseconds up to one second. These extended time scales render traditional molecular dynamics (MD) simulations impractical, as feasible simulation times on standard computational resources typically reach only into the picosecond range.[37] MD-based approaches such as QCxMS attempt to address this by simulating fragmentation processes at elevated effective internal energies compared to the experiment,[36] but even then, the simulated processes must span approximately six orders of magnitude in time. As a result, reaction network modeling based on rate constants represents a more viable strategy for predicting fragmentation pathways under these conditions.

2.3.2 Rice-Ramsperger-Kassel-Marcus Theory

Under the assumptions of QET, rate constants for unimolecular decompositions can be computed using Rice-Ramsperger-Kassel-Marcus (RRKM) theory,[39–41] depending on the internal energy E_{int} of an isolated ion:

$$k(E_{\text{int}}) = \frac{\sigma N^{\ddagger}(E_{\text{int}} - E_{\text{a}})}{h\rho(E_{\text{int}})}, \tag{2.36}$$

where σ is the reaction path degeneracy, h is Planck's constant, $N^{\ddagger}(E_{\text{int}} - E_{\text{a}})$ is the sum of states at the transition state, and $\rho(E_{\text{int}})$ is the density of states of the reactant ion.[41] The sum and density of states are often approximated by considering only vibrational states.[94]

Nevertheless, computing vibrational densities of states remains computationally demanding, as it requires second derivatives from accurate yet computationally expensive methods such as DFT to obtain reliable harmonic frequencies. Furthermore, the vibrational energy levels derived from these frequencies are highly sensitive to the molecular geometry, necessitating fully converged transition-state structures. However, automated transition-state search remains a significant challenge and may fail for certain reactions within a reaction network, potentially yielding unconverged or inaccurate transition-state geometries. In addition, barrierless reactions, i.e., dissociations lacking a clearly defined transition state, frequently occur, complicating the selection of an appropriate molecular geometry for computing the transition-state sum of states. Consequently, due to its complexity and sensitivity to input structures, the RRKM formalism is currently not suitable for integration into a

fully automated and robust workflow.[3,95]

2.3.3 Transition State Theory

An alternative approach to describing reaction rates is conventional transition state theory (TST). In this formalism, the reactant and transition state are assumed to be in thermodynamic equilibrium, leading to the following expression for the rate constant:

$$k(T) = \kappa \frac{k_B T}{h} \cdot e^{-\Delta G_a / k_B T}, \tag{2.37}$$

where k_B is the Boltzmann constant, ΔG_a is the Gibbs free energy of activation, and κ is the transmission coefficient,[96] which accounts for the probability that a system crossing the transition state will re-cross back to the reactant. Since we are primarily interested in relative rate constants, we assume $\kappa = 1$ for all reactions in the network.

To apply TST in the context of isolated molecules, the microcanonical internal energy is converted to an effective temperature using

$$T = \frac{E_{int}}{n_{vib} \cdot k_B}, \tag{2.38}$$

assuming thermal equilibrium among the vibrational modes. Although the molecules are isolated under low-pressure conditions in a mass spectrometer, the large number of accessible vibrational states justifies the use of this thermal approximation.[97] A comparison of rate constants calculated using RRKM theory and transition state theory (TST) reveals a very similar dependence of both approaches on internal energy E and the associated effective temperature T.[3]

The survival yield of a fragment, defined as the ratio of the final intensity I to the initial intensity I_0, follows the first-order rate law for unimolecular reactions:

$$\frac{I}{I_0} = e^{-k(E)t}, \tag{2.39}$$

where $t \approx 50\ \mu s$ is the typical flight time in the spectrometer.[26]

2.3.4 Internal Energy Distribution

The internal energy of the ions in the spectrometer is induced by the ionization process.[98] In standard positive mode EI-MS, an electron beam collides with the molecule, ejects an electron, and ionizes it.

$$M + e^- \rightarrow M^{\bullet +} + 2e^- \tag{2.40}$$

The ionization potential (IP) is herein the energy required to eject an electron e from Molecule M. Modeling this so-called (e,2e) process explicitly is highly complex and not feasible for routine applications. Instead, the internal energy transferred during ionization, referred to as impact excess energy (IEE), is modeled using an empirically determined energy distribution. QCxMS2 uses similar to QCxMS, a Poisson-type function, as given by

$$P(E) = \frac{\exp[cE(1 + \ln(b/cE)) - b]}{\sqrt{(cE + 1)}}, \tag{2.41}$$

where $P(E)$ represents the probability that the impact excess energy (IEE) is equal to E. The parameters a, b, and c are defined as approximately 0.2 eV, 1.0, and $1/(aN_{\mathrm{el}})$, respectively, where N_{el} is the number of electrons in the molecule. The maximum possible IEE is given by $E_{\mathrm{impact}} - \epsilon_{\mathrm{HOMO}}$, where E_{impact} is an input parameter and represents the kinetic energy of the free electron, before impact. The energy of the highest occupied molecular orbital (HOMO), denoted as ϵ_{HOMO}, is computed by a QM calculation (usually DFT). In standard EI experiments, E_{impact} amounts to 70 eV. To simulate these conditions, the distribution in QCxMS2 is set to an average of about 0.8 eV per atom of the input molecule.

2.3.5 Distribution of Charges

The internal energy leads to fragmentation of the molecular ion:

$$\mathrm{M}^{\bullet+} \rightarrow \mathrm{A}^{+} + \mathrm{B}^{\bullet}. \tag{2.42}$$

In this process, the localization of the positive charge among the resulting fragments must be determined, as it dictates which fragment ion will appear in the experimental mass spectrum. This is achieved according to the ionization potentials (IPs) of the fragments, using Boltzmann weighting based on the energy differences $\Delta E_{\mathrm{SCF},i}$ between the neutral and charged states of fragment i:

$$P_i = \frac{\exp\left(-\frac{\Delta E_{\mathrm{SCF},i}}{k_{\mathrm{B}} T_{\mathrm{Av}}}\right)}{\sum_{j=1}^{N} \exp\left(-\frac{\Delta E_{\mathrm{SCF},j}}{k_{\mathrm{B}} T_{\mathrm{Av}}}\right)}, \tag{2.43}$$

where P_i is the probability of fragment i retaining the charge, and N is the total number of fragments considered. The average ion temperature, denoted T_{Av}, is computed from the average internal energy using Eq. 2.38.

2.3.6 Computation of Collision-Induced Dissociation Mass Spectrometry

CID-MS differs from EI-MS in two major aspects. First, it employs a "soft" ionization technique, typically electrospray ionization (ESI), which leads to a different internal energy distribution.[98] This also introduces the need to account for different protomeric forms generated during the ESI process, which will be discussed in the following section. Second, fragmentation is induced via collisions with a neutral collision gas in a dedicated collision chamber. This collision-based mechanism further alters the internal energy distribution compared to EI-MS.

Fragmentation in CID is initiated by collisions of the ion with neutral gas atoms N

$$[\mathrm{M{+}H}]^{+} + \mathrm{N} \rightarrow [\mathrm{M{+}H}]^{+*} + \mathrm{N} \rightarrow \mathrm{A}^{+} + \mathrm{B}, \tag{2.44}$$

where $[\mathrm{M{+}H}]^{+*}$ denotes the collisionally activated ion. The internal energy after collision is given by:

$$E_{\mathrm{int}}([\mathrm{M{+}H}]^{+*}) = E_{\mathrm{int}}([\mathrm{M{+}H}]^{+}) + E_{\mathrm{coll}}. \tag{2.45}$$

The maximum value of E_{coll} is equal to the center-of-mass energy E_{com}, defined as

$$E_{\mathrm{com}} = \frac{m_{\mathrm{N}}}{m_{\mathrm{N}} + m_{\mathrm{P}}} E_{\mathrm{kin}}, \tag{2.46}$$

where E_{kin} is the kinetic energy of the precursor ion in the so-called laboratory frame, and m_{N} and m_{P} are the masses of the collision gas and precursor ion, respectively.[95] The fraction of E_{com} that is converted into internal energy is determined by the empirical collision inelasticity factor η:

$$E_{\mathrm{coll}} = \eta E_{\mathrm{com}}. \tag{2.47}$$

Based on experimental studies, η is typically around 0.5.[99] Using this value in QCxMS yields good agreement with experimental results.[37]

It is assumed that QET also holds in CID, i.e., the energy introduced by the collision redistributes rapidly via IVR before fragmentation takes place. This assumption is supported by MD-based QCxMS simulations, which have successfully reproduced key fragmentation pathways using an appropriately chosen internal energy distribution – without requiring explicit modeling of individual collision events.[37] Based on these findings, QCxMS2 also relies solely on modifying the internal energy distribution compared to EI-MS to simulate CID spectra.

As in QCxMS, a normally distributed internal energy model is used in QCxMS2 to approximate the activation energies arising from both the ESI process and subsequent collisions. This distribution is generated using the Box-Muller method,[100] resulting in a standard normal distribution:

$$P(E) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(E - E_{\mathrm{avg}})^2}{2\sigma^2}\right), \tag{2.48}$$

where the standard deviation is set to $\sigma = 0.2\,\mathrm{eV}$. The average internal energy, E_{avg}, is controlled by the *esiatom* parameter, which is typically set between 0.2 and 0.6 eV per atom, depending on the experimental setup.

2.3.7 Protonation Site Screening

During the ESI process, multiple protomers can form, exhibiting different fragmentation patterns.[101,102] These protomers do not necessarily follow a thermal distribution and higher-energy protomers can significantly contribute to the observed mass spectrum.[37,103] Therefore, in CID-MS, several protomers must be considered for each compound. To identify them in an automated manner, the protonation site screening procedure implemented in the CREST program[7,104,105] can be employed. This procedure uses localized molecular orbitals (LMOs), obtained from a GFN*n*-xTB[75] calculation, to identify lone pairs and π-centers as potential protonation sites. A proton is then placed at each of these sites, and the resulting structures are optimized and ranked by energy.[7,104,105] To capture all relevant protomers, an energy window of approximately 60 kcal mol^{-1} at the GFN2-xTB level (accounting for the method's estimated inaccuracy) and about 40 kcal mol^{-1} at the DFT level is typically used.[4]

2.3.8 Mass Spectral Fragmentation Tool

For the reaction network discovery approach of QCxMS2, it is crucial to generate all possible fragmentation pathways to get all relevant spectral peaks. To achieve this in a general and unbiased manner, no heuristic rules are applied. Instead, the fragment generator *MSREACT*, which is also implemented in *CREST*, is used.[7]

Herein, possible products are generated by applying constraining potentials between atom pairs, followed by optimization using the efficient SQM method GFN2-xTB. By default, atom pairs separated by up to three covalent bonds are subjected to elongation using a repulsive harmonic potential given by

$$V(r_{ij}) = \frac{1}{2} k_r (r_{ij} - r_0)^2, \tag{2.49}$$

where r_0 is set to 1.5 times the sum of the covalent radii of the two atoms plus the number of covalent bonds between them. The force constant k_r is set to $0.05\,\mathrm{E_h/Bohr^2}$. The number of bonds between two atoms is determined from the molecular graph (topology), using the Floyd-Warshall algorithm.[106] Subsequent geometry optimizations are performed with GFN2-xTB at an elevated electronic temperature of 5000 K. This promotes the formation of open-shell radicals and partially accounts for the multi-reference character of the open-shell species commonly encountered in (EI-)MS.[36] To enable the generation of frequently observed products resulting from hydrogen rearrangements, further optimizations are carried out with attractive harmonic potentials (k_r set to $-0.05\,\mathrm{E_h/Bohr^2}$) applied between hydrogen atoms and potential protonation sites. These sites are identified via LMOs obtained from GFN2-xTB, as described in the previous section, and are selected if a hydrogen atom lies within a default cutoff distance of 4 Å. This fragment generator showed good results within the QCxMS2 program, yielding in most cases all relevant fragments.[3,4]

CHAPTER 3

Efficient Computation of the Interaction Energies of Very Large Non-covalently Bound Complexes

Johannes Gorges,[†‡] Benedikt Bädorf,[†‡], Stefan Grimme,[†] Andreas Hansen[†] *Received: 21 July 2022*
Published online: 30 November 2022

Reprinted in Appendix A from Ref. J. Gorges, B. Bädorf, S. Grimme, and A. Hansen, *Efficient Computation of the Interaction Energies of Very Large Non-covalently Bound Complexes*, Synlett **34**.10 (2022) 1135, DOI: 10.1055/s-0042-1753141 with permission from Thieme.

Own contributions

- Conducting parts of the force-field and DFT calculations
- Interpretation of the results
- Co-writing and revising the manuscript

[†]Mulliken Center for Theoretical Chemistry, Universität Bonn, Beringstr. 4, D-53115 Bonn, Germany
[‡]These authors contributed equally.

The first work presented in this thesis investigates the accurate yet efficient quantum-chemical description of NCIs in very large supramolecular complexes with up to 2000 atoms. NCIs, including hydrogen bonding, π-π stacking, halogen bonding, and particularly LD, are central to supramolecular chemistry and play a key role in practically relevant applications such as drug delivery,[107,108] catalysis,[109,110] and the design of artificial molecular motors.[111] Modeling these interactions with quantum chemistry remains computationally demanding, especially for systems comprising several hundred to thousands of atoms.

To address this challenge, computationally efficient yet reasonably accurate methods such as SQM methods and even more approximate FF methods are often used. However, due to their high degree of empiricism, these methods must be carefully parametrized and benchmarked against more accurate reference techniques. For this purpose, a new benchmark set named LNCI16 is introduced, consisting of 16 supramolecular complexes with up to 1988 atoms. The benchmark covers diverse interaction motifs and molecular topologies, enabling a systematic assessment of different computational methods with respect to their performance in describing NCIs. To ensure a meaningful and consistent evaluation, a purely theoretical comparison of gas-phase interaction energies is performed. Interaction energies are computed in the supramolecular approach, i.e., as the energy difference between the complex and its unrelaxed fragments, without accounting for relaxation energies.

For the computation of reference energies, the range-separated composite hybrid DFA ωB97X-3c is employed. This method combines a molecularly optimized valence double-zeta basis set (vDZP), large-core effective core potentials, and the D4 dispersion correction. The accuracy of the method has been benchmarked on established NCI benchmarks, showing excellent agreement to CCSD(T) reference data.[112] Due to its computational efficiency and low memory demand, even the calculations of the largest systems in the LNCI16 are still computationally feasible. The performance of several classes of computational methods, introduced in Chapter 1, is evaluated: composite DFT methods, SQM methods, and FFs. Due to the broad range of interaction energies in the dataset, relative mean absolute deviations (relMAD) were employed as the primary metric for comparison. In this context, a relMAD below 5% is considered "chemically accurate". Furthermore, since the tested composite DFT methods are not necessarily less accurate than the reference ωB97X-3c, the benchmark is more relevant for assessing the SQM and FF methods. Among the composite DFT methods, r^{2}SCAN-3c[113] achieves the best agreement with a relMAD of 6.6%. However, convergence issues of this meta-GGA-based composite method arise for systems with small HOMO-LUMO gaps. PBEh-3c[114] performs only slightly worse (relMAD 8.5%) while converging for all systems. Among the SQM methods, GFN2-xTB[76] stands out with a relMAD of 11.1%, showing that it can describe large NCI complexes with high accuracy at low computational cost. In the force-field domain, GFN-FF[115] yields the most accurate results (relMAD 34.7%), followed closely by GAFF (37.0%). To assess the computational efficiency of the employed methods, theoretical timings for the largest system in the set, a nylon polymer complex comprising 1988 atoms, were normalized to a single CPU core. Composite DFT methods required several days to weeks of wall time, SQM methods completed within a few minutes, and force fields such as GAFF and GFN-FF finished in a matter of seconds. This demonstrates the trade-off between accuracy and efficiency and provides practical guidance on method selection depending on system size and the quantity of interest. In conclusion, the LNCI16 benchmark set serves as a widely applicable test set for computational methods targeting large supramolecular systems. The results of this study are relevant not only for method selection in the modeling of such very large supramolecular complexes but also for future method development, in particular in the context of NCIs and robustness for extended system sizes.

CHAPTER 4

Reliable Prediction of Association (Free) Energies of Supramolecular Complexes with Heavy Main Group Elements – the HS13L Benchmark Set

Johannes Gorges,[†] Stefan Grimme[†] Andreas Hansen[†]

Received: 31 August 2022
Published online: 18 November 2022

Reproduced in Appendix B from Ref. J. Gorges, S. Grimme, and A. Hansen, *Reliable prediction of association (free) energies of supramolecular complexes with heavy main group elements - the HS13L benchmark set*, Phys. Chem. Chem. Phys. **24**.47 (2022) 28831, DOI: 10.1039/d2cp04049b with permission from the Royal Society of Chemistry.

Own contributions

- Compiling the HS13L benchmark set
- Performing all density functional theory, semiempirical quantum mechanical, and force-field calculations
- Interpretation of the results
- Writing the manuscript

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

Whereas the previous work focused solely on theoretical gas-phase interaction energies of NCI complexes, this study extends the analysis to experimentally accessible association free energies in solution. This also includes the computation of geometry relaxation energy upon complex formation, as well as thermal and solvation effects, as described in Section 2.2. To address the scarcity of benchmark data for such systems, a new benchmark set named HS13L was compiled. It comprises 13 large supramolecular complexes with system sizes ranging from 37 to 266 atoms and includes heavy main group elements such as Zn, Se, Te, P, As, Br, and I. For all systems, experimentally measured association Gibbs free energies are available, ranging from -1.9 to $-9.2\,\mathrm{kcal\,mol^{-1}}$. These complexes are of significant interest in various areas of chemistry,[13] including drug delivery[116] and use as reaction containers.[117] However, their computational modeling remains challenging due to their large size and conformational flexibility. This necessitates highly accurate and efficient theoretical workflows capable of treating systems with several hundred atoms. In addition, the presence of heavy atoms and the complexity of their characteristic NCI motifs, including halogen, chalcogen, pnictogen, and tetrel bonding,[118] require rigorous benchmarking of quantum chemical methods to assess their accuracy.

Association free energies were computed using an automated, multilevel workflow combining the CREST[7,105] program for conformer generation and the CENSO[119] program for energetic refinement. Conformer ensembles were generated with GFN2-xTB[76] and reranked to find the lowest-energy structures using multiple sorting steps at different levels of theory in CENSO. Final energy rankings employed the composite DFT method r^{2}SCAN-3c[113], thermostatistical corrections were computed using the single-point Hessian approach[85] and the mRRHO model[83] at the GFN2-xTB level, and solvation free energies were obtained using the COSMO-RS implicit solvation model.[120,121] Charged complexes (systems **10**-**13**) were neutralized by adding either Na^+ or Cl^- counterions with the automated interaction site screening (aISS) algorithm[122] in combination with the xtb-IFF force field,[123] resulting in the HS13L-CI variant. Neutralization significantly improved agreement with experiment, as implicit solvation models tend to show larger deviations for ionic systems. The final protocol yields a mean absolute deviation (MAD) of $2.0\,\mathrm{kcal\,mol^{-1}}$ for HS13L-CI, supporting its reliability. Among the tested solvation models, COSMO-RS performed best and SMD[124] also gave reasonable results ($3.9\,\mathrm{kcal\,mol^{-1}}$ MAD), whereas CPCM[125] ($14\,\mathrm{kcal\,mol^{-1}}$ MAD) and ALPB[126] ($13.2\,\mathrm{kcal\,mol^{-1}}$ MAD) showed larger errors. For CPCM, this is likely due to the absence of non-electrostatic contributions, and for ALPB, the deviations may stem from limitations in the underlying GFN2-xTB method or solvation parameterization. The single-point Hessian approach proved robust for thermostatistical corrections, whereas relaxed frequency calculations with GFN2-xTB, GFN1-xTB,[127] PM6-D3H4X,[128] and GFN-FF[115] also gave satisfactory results, showing that DFT-level frequencies are not required for accurate predictions. To validate the gas-phase component independently, high-level DLPNO-CCSD(T1) reference values extrapolated to the complete basis set limit were used. Various DFAs with different dispersion corrections were benchmarked. Notably, r^{2}SCAN-3c showed a deviation of only $3.4\,\mathrm{kcal\,mol^{-1}}$, indicating that its agreement with experiment is not merely due to error cancellation. Surprisingly, hybrid and double-hybrid DFAs did not significantly outperform GGA or meta-GGA functionals, contrary to expectations based on Jacob's ladder (see Section 2.1.3).

In conclusion, the HS13L and HS13L-CI benchmark sets fill a critical gap in validating quantum chemical methods for large, supramolecular systems with heavy elements and charges. The workflow based on CREST, CENSO, GFN2-xTB, r^{2}SCAN-3c, and COSMO-RS provides a useful balance between accuracy and efficiency for reliably predicting association free energies in solution, even for such challenging systems as investigated here.

CHAPTER 5

QCxMS2 – a Program for the Calculation of Electron Ionization Mass Spectra via Automated Reaction Network Discovery

Johannes Gorges[†], Stefan Grimme[†]

Received: 23 January 2025
Published online: 03 March 2025

Reprinted in Appendix C from
J. Gorges and S. Grimme, *QCxMS2 - a program for the calculation of electron ionization mass spectra via automated reaction network discovery*, Phys. Chem. Chem. Phys. **27**.14 (2025) 6899, DOI: 10.1039/d5cp00316d with permission from the Royal Society of Chemistry.

Own contributions

- Development of the QCxMS2 software
- Performing calculations with QCxMS2 using various program parameters and quantum chemical methods
- Interpretation of the results
- Writing the manuscript

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

In this work, a new program was developed for the computation of EI-MS using QC calculations. Accurate EI-MS prediction is of great importance for analytical chemistry since interpretation of experimental spectra is challenging and often unsuccessful.[30] Currently, many data-driven ML approaches exist, but their performance depends on the training data and is uncertain for unknown compounds with potentially novel fragmentation pathways.[34] In contrast, QC-based methods offer a more general and reliable approach. The only fully automated program available for EI-MS computation is QCxMS,[36,129] which has shown good results for a broad range of compounds.[130] However, its reliance on MD simulations limits both the feasible simulation time and level of theory, consequently affecting spectral accuracy. To improve the accuracy, a new program, QCxMS2, was developed. It replaces MD simulations with a reaction network exploration approach. This enables the use of more accurate QC methods for the evaluation of reaction energetics and overcomes the time limitations inherent to MD approaches. Reaction rates of the fragmentation reactions are derived from the respective reaction barriers using TST,[96] as described in Section 2.3. Kinetic modeling of the resulting reaction network via Monte Carlo simulation over a molecular size-dependent internal energy distribution yields relative fragment intensities and thus a computed mass spectrum. QCxMS2 is an advanced script that employs a composite QC protocol using geometry and minimum energy reaction path optimizations as well as single-point calculations to compute reaction energies and barriers for the generated reaction network. The approach is fully automated and integrates external tools such as the MSREACT mode of CREST[7] for fragment generation, MolBar[131] for duplicate filtering, and ORCA[132] for reaction path searches and DFT calculations. The new program was tested on a benchmark set of 16 chemically diverse organic and inorganic molecules, including linear alkanes, alkenes, alcohols, ketones, esters, heterocycles, and compounds with the elements P, S, and Al. Spectral similarity between theoretical and experimental spectra was quantified using the entropy similarity score (ESS).[133,134] It ranges from 0 (no agreement) to 1 (perfect agreement), and an ESS above 0.75 can be regarded as sufficiently good agreement with the experiment for typical applications.[133] Using GFN2-xTB[76] for both geometries and energies, QCxMS2 achieves an average ESS of 0.670. Refining the reaction barriers at the RSH DFT level ωB97X-3c[112] increases the average ESS to 0.700, while using ωB97X-3c also for geometry optimization further improves it to 0.730. These values are significantly better than those obtained with the predecessor QCxMS, which yields an average ESS of 0.622 when using GFN2-xTB. QCxMS2 is also more robust, with a minimum ESS of 0.527 across the test set, compared to the minimum value of only 0.1 for QCxMS. Especially for molecules that are inaccurately described by SQM methods like GFN2-xTB, higher accuracy can be achieved by refining reaction barriers at the DFT level. Furthermore, fragments stemming from rearrangement reactions, which are often underestimated in MD-based simulations due to their short timescales, are captured reliably in QCxMS2. For only one molecule, the important main peak was missing, presumably because the MSREACT fragment generator relies on GFN2-xTB, which is too inaccurate for this particular case. For a typical metabolite-like molecule such as caffeine with 25 atoms, QCxMS2 takes 3.7 hours on 16 CPU cores using GFN2-xTB, and 15.7 hours if barriers are refined with ωB97X-3c. Although QCxMS is faster with a runtime of 1 hour for the same molecule at the GFN2-xTB level, QCxMS2 is more efficient than QCxMS when aiming for high accuracy, since DFT-level MD simulations are practically unfeasible for routine applications and refinement of barriers via single-point calculations is not possible.

In conclusion, QCxMS2 is a robust and promising program for the quantum chemical calculation of EI-MS spectra. The results obtained for a chemically diverse test set demonstrate that automated reaction network exploration is a viable and accurate alternative to MD-based simulations.

CHAPTER 6

Evaluation of the QCxMS2 Method for the Calculation of Collision-Induced-Dissociation Spectra via Automated Reaction Network Exploration

Johannes Gorges[†], Marianne Engeser[‡], Stefan Grimme[†]

Submitted to the Journal of the American Society of Mass Spectrometry: 10 July 2025 – Published online: 4 September 2025

Reprinted in Appendix D from
J. Gorges, M. Engeser, and S. Grimme, *Evaluation of the QCxMS2 method for the calculation of collision-induced-dissociation spectra via automated reaction network exploration*, ChemRxiv Prepr. (2025), DOI: 10.26434/chemrxiv-2025-gcws2 – License: CC BY-NC-ND 4.0.

Own contributions

- Implementation of the CID mode in the QCxMS2 software
- Performing calculations with QCxMS2 using various program parameters and quantum chemical methods
- Interpretation of the results
- Writing the manuscript

[†]Mulliken Center for Theoretical Chemistry, Universität Bonn, Beringstr. 4, D-53115 Bonn, Germany
[‡]Kekulé Institute for Organic Chemistry and Biochemistry, Gerhard-Domagk-Str. 1, 53121 Bonn, Germany

In the final work of this thesis, the QCxMS2 program, whose development for EI-MS was described in the previous chapter, is extended to enable the calculation of CID-MS. CID-MS is often combined with soft ionization techniques such as ESI and allows the analysis of a broader range of labile, nonvolatile compounds, including metabolites, peptides, and drug-like molecules,[27,135,136] and is therefore arguably of even greater practical relevance than EI-MS. Current generally applicable QC-based approaches, such as QCxMS[37] and CID-MD,[38] rely on MD simulations and lack either the accuracy or computational efficiency needed for routine use in automated structure elucidation workflows. In contrast to EI-MS, which operates under standardized conditions at a typical ionization energy of 70 eV, a wide variety of experimental setups exist for CID-MS that are often tailored to the compound being analyzed to produce information-rich spectra with characteristic fragmentation patterns. The internal energy of ions in CID-MS depends on parameters such as the applied voltage during ESI and the conditions in the collision chamber, including the type of collision gas, its pressure and temperature, and the specified collision energies.[95] Furthermore, ESI can generate multiple protomers of a single molecule, each of which may follow distinct fragmentation pathways.[137]

These challenges are addressed in QCxMS2 by approximating the internal energy with a normal distribution scaled by the number of atoms in the input molecule. This serves as a proxy for the ionization and collisional activation processes, enabling efficient simulation of fragmentation without explicitly modeling individual collisions. The approach relies on the assumption from QET that the most relevant fragmentations occur after the collision energy has equilibrated throughout the molecule. To evaluate the method, a test set of 13 organic compounds was compiled, including amino acids, amines, aromatic drugs, and a phosphorus-containing heterocycle. For each molecule, both CID at 20 eV and higher-energy collisional dissociation (HCD) spectra at 70 eV were measured under consistent experimental conditions to avoid the inconsistencies commonly found in spectral libraries, where spectra are recorded under differing experimental setups.

The theoretical predictions were compared to these reference spectra using the entropy similarity score (ESS). On average, QCxMS2 achieves at the composite level of ωB97X-3c barriers and GFN2-xTB geometries good ESS values of 0.687 for the HCD spectra and 0.773 for the CID spectra, demonstrating its capability to reliably reproduce key fragmentation patterns across a chemically diverse set. As in the previous work shown for EI-MS, QCxMS2 also achieves here significantly higher accuracy than QCxMS, yielding average ESS values of only 0.377 for HCD and 0.626 for CID. Different average internal energies were tested with QCxMS2, which can be done efficiently without rerunning expensive QC calculations, an advantage over MD-based approaches, which require complete recalculations. Herein, the optimal average internal energy per atom was found to be molecule-specific, although reasonably good values could be achieved by using 0.45 eV per atom for the HCD spectra. Furthermore, different protonation sites of the molecules were systematically investigated, resulting in a total of 41 protomers computed across the test set. In line with previous studies,[37,38,103,138] it was found that for many compounds (seven out of 13), protomers of up to $\approx 40\,\mathrm{kcal\,mol^{-1}}$ in relative free energy yield better agreement with the experiment compared to the lowest-energy one. For several molecules, QCxMS2 also predicts interconversion between protomers, validating the mobile proton theory.[139]

In conclusion, this work extends QCxMS2 to the domain of CID-MS and demonstrates its ability to reliably predict fragmentation spectra across a broad range of compounds while maintaining reasonable computational cost. With theoretical spectra now available for both EI- and CID-MS, this work establishes QCxMS2 as a powerful tool for supporting automated structure elucidation workflows.

CHAPTER 7

Summary and Outlook

Computational chemistry has become an increasingly powerful tool, driving advances in a wide range of research fields. Beyond the ongoing development of electronic structure and energy prediction methods – such as DFAs, SQM methods, FFs, and MLIPs – the design and implementation of automated computational workflows that integrate these methods are equally important. Such workflows make complex molecular property calculations routinely accessible, even to non-expert users, and enable scientific insights that would otherwise require prohibitively large amounts of manual effort or even remain entirely out of reach. This thesis focused on both the development of new computational workflows and their application to evaluate reliability and applicability across a wide range of chemical systems. Specifically, it addressed two key areas: The accurate computation of NCIs in supramolecular systems and the quantum chemical prediction of mass spectra.

The topic of NCIs in large molecular systems was investigated within the LNCI16 benchmark study, as described in Chapter 3. This benchmark includes various efficient computational methods for calculating gas-phase interaction energies in 16 large molecular complexes, comprising up to approximately 2000 atoms and covering a wide range of interaction motifs. The methods were evaluated against theoretical reference values obtained using the composite range-separated hybrid DFT method ωB97X-3c, enabling a clearly defined and consistent comparison. Among the tested approaches, the composite meta-GGA DFA r^{2}SCAN-3c delivered good overall accuracy but exhibited occasional convergence issues. In contrast, the hybrid composite method PBEh-3c also provided good accuracy while demonstrating more stable convergence behavior. Among the more efficient SQM methods, GFN2-xTB proved highly accurate, establishing itself as a valuable tool for modeling large NCI complexes. Regarding FF methods, both GFN-FF and GAFF performed well, making them suitable for high-throughput screening of large molecular assemblies. Owing to its scale and diversity, the LNCI16 benchmark set serves as a robust testing ground for evaluating the accuracy and reliability of current and newly developed computational methods. For instance, it has already been employed in the development of the recently introduced g-xTB method. Since interactions accumulate in such large systems, inconsistencies or inaccuracies in individual energy contributions, which might remain undetected in smaller molecules, can be systematically identified with this benchmark set.

In the following study, the computation of association free energies was investigated, which includes – in addition to gas-phase energies – also the calculation of thermal and solvation effects. To this end, the HS13L benchmark set was compiled, consisting of 13 supramolecular complexes with up to 266 atoms and featuring heavy main-group elements, for which reliable benchmark data are scarce. These

complexes exhibit unusual binding motifs involving heavy elements, making the set both challenging and informative. Importantly, experimental association free energies are available for all complexes. To compute these values, conformer ensembles were first generated using the CREST program at the GFN2-xTB level and subsequently re-ranked with the more accurate r^2SCAN-3c method via the CENSO program. Thermal contributions were calculated at the GFN2-xTB level, while solvation effects were modeled using COSMO-RS. The final computed association free energies showed excellent agreement with experimental values. These results confirm the suitability of the CREST+CENSO workflow for accurately predicting association free energies in complex supramolecular systems, which is of considerable relevance across chemical and biological applications. Furthermore, high-level DLPNO-CCSD(T1) reference values extrapolated to the CBS limit were generated, enabling direct comparison of gas-phase binding energies to assess the accuracy of electronic energy prediction methods. Beyond serving as a benchmark for electronic energy predictions, similar to LNCI16, the HS13L set also provides a valuable testing ground for evaluating methods used to compute thermal and solvation corrections. This is particularly important given the scarcity of experimental data for large solvated systems and the challenges associated with generating reliable high-level theoretical reference data for this free energy contribution. Herein, the availability of experimental association free energies enhances the value of the HS13L by anchoring theoretical predictions in reality. While the possibility of error cancellation between the separate methods must be considered when comparing multi-component properties like the free energy to experimental values, the high degree of agreement achieved using state-of-the-art methods lends strong support to the overall computational approach. This is especially relevant in light of the ongoing discrepancies between the high-level theoretical reference methods, coupled-cluster theory and quantum diffusion Monte Carlo, when applied to large NCI systems.[140] In such contexts, experimental reference data remain essential for guiding methodological development and ensuring robust validation. For the HS13L benchmark set, the coupled-cluster values - combined with solvation and thermal contributions - showed good agreement with experimental data, indicating the reliable accuracy of this method also for larger systems.

Another central focus of this thesis was the quantum chemical calculation of mass spectra. Despite the critical role of MS in many areas of science, a generally applicable and sufficiently accurate QC-based program for spectrum prediction has so far been lacking. To address this, a new program called QCxMS2 was developed as the successor of QCxMS, which is the only other fully automated QC-based EI-MS simulation program currently available. As in the HS13L study described above, direct comparison with experimental data is essential here to ensure the validity of the developed program. QCxMS2 introduces a new paradigm in mass spectrum simulation by constructing static reaction networks and computing reaction barriers, instead of relying on MD simulations as its predecessor QCxMS. This approach overcomes the inherent limitations of MD-based approaches, enabling the use of more accurate methods and circumventing the issue of limited simulation time. Thus, it allows for greater accuracy and efficiency, while also offering mechanistic insight into fragmentation processes. QCxMS2 was initially developed for EI-MS, where experimental reference data are more consistent due to standardized conditions. On a chemically diverse test set, QCxMS2 demonstrated improved accuracy and robustness compared to its predecessor, QCxMS. Herein, an important finding was that using higher levels of theory leads to more accurate spectra, supporting the overall foundations of the approach. In particular, molecules that pose challenges for current SQM methods, such as GFN2-xTB, could be more accurately described by employing the more reliable composite DFT method ωB97X-3c, while still maintaining affordable computational costs. The advantage of the reaction network approach compared to the MD-based one was demonstrated

by the ability to perform costly geometry optimizations and reaction path searches using low-cost SQM methods, while reaction energies and barriers could be refined through still feasible single-point calculations at the more accurate DFT level. Moreover, fragments resulting from rearrangement reactions were more reliably captured by QCxMS2. In only a few cases, important signals were missing, presumably due to fragments not generated by the MSREACT fragment generator – an issue that may be resolved in future work as discussed below.

The QCxMS2 program was also extended to simulate CID-MS, which is nowadays more widely used than EI-MS and thus arguably of greater practical relevance. Given the inconsistencies in existing CID-MS spectral libraries stemming from the wide variety of experimental setups, a diverse test set of experimental spectra was compiled in collaboration with the Engeser group in the organic chemistry department to ensure accurate and reliable reference data. Even more pronounced than for EI-MS, QCxMS2 outperforms QCxMS on the test set, highlighting its potential for reliable CID-MS simulations as well. A key advantage of the reaction network approach followed in QCxMS2 is that it allows different energy regimes to be rapidly explored without the need to rerun costly quantum chemical simulations, as required in MD-based approaches like QCxMS. This is particularly beneficial for CID-MS applications, given the high diversity of experimental setups currently used. In this context, different energy distributions may be tested in future works, which consider more aspects from the experiment, such as the type of the employed collision gas, its pressure and temperature, as well as the mass and collisional cross section of the input molecule to arrive at an energy distribution better approximating the transferred collisional activation energies. Also, collisional cooling processes could be considered. Initial attempts in this direction were not successful and were discarded for this thesis but may be worth revisiting in the future to improve the quality of the computed spectra. Furthermore, the challenge of genuine prediction of a CID spectrum remains, as it is not *a priori* known which protomer is most relevant for a spectrum. Thus, all currently available QC-based approaches require the computation of all protomers in a reasonable free energy window of approximately 40 kcal mol^{-1}. As with the MD-based approaches, no correlation between relative protomer energies and their distribution in the resulting spectrum could be found. However, QCxMS2 can generally predict interconversion between protomers through rearrangements better than the MD-based approaches, and thus gives more insight into this problem. As a result, QCxMS2 already provides strong support for mobile proton theory, but further research in this direction has to be conducted to potentially enable true prediction of a CID spectrum in the future. Overall, the good results both for EI- and CID-MS underscore the value and versatility of QCxMS2 as an accurate and robust tool for the quantum chemical calculations of mass spectra.

Nevertheless, QCxMS2 does not yet fully resolve the problem of mass spectra simulation, and deviations from experimental spectra remain. One of its key strengths lies in its modular and extensible architecture, which makes it well-suited to incorporate advances from across computational chemistry. A substantial part of this thesis, although not covered in detail, was devoted to the development of QCxMS2 as open-source software in the Fortran programming language. This makes the tool accessible for integration into broader computational workflows, a role that its predecessor QCxMS already played successfully. Open-source availability also invites participation from the scientific community for future development and adaptation. Several areas offer particularly strong potential for future improvement:

1. A more comprehensive yet still efficient fragment generation algorithm that captures all relevant dissociation pathways.

2. Implementation of newly developed QC methods or MLIPs for more accurate prediction of geometries, energies, and vibrational properties.

3. Implementation of improved transition state search algorithms that are both computationally efficient and robust.

All of these aspects represent active areas of research, and further developments are expected to increase the accuracy and applicability of QCxMS2. Herein, the greatest advancement is expected from Point 2, since it has been demonstrated for both CID-MS and EI-MS that the quality of the simulated spectra is highly sensitive to the level of theory used. In particular, the inclusion of the currently developed g-xTB method has already been shown in initial tests to provide spectra of DFT quality at tight-binding costs. Due to its efficiency, efficient low-cost methods such as g-xTB could also be employed to improve the fragment generation step, which likewise benefits from a more accurate description of the PES.

Looking forward, QCxMS2 offers promising opportunities for integration into ML-based *de novo* structure prediction workflows as schematically shown in the introduction in Figure 1.2. Candidate structures proposed by an ML model for an experimental spectrum of an unknown compound are evaluated with QCxMS2 to determine the most probable structure. In addition, an active learning approach could be followed. The simulated spectrum would then be compared to the measured one, providing a feedback signal to re-rank candidates or guide further model refinement. This combination of data-driven structure generation and physics-based validation could significantly improve the reliability of automated structure elucidation. If realized, such a pipeline would represent a major advance in analytical chemistry, with broad implications for fields like metabolomics and materials science.

Beyond its primary application in mass spectrometry, QCxMS2 may also serve as a benchmark platform for automated reaction discovery more generally. Because a mass spectrum encodes all relevant unimolecular fragmentation pathways and can be experimentally verified, it provides a powerful and interpretable validation target. As such, QCxMS2 is well-suited for benchmarking transition state search algorithms and electronic structure methods for barrier estimation, offering valuable insights for broader applications such as catalyst discovery and reaction mechanism design. In these areas, the completeness of a proposed reaction network is often challenging to verify, and confidence in the results can be provided by the use of experimental mass spectra as validation for computational workflows.

In conclusion, this thesis advances the field of computational chemistry by providing new insights into the accurate modeling of NCIs and by introducing a robust, next-generation program for the quantum chemical prediction of MS.

List of Acronyms

CID collision-induced dissociation
CBS complete basis set
CASE computer-assisted structure elucidation
DFA density functional approximation
DFT density functional theory
EI electron ionization
ESI electrospray ionization
ESS entropy similarity score
FF force field
GTO Gaussian-type orbital
HCD higher-energy collisional dissociation
IEE impact excess energy
IVR intramolecular vibrational energy redistribution
IP ionization potential
KS-DFT Kohn-Sham density functional theory
LMO localized molecular orbital
LD London dispersion
MLIP machine-learned interatomic potential
MS mass spectrometry
MAD mean absolute deviation
MD molecular dynamics
NCI non-covalent interaction
QC quantum chemistry
QET quasi-equilibrium theory
relMAD relative mean absolute deviation
RRHO rigid-rotor-harmonic-oscillator
SQM semiempirical quantum mechanical
TST transition state theory
WFT wave function theory

Bibliography

[1] J. Gorges, S. Grimme, and A. Hansen, *Reliable prediction of association (free) energies of supramolecular complexes with heavy main group elements - the HS13L benchmark set*, Phys. Chem. Chem. Phys. **24** (2022) 28831, DOI: 10.1039/d2cp04049b.

[2] J. Gorges, B. Bädorf, S. Grimme, and A. Hansen, *Efficient Computation of the Interaction Energies of Very Large Non-covalently Bound Complexes*, Synlett **34** (2022) 1135, DOI: 10.1055/s-0042-1753141.

[3] J. Gorges and S. Grimme, *QCxMS2 - a program for the calculation of electron ionization mass spectra via automated reaction network discovery*, Phys. Chem. Chem. Phys. **27** (2025) 6899, DOI: 10.1039/d5cp00316d.

[4] J. Gorges, M. Engeser, and S. Grimme, *Evaluation of the QCxMS2 method for the calculation of collision-induced-dissociation spectra via automated reaction network exploration*, ChemRxiv Prepr. (2025), DOI: 10.26434/chemrxiv-2025-gcws2.

[5] J. Gorges, M. Engeser, and S. Grimme, *Evaluation of the QCxMS2 Method for the Calculation of Collision-Induced Dissociation Spectra via Automated Reaction Network Exploration*, Journal of the American Society for Mass Spectrometry **36** (2025) 2276, DOI: 10.1021/jasms.5c00234.

[6] 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.

[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] R. B. Ouma, S. M. Ngari, and J. K. Kibet, *A review of the current trends in computational approaches in drug design and metabolism*, Discov. public Heal. **21** (2024) 108, DOI: 10.1186/s12982-024-00229-3.

[9] N. Marzari, A. Ferretti, and C. Wolverton, *Electronic-structure methods for materials design*, Nat. Mater. **20** (2021) 736, DOI: 10.1038/s41563-021-01013-3.

[10] M. Bursch, J. M. Mewes, A. Hansen, and S. Grimme, *Best-Practice DFT Protocols for Basic Molecular Computational Chemistry*, Angew. Chemie - Int. Ed. **61** (2022), DOI: 10.1002/anie.202205735.

[11] P. Deglmann, A. Schäfer, and C. Lennartz, *Application of quantum calculations in the chemical industry - An overview*, Int. J. Quantum Chem. **115** (2015) 107, DOI: 10.1002/qua.24811.

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

[13] I. V. Kolesnichenko and E. V. Anslyn, *Practical applications of supramolecular chemistry*, Chem. Soc. Rev. **46** (2017) 2385, DOI: 10.1039/c7cs00078b.

[14] P. L. Urban, *Quantitative mass spectrometry: An overview*, Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. **374** (2016) 20150382, DOI: 10.1098/rsta.2015.0382.

[15] J. D. Watson and F. H. Crick, *Molecular structure of nucleic acids: A Structure for deoxyribose nucleic acid*, 50 Years DNA **171** (2016) 83, DOI: 10.1038/nature01396.

[16] E. V. Anslyn, *Supramolecular analytical chemistry*, J. Org. Chem. **72** (2007) 687, DOI: 10.1021/jo0617971.

[17] L. Szente and J. Szejtli, *Highly soluble cyclodextrin derivatives: Chemistry, properties, and trends in development*, Adv. Drug Deliv. Rev. **36** (1999) 17, DOI: 10.1016/S0169-409X(98)00092-1.

[18] E. Arunkumar, C. C. Forbes, B. C. Noll, and B. D. Smith, *Squaraine-derived rotaxanes: Sterically protected fluorescent near-IR dyes*, J. Am. Chem. Soc. **127** (2005) 3288, DOI: 10.1021/ja042404n.

[19] J. R. Turkington, P. J. Bailey, J. B. Love, A. M. Wilson, and P. A. Tasker, *Exploiting outer-sphere interactions to enhance metal recovery by solvent extraction*, Chem. Commun. **49** (2013) 1891, DOI: 10.1039/c2cc37874d.

[20] F. Jensen, *Introduction to Computational Chemistry*, Vol. 2, Wiley, 2007.

[21] C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher, and S. Grimme, *Extended tight-binding quantum chemistry methods*, Wiley Interdiscip. Rev. Comput. Mol. Sci. **11** (2021) e01493, DOI: 10.1002/wcms.1493.

[22] 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*, Appl. Phys. Rev. **5** (2018) 31104, DOI: 10.1063/1.5020808.

[23] 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.

[24] B. M. Wood, M. Dzamba, X. Fu, M. Gao, M. Shuaibi, L. Barroso-Luque, K. Abdelmaqsoud, V. Gharakhanyan, J. R. Kitchin, D. S. Levine, et al., *UMA: A Family of Universal Models for Atoms*, arXiv preprint arXiv:2506.23971 (2025), DOI: 10.48550/arXiv.2506.23971.

[25] C. Villot, F. Ballesteros, D. Wang, and K. U. Lao, *Coupled Cluster Benchmarking of Large Noncovalent Complexes in L7 and S12L as Well as the C60Dimer, DNA-Ellipticine, and HIV-Indinavir*, J. Phys. Chem. A **126** (2022) 4326, DOI: 10.1021/acs.jpca.2c01421.

[26] J. H. Gross, *Mass Spectrometry: A Textbook*, Springer Science & Business Media, 2017 1, DOI: 10.1007/978-3-319-54398-7.

[27] J. B. Fenn, *Electrospray wings for molecular elephants (Nobel lecture)*, Angew. Chemie - Int. Ed. **42** (2003) 3871, DOI: 10.1002/anie.200300605.

[28] K. Håkansson and J. S. Klassen, *Ion Activation Methods for Tandem Mass Spectrometry*, Electrospray MALDI Mass Spectrom. Fundam. Instrumentation, Pract. Biol. Appl. Second Ed. **39** (2012) 571, DOI: 10.1002/9780470588901.ch16.

[29] T. Kind, H. Tsugawa, T. Cajka, Y. Ma, Z. Lai, S. S. Mehta, G. Wohlgemuth, D. K. Barupal, M. R. Showalter, M. Arita, and O. Fiehn, *Identification of small molecules using accurate mass MS/MS search*, Mass Spectrom. Rev. **37** (2018) 513, DOI: 10.1002/mas.21535.

[30] R. R. Da Silva, P. C. Dorrestein, and R. A. Quinn, *Illuminating the dark matter in metabolomics*, Proc. Natl. Acad. Sci. U. S. A. **112** (2015) 12549, DOI: 10.1073/pnas.1516878112.

[31] Y. Lai et al., *High-Resolution Mass Spectrometry for Human Exposomics: Expanding Chemical Space Coverage*, Environ. Sci. Technol. **58** (2024) 12784, DOI: 10.1021/acs.est.4c01156.

[32] J. Lederberg, "How DENDRAL was conceived and born," *Proc. ACM Conf. Hist. Med. Informatics, HMI 1987*, vol. 1987-Janua, 1987 5, DOI: 10.1145/41526.41528.

[33] R. K. Lindsay, E. A. Feigenbaum, B. G. Buchanan, and J. Lederberg, *Applications of artificial intelligence for chemical inference: the DENDRAL Project*, (No Title) **1392** (1980) 5962.

[34] M. Bohde, M. Manjrekar, R. Wang, S. Ji, and C. W. Coley, *DiffMS: Diffusion Generation of Molecules Conditioned on Mass Spectra*, arXiv Prepr. arXiv2502.09571 (2025).

[35] R. Bushuiev, A. Bushuiev, N. F. de Jonge, A. Young, F. Kretschmer, R. Samusevich, J. Heirman, F. Wang, L. Zhang, K. Dührkop, M. Ludwig, N. A. Haupt, A. Kalia, C. Brungs, R. Schmid, R. Greiner, B. Wang, D. S. Wishart, L. P. Liu, J. Rousu, W. Bittremieux, H. Rost, T. D. Mak, S. Hassoun, F. Huber, J. J. van der Hooft, M. A. Stravs, S. Böcker, J. Sivic, and T. Pluskal, *MassSpecGym: A benchmark for the discovery and identification of molecules*, Adv. Neural Inf. Process. Syst. **37** (2024) 110010.

[36] S. Grimme, *Towards first principles calculation of electron impact mass spectra of molecules*, Angew. Chemie - Int. Ed. **52** (2013) 6306, DOI: 10.1002/anie.201300158.

[37] J. Koopman and S. Grimme, *From QCEIMS to QCxMS: A Tool to Routinely Calculate CID Mass Spectra Using Molecular Dynamics*, J. Am. Soc. Mass Spectrom. **32** (2021) 1735, DOI: 10.1021/jasms.1c00098.

[38] J. Lee, D. J. Tantillo, L. P. Wang, and O. Fiehn, *Predicting Collision-Induced-Dissociation Tandem Mass Spectra (CID-MS/MS) Using Ab Initio Molecular Dynamics*, J. Chem. Inf. Model. **64** (2024) 7470, DOI: 10.1021/acs.jcim.4c00760.

[39] L. S. Kassel, *Studies in homogeneous gas reactions II: Introduction of quantum theory*, J. Phys. Chem. **32** (1928) 1065, DOI: 10.1021/j150289a011.

[40] O. K. Rice and H. C. Ramsperger, *Theories of unimolecular gas reactions at low pressures. II*, J. Am. Chem. Soc. **50** (1928) 617, DOI: 10.1021/ja01390a002.

[41] R. A. Marcus, *Unimolecular dissociations and free radical recombination reactions*, J. Chem. Phys. **20** (1952) 359, DOI: 10.1063/1.1700424.

[42] H. M. Rosenstock, M. B. Wallenstein, A. L. Wahrhaftig, and H. Eyring, *Absolute Rate Theory for Isolated Systems and the Mass Spectra of Polyatomic Molecules*, Proc. Natl. Acad. Sci. **38** (1952) 667, DOI: 10.1073/pnas.38.8.667.

[43] C. A. Bauer and S. Grimme, *How to Compute Electron Ionization Mass Spectra from First Principles*, J. Phys. Chem. A **120** (2016) 3755, DOI: 10.1021/acs.jpca.6b02907.

[44] *Program package for the quantum mechanical calculation of EI mass spectra using automated reaction network exploration* qcxms2, https://github.com/grimme-lab/QCxMS2.

[45] A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*, First, Dover Publications, 1996.

[46] E. Schrödinger, *Quantisierung als Eigenwertproblem*, Ann. Phys. **384** (1926) 361, DOI: 10.1002/andp.19263840404.

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

[48] P. A. Dirac, *A new notation for quantum mechanics*, Math. Proc. Cambridge Philos. Soc. **35** (1939) 416, DOI: 10.1017/S0305004100021162.

[49] 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.

[50] K. Raghavachari, G. W. Trucks, J. A. Pople, and M. Head-Gordon, *Reprint of: A fifth-order perturbation comparison of electron correlation theories*, Chem. Phys. Lett. **589** (2013) 37, DOI: 10.1016/j.cplett.2013.08.064.

[51] R. Nityananda, P. Hohenberg, and W. Kohn, *Inhomogeneous electron gas*, Resonance **22** (2017) 809, DOI: 10.1007/s12045-017-0529-3.

[52] N. Schuch and F. Verstraete, *Computational complexity of interacting electrons and fundamental limitations of density functionaltheory*, Nat. Phys. **5** (2009) 732, DOI: 10.1038/nphys1370.

[53] W. Kohn and L. J. Sham, *Self-consistent equations including exchange and correlation effects*, Phys. Rev. **140** (1965) A1133, DOI: 10.1103/PhysRev.140.A1133.

[54] J. P. Perdew, "Jacob's ladder of density functional approximations for the exchange-correlation energy," *AIP Conf. Proc.* AIP, 2003 1, DOI: 10.1063/1.1390175.

[55] 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.

[56] J. W. Furness, A. D. Kaplan, J. Ning, J. P. Perdew, and J. Sun, *Accurate and Numerically Efficient r2SCAN Meta-Generalized Gradient Approximation*, J. Phys. Chem. Lett. **11** (2020) 8208, DOI: 10.1021/acs.jpclett.0c02405.

[57] A. D. Becke, *Density-functional thermochemistry. III. The role of exact exchange*, J. Chem. Phys. **98** (1993) 5648, DOI: 10.1063/1.464913.

[58] 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.

[59] S. H. Vosko, L. Wilk, and M. Nusair, *Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis*, Can. J. Phys. **58** (1980) 1200, DOI: 10.1139/p80-159.

[60] 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.

[61] 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.

[62] S. Grimme, *Semiempirical hybrid density functional with perturbative second-order correlation*, J. Chem. Phys. **124** (2006) 034108, DOI: 10.1063/1.2148954.

[63] S. Grimme, A. Hansen, J. G. Brandenburg, and C. Bannwarth, *Dispersion-Corrected Mean-Field Electronic Structure Methods*, Chem. Rev. **116** (2016) 5105, DOI: 10.1021/acs.chemrev.5b00533.

[64] 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.

[65] 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.

[66] E. Caldeweyher, C. Bannwarth, and S. Grimme, *Extension of the D3 dispersion coefficient model*, J. Chem. Phys. **147** (2017) 34112, DOI: 10.1063/1.4993215.

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

[68] 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.

[69] 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.

[70] E. R. Johnson and A. D. Becke, *A post-Hartree-Fock model of intermolecular interactions: Inclusion of higher-order corrections*, J. Chem. Phys. **124** (2006) 174104, DOI: 10.1063/1.2190220.

[71] 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.

[72] B. V. Derjaguin, *The force between molecules*, Prog. Surf. Sci. **40** (1992) 151, DOI: 10.1016/0079-6816(92)90041-F.

[73] F. Neese and E. F. Valeev, *Revisiting the atomic natural orbital approach for basis sets: Robust systematic basis sets for explicitly correlated and conventional correlated ab initio methods?* J. Chem. Theory Comput. **7** (2011) 33, DOI: 10.1021/ct100396y.

[74] M. Müller, A. Hansen, and S. Grimme, *An atom-in-molecule adaptive polarized valence single-ζ atomic orbital basis for electronic structure calculations*, J. Chem. Phys. **159** (2023), DOI: 10.1063/5.0172373.

[75] C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher, and S. Grimme, *Extended tight-binding quantum chemistry methods*, Wiley Interdiscip. Rev. Comput. Mol. Sci. **11** (2021) e01493, DOI: 10.1002/wcms.1493.

[76] 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.

[77] N. D. Mermin, *Thermal properties of the inhomogeneous electron gas*, Phys. Rev. **137** (1965) 1441, DOI: 10.1103/PhysRev.137.A1441.

[78] 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.

[79] V. K. Prasad, A. Otero-De-La-Roza, and G. A. Dilabio, *Fast and Accurate Quantum Mechanical Modeling of Large Molecular Systems Using Small Basis Set Hartree-Fock Methods Corrected with Atom-Centered Potentials*, J. Chem. Theory Comput. **18** (2022) 2208, DOI: 10.1021/acs.jctc.1c01128.

[80] H. Neugebauer, B. Bädorf, S. Ehlert, A. Hansen, and S. Grimme, *High-throughput screening of spin states for transition metal complexes with spin-polarized extended tight-binding methods*, J. Comput. Chem. **44** (2023) 2120, DOI: 10.1002/jcc.27185.

[81] S. Chimie physique, *Journal de chimie physique 1995*, vol. 18, H. Kündig.

[82] W. Heisenberg, *Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik*, Zeitschrift für Phys. **43** (1927) 172, DOI: 10.1007/BF01397280.

[83] S. Grimme, *Supramolecular binding thermodynamics by dispersion-corrected density functional theory*, Chem. - A Eur. J. **18** (2012) 9955, DOI: 10.1002/chem.201200497.

[84] J. D. Chai and M. Head-Gordon, *Systematic optimization of long-range corrected hybrid density functionals*, J. Chem. Phys. **128** (2008) 84106, DOI: 10.1063/1.2834918.

[85] S. Spicher and S. Grimme, *Single-Point Hessian Calculations for Improved Vibrational Frequencies and Rigid-Rotor-Harmonic-Oscillator Thermodynamics*, J. Chem. Theory Comput. **17** (2021) 1701, DOI: 10.1021/acs.jctc.0c01306.

[86] 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.

[87] 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.

[88] 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.

[89] 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.

[90] 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.

[91] 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.

[92] M. Dantus, *Tracking Molecular Fragmentation in Electron-Ionization Mass Spectrometry with Ultrafast Time Resolution*, Acc. Chem. Res. **57** (2024) 845, DOI: 10.1021/acs.accounts.3c00713.

[93] M. Ruberti and V. Averbukh, *Advances in modeling attosecond electron dynamics in molecular photoionization*, Wiley Interdiscip. Rev. Comput. Mol. Sci. **13** (2023) e1673, DOI: 10.1002/wcms.1673.

[94] T. Baer and P. M. Mayer, *Statistical Rice-Ramsperger-Kassel-Marcus quasiequilibrium theory calculations in mass spectrometry*, J. Am. Soc. Mass Spectrom. **8** (1997) 103, DOI: 10.1016/S1044-0305(96)00212-7.

[95] L. Drahos and K. Vékey, *Mass kinetics: A theoretical model of mass spectra incorporating physical processes, reaction kinetics and mathematical descriptions*, J. Mass Spectrom. **36** (2001) 237, DOI: 10.1002/jms.142.

[96] H. Eyring, *The activated complex in chemical reactions*, J. Chem. Phys. **3** (1935) 63, DOI: 10.1063/1.1749604.

[97] M. Barbatti, *Defining the temperature of an isolated molecule*, J. Chem. Phys. **156** (2022) 204304, DOI: 10.1063/5.0090205.

[98] K. Vékey, *Internal energy effects in mass spectrometry*, J. Mass Spectrom. **31** (1996) 445, DOI: 10.1002/(SICI)1096-9888(199605)31:5<445::AID-JMS354>3.0.CO;2-G.

[99] Á. Kuki, G. Shemirani, L. Nagy, B. Antal, M. Zsuga, and S. Kéki, *Estimation of activation energy from the survival yields: Fragmentation study of leucine enkephalin and polyethers by tandem mass spectrometry*, J. Am. Soc. Mass Spectrom. **24** (2013) 1064, DOI: 10.1007/s13361-013-0635-8.

[100] G. E. P. Box and M. E. Muller, *A Note on the Generation of Random Normal Deviates*, Ann. Math. Stat. **29** (1958) 610, DOI: 10.1214/aoms/1177706645.

[101] K. B. Shelimov, D. E. Clemmer, R. R. Hudgins, and M. F. Jarrold, *Protein structure in Vacuo: Gas-phase conformations of BPTI and cytochrome c*, J. Am. Chem. Soc. **119** (1997) 2240, DOI: 10.1021/ja9619059.

[102] M. F. Jarrold, *Unfolding, refolding, and hydration of proteins in the gas phase*, Acc. Chem. Res. **32** (1999) 360, DOI: 10.1021/ar960081x.

[103] J. Lee, D. J. Tantillo, L. P. Wang, and O. Fiehn, *Impact of Protonation Sites on Collision-Induced Dissociation-MS/MS Using CIDMD Quantum Chemistry Modeling*, J. Chem. Inf. Model. **64** (2024) 7457, DOI: 10.1021/acs.jcim.4c00761.

[104] P. Pracht, C. A. Bauer, and S. Grimme, *Automated and efficient quantum chemical determination and energetic ranking of molecular protonation sites*, J. Comput. Chem. **38** (2017) 2618, DOI: 10.1002/jcc.24922.

[105] 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.

[106] R. W. Floyd, *Algorithm 97: shortest path*, Commun. ACM **5** (1962) 345, DOI: 10.1145/367766.368168.

[107] A. Bom, M. Bradley, K. Cameron, J. K. Clark, J. Van Egmond, H. Feilden, E. J. MacLean, A. W. Muir, R. Palin, D. C. Rees, and M. Q. Zhang, *A novel concept of reversing neuromuscular block: Chemical encapsulation of rocuronium bromide by a cyclodextrin-based synthetic host*, Angew. Chemie - Int. Ed. **41** (2002) 265, DOI: 10.1002/1521-3773(20020118)41:2<265::AID-ANIE265>3.0.CO;2-Q.

[108] K. Suresh, V. López-Mejías, S. Roy, D. F. Camacho, and A. J. Matzger, *Leveraging Framework Instability: A Journey from Energy Storage to Drug Delivery*, Synlett **31** (2020) 1573, DOI: 10.1055/s-0040-1707139.

[109] R. J. Phipps, *Cluster Preface: Non-Covalent Interactions in Asymmetric Catalysis*, Synlett **27** (2016) 1024, DOI: 10.1055/s-0035-1561933.

[110] P. Renzi and M. Bella, *Design of Experiments: A rational approach toward non-covalent asymmetric organocatalysis*, Synlett **28** (2017) 306, DOI: 10.1055/s-0036-1588654.

[111] S. Kassem, T. Van Leeuwen, A. S. Lubbe, M. R. Wilson, B. L. Feringa, and D. A. Leigh, *Artificial molecular motors*, Chem. Soc. Rev. **46** (2017) 2592, DOI: 10.1039/c7cs00245a.

[112] M. Müller, A. Hansen, and S. Grimme, *ω B97X-3c: A composite range-separated hybrid DFT method with a molecule-optimized polarized valence double- ζ basis set*, J. Chem. Phys. **158** (2023), DOI: 10.1063/5.0133026.

[113] S. Grimme, A. Hansen, S. Ehlert, and J. M. Mewes, *R2SCAN-3c: A "swiss army knife" composite electronic-structure method*, J. Chem. Phys. **154** (2021) 064103, DOI: 10.1063/5.0040021.

[114] 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.

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

[116] M. V. Rekharsky and Y. Inoue, *Complexation thermodynamics of cyclodextrins*, Chem. Rev. **98** (1998) 1875, DOI: 10.1021/cr970015o.

[117] D. Phlip, *Supramolecular chemistry: Concepts and perspectives. By J.-M. Lehn, VCH, Weinheim 1995, x, 271 pp., softcover, DM 58.00, ISBN 3-527-2931 1-6*, Adv. Mater. **8** (1996) 866, DOI: 10.1002/adma.19960081029.

[118] M. A. Pitt and D. W. Johnson, *Main group supramolecular chemistry*, Chem. Soc. Rev. **36** (2007) 1441, DOI: 10.1039/b610405n.

[119] 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.

[120] 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.

[121] A. Klamt, *The COSMO and COSMO-RS solvation models*, Wiley Interdiscip. Rev. Comput. Mol. Sci. **1** (2011) 699, DOI: 10.1002/wcms.56.

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

[123] 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.

[124] A. V. Marenich, C. J. Cramer, and D. G. Truhlar, *Generalized born solvation model SM12*, J. Chem. Theory Comput. **9** (2013) 609, DOI: 10.1021/ct300900e.

[125] 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.

[126] 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.

[127] 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.

[128] P. S. Brahmkshatriya, P. Dobes, J. Fanfrlik, J. Rezac, K. Paruch, A. Bronowska, M. Lepsík, and P. Hobza, *Quantum Mechanical Scoring: Structural and Energetic Insights into Cyclin-Dependent Kinase 2 Inhibition by Pyrazolo[1,5-a]pyrimidines*, Curr. Comput. Aided-Drug Des. **9** (2013) 118, DOI: 10.2174/1573409911309010011.

[129] J. Koopman and S. Grimme, *Calculation of Electron Ionization Mass Spectra with Semiempirical GFNn-xTB Methods*, ACS Omega **4** (2019) 15120, DOI: 10.1021/acsomega.9b02011.

[130] V. Ásgeirsson, C. A. Bauer, and S. Grimme, *Quantum chemical calculation of electron ionization mass spectra for general organic and inorganic molecules*, Chem. Sci. **8** (2017) 4879, DOI: 10.1039/c7sc00601b.

[131] N. van Staalduinen and C. Bannwarth, *MolBar: a molecular identifier for inorganic and organic molecules with full support of stereoisomerism*, Digit. Discov. **3** (2024) 2298, DOI: 10.1039/d4dd00208c.

[132] 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.

[133] Y. Li, T. Kind, J. Folz, A. Vaniya, S. S. Mehta, and O. Fiehn, *Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification*, Nat. Methods **18** (2021) 1524, DOI: 10.1038/s41592-021-01331-z.

[134] Y. Li and O. Fiehn, *Flash entropy search to query all mass spectral libraries in real time*, Nat. Methods **20** (2023) 1475, DOI: 10.1038/s41592-023-02012-9.

[135] P. Kebarle and U. H. Verkcerk, *Electrospray: From Ions in solution to Ions in the gas phase, what we know now*, Mass Spectrom. Rev. **28** (2009) 898, DOI: 10.1002/mas.20247.

[136] M. Schäfer, M. Drayß, A. Springer, P. Zacharias, and K. Meerholz, *Radical cations in electrospray mass spectrometry: Formation of open-shell species, examination of the fragmentation behaviour in ESI-MSn and reaction mechanism studies by detection of transient radical cations*, European J. Org. Chem. **2007** (2007) 5162, DOI: 10.1002/ejoc.200700199.

[137] D. Fu, S. G. Habtegabir, H. Wang, S. Feng, and Y. Han, *Understanding of protomers/deprotomers by combining mass spectrometry and computation*, Anal. Bioanal. Chem. **415** (2023) 3847, DOI: 10.1007/s00216-023-04574-1.

[138] J. Koopman and S. Grimme, *Calculation of Mass Spectra with the QCxMS Method for Negatively and Multiply Charged Molecules*, J. Am. Soc. Mass Spectrom. **33** (2022) 2226, DOI: 10.1021/jasms.2c00209.

[139] A. R. Dongré, J. L. Jones, Á. Somogyi, and V. H. Wysocki, *Influence of peptide composition, gas-phase basicity, and chemical modification on fragmentation efficiency: Evidence for the mobile proton model*, J. Am. Chem. Soc. **118** (1996) 8365, DOI: 10.1021/ja9542193.

[140] Y. S. Al-Hamdani, P. R. Nagy, A. Zen, D. Barton, M. Kállay, J. G. Brandenburg, and A. Tkatchenko, *Interactions between large molecules pose a puzzle for reference quantum mechanical methods*, Nat. Commun. **12** (2021) 3927, DOI: 10.1038/s41467-021-24119-3.

Appendix

APPENDIX A

Appendix: Efficient Computation of the Interaction Energies of Very Large Non-covalently Bound Complexes

Johannes Gorges,[†‡] Benedikt Bädorf,[†‡], Stefan Grimme,[†] Andreas Hansen[†] *Received: 21 July 2022*
Published online: 30 November 2022

Reprinted in Appendix A from Ref. J. Gorges, B. Bädorf, S. Grimme, and A. Hansen, *Efficient Computation of the Interaction Energies of Very Large Non-covalently Bound Complexes*, Synlett **34**.10 (2022) 1135, DOI: 10.1055/s-0042-1753141 with permission from Thieme.

Own contributions

- Conducting parts of the force-field and DFT calculations
- Interpretation of the results
- Co-writing and revising the manuscript

[†]Mulliken Center for Theoretical Chemistry, Universität Bonn, Beringstr. 4, D-53115 Bonn, Germany
[‡]These authors contributed equally.

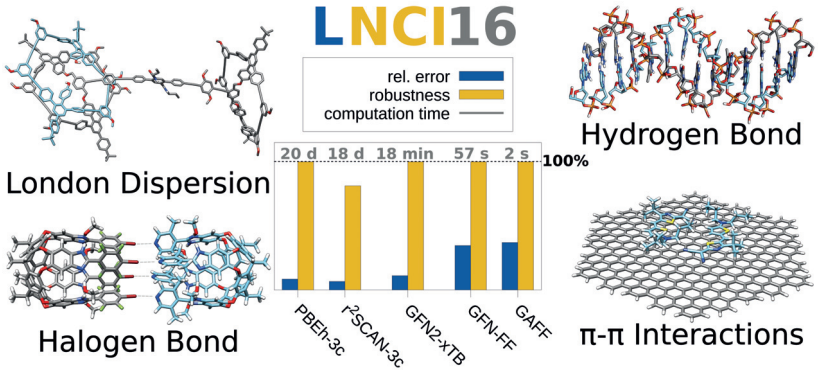
Synlett J. Gorges et al. Cluster

Efficient Computation of the Interaction Energies of Very Large Non-covalently Bound Complexes

Johannes Gorges◊
Benedikt Bädorf◊
Stefan Grimme
Andreas Hansen*

Mulliken Center for Theoretical Chemistry, Institute for Physical and Theoretical Chemistry, University of Bonn, Beringstr. 4, 53115 Bonn, Germany
hansen@thch.uni-bonn.de
◊These authors contributed equally

Published as part of the Cluster
Dispersion Effects



Received: 21.07.2022
Accepted after revision: 04.10.2022
Published online: 29.11.2022 (Version of Record)
DOI: 10.1055/s-0042-1753141; Art ID: ST-2022-07-0335-C

Abstract We present a new benchmark set consisting of 16 large non-covalently bound systems (LNCI16) ranging from 380 up to 1988 atoms and featuring diverse interaction motives. Gas-phase interaction energies are calculated with various composite DFT, semi-empirical quantum mechanical (SQM), and force field (FF) methods and are evaluated using accurate DFT reference values. Of the employed QM methods, PBEh-3c proves to be the most robust for large systems with a relative mean absolute deviation (relMAD) of 8.5% with respect to the reference interaction energies. r²SCAN-3c yields an even smaller relMAD, at least for the subset of complexes for which the calculation could be converged, but is less robust for systems with smaller HOMO–LUMO gaps. The inclusion of Fock-exchange is therefore important for the description of very large non-covalent interaction (NCI) complexes in the gas phase. GFN2-xTB was found to be the best performer of the SQM methods with an excellent result of only 11.1% deviation. From the assessed force fields, GFN-FF and GAFF achieve the best accuracy. Considering their low computational costs, both can be recommended for routine calculations of very large NCI complexes, with GFN-FF being clearly superior in terms of general applicability. Hence, GFN-FF may be routinely applied in supramolecular synthesis planning.

Key words non-covalent interaction energies, benchmarking, dispersion, composite methods, semi-empirical methods, force fields

1 Introduction

Supramolecular chemistry finds application in many areas of chemistry,[1] such as in drug delivery,[2,3] the design of artificial molecular motors,[4] and in catalysis.[5,6] The structures and functionalities of these compounds are mainly governed by non-covalent interactions (NCIs). Of the various interaction types observed in supramolecular complexes, such as hydrogen and halogen bonding, π–π-stacking, and ion-dipolar interactions, London dispersion (LD) forces contribute a large amount of the interaction energy in many cases.[7] This comparably weak interaction type is omnipresent in chemistry and biology. Since it depends on the contact surface of the respective complex, LD interactions can equate to sizeable energy contributions for large systems and must not be neglected.[8,9]

Computational chemistry has become a powerful tool for modeling structures and predicting the binding motifs of NCI complexes.[10] Herein, a major challenge for theoretical methods is the accurate description of LD effects.[11] Computationally efficient approaches, such as the D3[12,13] and D4[14] dispersion schemes, the exchange-hole dipole moment (XDM)[15,16] approach or the non-local VV10 correction,[17] can be used for the description of this long-range electronic correlation effect. Combined with accurate density functional approximations (DFAs), the challenging task of reliably predicting the interaction energies of large supramolecular complexes consisting of up to 2000 atoms becomes feasible.[18,19] For the system size given in the LNCI16 set, dispersion contributions beyond the pairwise attributions considered become increasingly important[20] and can be efficiently modeled by the three-body Axilrod–Teller–Muto (ATM)[21,22] term in D3 or D4, and the many-body dispersion (MBD)[23,24] correction scheme.[25] For large π systems with small HOMO–LUMO gaps, as is more often the case with the binding of adsorbates on metal surfaces, higher-order dispersion terms can become more important.[26]

Another crucial point for supramolecular complexes is the basis set superposition error (BSSE),[27] which leads to an overestimation of interaction energies if DFAs or wave

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

function theory (WFT) methods are used in combination with smaller basis sets. In most cases, the use of basis sets with at least triple ζ quality is necessary to obtain a diminishing BSSE.[25] This becomes computationally unfeasible for systems with more than 500 atoms. Therefore, several empirical approaches have been developed to reduce the BSSE, such as the geometrical counterpoise correction (gCP)[28] and the related beyond pairwise approach of the DFT-C method.[29]

A technical problem for the DFT calculations of large molecules in the gas phase, i.e., without electrostatic screening via a proper solvent model, is that the HOMO–LUMO gap often diminishes with increasing system sizes, even for mostly chemically saturated proteins.[30] In combination with the notorious underestimation of HOMO–LUMO gaps by common (meta-)GGA DFAs, the gaps may even approach zero, leading to very unstable self-consistent field (SCF) iterations and unreliable results. While hybrid DFAs suffer less from this problem due to mixing in of exact exchange into the functional, their serious drawback is the order of magnitude higher computational cost.

Due to their robustness and low computational costs, efficient semi-empirical quantum mechanical (SQM) methods are powerful tools for the structural and energetic screening of large molecular systems.[31,32] For example, in recent studies,[33,34] SQM methods and computationally much cheaper but also much more simplified atomistic force field (FF) methods have been used for modeling even larger supramolecular structures such as those frequently encountered in structural biology.[35]

As the much lower computation time of SQM and even more FF methods comes at the cost of a higher degree of empiricism, it is crucial to carefully benchmark SQM and FF methods against accurate reference values.[36] For this purpose, several important NCI benchmark sets have been composed, such as L7,[18] S30L,[19] and the 'extra-large' EXL8,[37] covering systems with up to 1027 atoms. A subset of the latter was used in a recent DFT benchmark study.[38] However, except for EXL8, which has limited statistical validity due to the small number of systems, there are, to our knowledge, no benchmark sets with NCI complexes with significantly more than 250 atoms, although these systems are just as relevant and occur in many areas of chemistry and structural biology.[8,9] Therefore, we propose the new LNCI16 benchmark set with systems ranging from 380 up to 1988 atoms. By covering diverse interactions such as hydrogen bonding, halogen bonding, π–π-stacking, and with a main focus on London dispersion bound complexes, this set aims to represent the great diversity in supramolecular chemistry. By comparison to accurate DFT reference values, various efficient composite DFT methods as well as low-cost SQM and FF methods are evaluated.

In this work, the interaction energy (E_{int}) is calculated via the supermolecular approach:

$$E_{int} = E(AB) - E(A) - E(B) \quad (1),$$

where E is the gas-phase electronic energy of complex AB, host A, and guest B, respectively. This approach is generally applicable for any system size and is only limited by the computational costs of the employed method for the calculation of the complex. Since we aim at a theoretical benchmark, i.e., not comparing with experimentally measured association energies, all energies are calculated in the geometry of the complex and thus neglect the fragment (monomer) relaxation energy. In the gas phase, the NCIs calculated by various methods can unambiguously be compared with each other, which would not be possible in solvated-state calculations for which no common model exists. The calculation of gas-phase energies for charged systems turned out to be problematic in many cases due to the missing screening effects of a solvent, thus leading to almost vanishing HOMO–LUMO gaps (<1 eV). Therefore, many systems, especially those with higher charges as often present in proteins, had to be excluded from the benchmark set.

After a description of the test set in the next section, we summarize the computational details followed by a statistical evaluation of the employed DFT, SQM, and FF methods. Finally, we discuss computational timings and give method recommendations for computation of the interaction energies of large NCI complexes.

2 The LNCI16 Benchmark Set

Figure 1 shows the 16 optimized complex structures contained within the benchmark set, details of which are briefly described below.

The first three systems all include boron atoms, with systems **1** and **2** consisting of the same porous organic cage host. The apolar guest benzene is mostly bound via dispersion interactions (BpocBenz). However, the host is also able to form hydrogen bonds with the polar guest methanol, as in BpocMeOH.[39] System **3** is a boron–nitrogen nanotube with an α-ALA guest molecule that is mostly bound by LD interactions.[40] Gramicidin A (**4**) is an ionophoric antibiotic, which forms helical dimers that are connected by hydrogen bonds.[41] H-bonds are also the main binding motif in system **5** consisting of a linear oligocarbamate guest molecule and a helical host.[42] DNA is arguably the most notable supramolecular complex and is predominantly bound by H-bonds. Hence, a neutralized cutout of this important system is also included in our benchmark set. Next is a cutout of a protein–ligand complex of the Src homology 3 (SH3) protein (**7**), which is able to recognize specific proline sequences.[43] Systems **3**–**7** are part of the EXL8 benchmark set reported by Ni et.al.[37] and were taken in their original geometries.

With the following two protein-ligand cutouts, charged systems were also considered in the test set. The first is the tyrosine-protein kinase 2 (TYK2) with an ejm46 ligand in the binding pocket, which is of interest for the treatment of

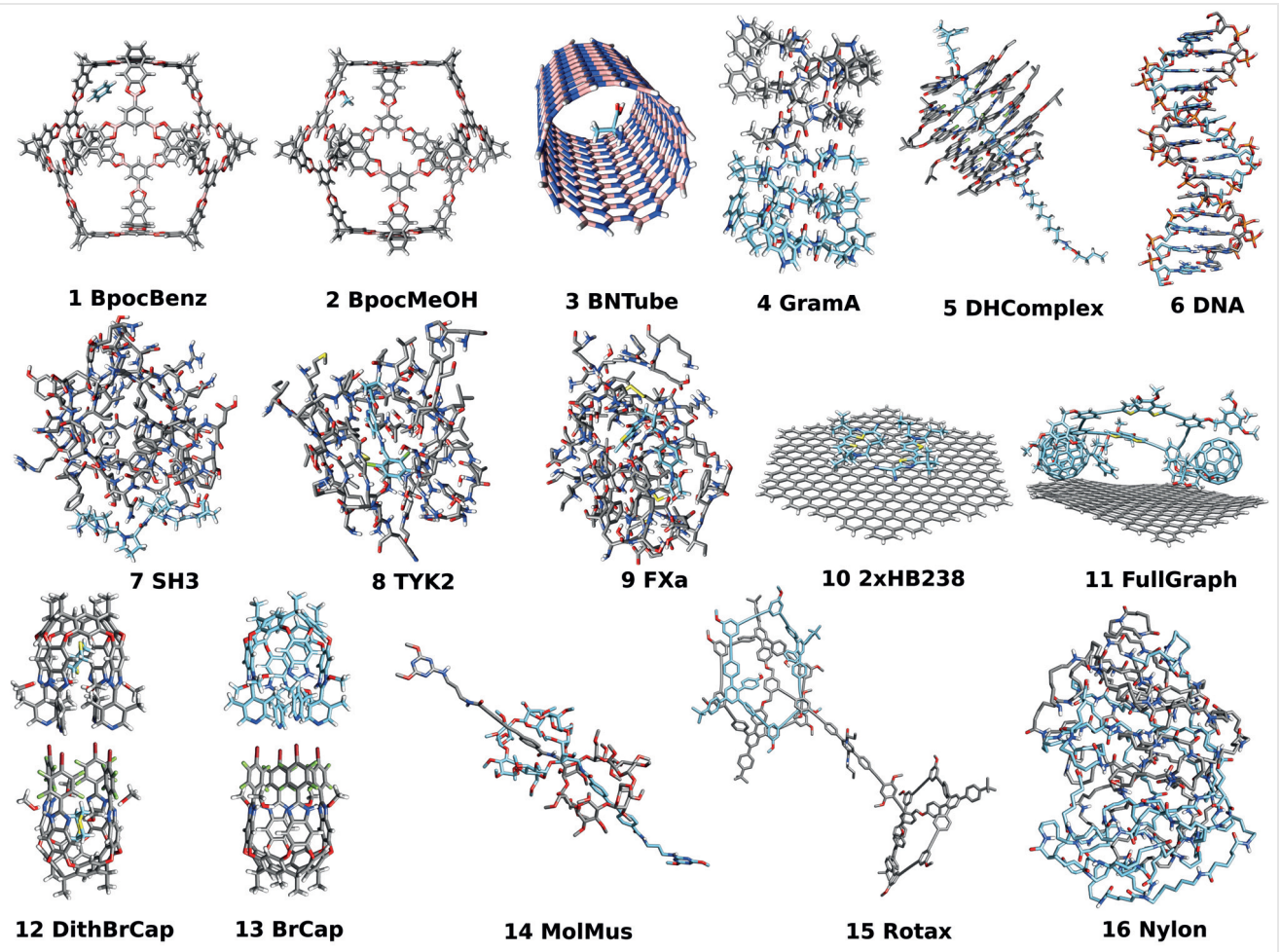


Figure 1 Structures and names of the complexes comprising the LNCI16 benchmark set (systems 3–7 from EXL8). Carbon atoms in the host molecules are colored light gray while they are depicted in light blue in the guest molecules.

inflammatory diseases.[44] In the second system, Rivaroxaban, an anticoagulant drug, is bound to the activated serine protease factor X (FXa).[44] Another field in which noncovalent interactions play a major role is the adsorption processes of molecules on surfaces.[45] System **10** shows such adsorption which is dominated by π–π-interactions between a graphene sheet and two dipolar donor–acceptor dye molecules (2xHB238).[46] These dyes belong to a class of polymethine dyes (also called merocyanines). System **11** consists of an interaction between a graphene sheet and a fullerene-based macrocycle co-adsorbate, which was synthesized by the Höger group.[47]

Furthermore, two systems consisting of a halogen-bonded capsule were chosen to further assess this challenging interaction type.[48,49] In DithBrCap, the interaction energy between two dithiane guest molecules and the capsule is investigated, while for BrCap the interaction through halogen bonding between the two parts of the capsule is computed.[50] System **14** is a model for a molecular muscle and represents an important class of molecular machines. The muscle is given in its 'contracted' form and is bound via hydrogen bonds and LD interactions between its two identical parts.[34] Another important class of supramolecular complexes is rotaxanes in which the host and guest molecules are mechanically interlocked. System **15** is a phenylacetylene-based complex belonging to this class with the major type of interaction being LD.[51] The final and largest system of the benchmark set is a snapshot taken from a molecular dynamics simulation of long nylon chains that are intertwined forming a nanoparticle via hydrogen bonding.[52] The behavior of such plastic nanoparticles is of interest in environmental chemistry.[53] Furthermore, the absorption of small molecules in clothing fabrics such as nylon is another field of interest.[54,55]

An overview of the complexes with their charges and calculated reference interaction energies (see Section 4) is given in Table 1.[56]

Table 1 Overview of the Investigated Complexes, Their Charges and Calculated ωB97X-3c[56] Reference Interaction Energies (kcal mol^{-1})[a]

Complex	Charge	Ref.	ΔE_{int} (kcal mol^{-1})
BpocBenz (1)	0	39	– 6.81
BpocMeOH (2)	0	39	– 6.19
BNTube (3)	0	39, 40	– 14.32
GramA (4)	0	39, 41	– 36.30
DHComplex (5)	0	39, 42	– 57.57
DNA (6)	0	39	–363.30
SH3 (7)	0	39, 43	– 25.65
TYK2 (8)	+1	44	– 49.03
FXa (9)	–2	44	–105.27
2xHB238 (10)	0	46	– 74.92
FullGraph (11)	0	47	– 74.13
DithBrCap (12)	0	50	– 45.63
BrCap (13)	0	50	– 21.12
MolMus (14)	0	34	– 62.58
Rotax (15)	0	51	– 55.89
Nylon (16)	0	52, 53	–566.23

[a] The respective reference from which the structure was taken is given. Additional details on the geometries can be found in the Supporting Information.

The inclusion of boron atoms in systems **1**–**3** may be problematic for many force field methods as they are rarely parameterized for each atom type. Nevertheless, these interesting systems were included because they make the benchmark set more diverse. Force fields may also be evaluated using the subset without boron atoms (systems **4**–**16**), which still includes a reasonable size of 13 systems compared to the L7 or EXL8 benchmark sets.

3 Computational Details

All methods evaluated in this work are given in Table 2. The PM6-D3H4X[57–59] and PM7[60] calculations were carried out with MOPAC2016.[61] GFN2-xTB,[32,62] GFN1-xTB,[63] GFN0-xTB[64] and GFN-FF[65] calculations were performed using xTB.[66] The xTB-IFF[67] energies were computed using the xTB-IFF program.[68] D3(BJ)[12,13,15] dispersion contributions for the B97M functional with and without the inclusion of the ATM term were conducted with the s-dftd3[69] standalone program. D4[14,70] dispersion energies, which include the ATM term by default, were calculated using the dft-d4[71] 3.4.0 standalone program. D4-MBD dispersion energies were computed with dft-d4 2.4.0.

The DFTB engine of the Amsterdam Modeling Suite (AMS)[72] was used to perform the DFTB calculations with the Quasinano2015 parametrization.[73,74] Additionally, the engine was employed for the calculation of GFN1-xTB charges, which were then fed into the AMS ForceField engine[75] for the UFF[76] calculations. The Open Babel program package[77,78] was used to perform the MMFF94[79,80] and Ghemical[81] calculations using the respective default charges. However, the default charge model of the Ghemical force field predicted wrong charges for the charged systems **8** and **9** (+2 instead of –2 for FXa and +3 instead of +1 in the case of TYK2). Open Babel was also employed for the GAFF[82] calculations, which were conducted with Gasteiger charges[83] as well as GFN2-xTB charges.

Table 2 Tested Composite QM, SQM, and FF Methods with the Respective Applied Dispersion Corrections[a]

Method	Dispersion	Ref.
Composite QM		
B97M-V-C	VV10	29, 91
PBEh-3c	D3(BJ)-ATM	88
r2SCAN-3c	D4	85
B97-3c	D3(BJ)-ATM	87
HF-3c	D3(BJ)	86
SQM		
DFTB(Quasinano)	D3(BJ)-ATM	73, 74
GFN2-xTB	D4	92
GFN1-xTB	D3(BJ)	63
PM7	D2	60
PM6-D3H4X	D3	57
GFN0-xTB	D4	94
FF		
GFN-FF	D4	65
UFF	LJ	76
xTB-IFF	D4	67
GAFF	LJ	82
MMFF94	Buf-14-7	79, 80
Ghemical	LJ	81

[a] Additional computational details are given in the Supporting Information.

TURBOMOLE (V. 7.5.1)[84] was used for the r^{2}SCAN-3c,[85] HF-3c,[86] B97-3c,[87] and PBEh-3c[88] calculations. The ωB97X-3c[56] reference values were calculated with the ORCA (V. 5.0.2)[89,90] quantum chemistry package. Due to convergence problems in Q-Chem using the B97M-V[17,91] DFA in combination with the def2-SVPD[92,93] basis set, single-point calculations for this method were performed with TURBOMOLE, while DFT-C[29] correction terms were computed with the Q-Chem program package.[94] The non-local VV10 correction was computed non-self-consistently.

4 Generation of Reference Values

The generation of reliable reference values is crucial for every theoretical benchmark study. For the system size covered by the LNCI16 set, this is an especially challenging task. The use of the common 'gold standard' (CCSD(T)), even with local approximations, which have been successfully used for NCI complexes of up to about 1000 atoms,[39,95] is not feasible for the systems comprised in the LNCI16 set because of the enormous computational cost. In this work, a newly developed, efficient, range-separated DFA composite method termed ωB97X-3c is applied to generate reference interaction energies. It employs a deeply contracted valence double-ζ atomic orbital (AO) basis set (vDZVP), which is specially optimized for molecules in combination with large core ECPs and a refitted D4 dispersion correction. Due to its molecular (DFT) optimization and the specially adapted D4 parameterization, this composite method is essentially BSSE free, despite its small basis set. Consequently, interaction energies from ωB97X-3c for existing NCI benchmark sets show very small deviations from the basis-set converged results of the parent method ωB97X-D4[96]/def2-QZVPP[92] with revised D4 parametrization[56] (Table 3). A detailed description and extensive evaluation of this DFA will be published in the near future.[56] Note that ωB97X-3c, with its specific D4 parametrization[56] in combination with the uniquely developed basis set ensuring a small BSSE, is among the most accurate DFAs ever tested for NCIs, and yet is computationally feasible for NCI complexes with a few thousand atoms. The respective mean absolute deviations (MADs) of the popular B3LYP-D4[97,98] method are given for comparison and are in most cases significantly larger. Based on this excellent performance for diverse NCI benchmark sets, ωB97X-3c is a suitable reference method, while still being affordable for the computation of systems of up to a few thousand atoms. Due to its range-separated hybrid DFA character, this method does not suffer from the aforementioned gap problem, and we observed no SCF convergence problems, even for small gap test systems with up to 2795 atoms. Since this presented benchmark study mainly focuses on SQM and FF methods, the estimated errors of ωB97X-3c are expected to be much smaller compared to the typical errors of the SQM and FF methods, thus enabling a meaningful evaluation of these low-cost methods.

Table 3 Mean Absolute Deviations (MAD) (kcal mol^{-1}) from Accurate Reference Values (Mostly of CCSD(T)/CBS Quality)[a]

	ωB97X-3c	ωB97X-D4/QZ	B3LYP-D4/QZ
S30L	1.7	1.9	5.3
IONPI19	1.0	1.0	1.2
L7	1.6	0.7	2.4
S66	0.3	0.1	0.3
R160x6	0.3	0.2	0.2
HB300SPX	0.3	0.2	0.5

[a] MAD values for ωB97X-3c, ωB97X-D4/QZ (with revised D4 parametrization[41]), and the popular B3LYP-D4/QZ method for NCI benchmark sets S30L,[19] IONPI19,[99] L7,[18,100] S66,[101] R160x6,[102] and HB300SPX.[103] ωB97X-3c and ωB97X-D4 values are taken from Ref. 56; B3LYP-D4 data was taken from Ref. 85 (see the Supporting Information for more detailed information); QZ corresponds to def2-QZVPP.

5 Results and Discussion

Considering the broad range of interaction energies covered in the LNCI16 set, the usual statistical error measures, such as mean deviation (MD) and mean absolute deviation (MAD), may be strongly biased by the large interaction energy cases. A downscaling of these very large interaction energies would have a certain arbitrariness, hence we decided to base our statistical evaluation on the relative deviations from the reference method. This was also suggested by Piecuch to enable a meaningful evaluation of the performance of DFT methods for NCIs,[104] considering a relative deviation below 5% as 'chemical accuracy' in this context. Additionally, we provide MDs and MADs for all the evaluated methods in the Supporting Information. An overestimation of the interaction energy ('overbinding') by a method results in a negative relative deviation, whereas an underestimation ('underbinding') is defined as a positive relative deviation. Although systems with a ωB97X-3c HOMO–LUMO gap below 3.5 eV were removed, convergence problems with the tested (meta-)GGA DFT methods r^{2}SCAN-3c, B97-3c, and B97M emerged for systems **9–11**, such that interaction energies could not be calculated with these methods for the corresponding complexes.

First, the contribution of higher-order dispersion terms than the pairwise attributions in the DFT-D framework is discussed for the B3LYP functional. This functional was chosen for the discussion as its D4 dispersion contribution is usually in good agreement with WFT dispersion energy estimates.[104] Figure 2 shows the relative contributions to the D4 energy of the three-body ATM term and of the many-body approach (MBD), including higher-order dispersion terms.[14] Interestingly, the ATM contribution to the interaction energies of systems BpocBenz, BpocMeOH, and BrCap is close to zero, which can be attributed to the small contact surface consisting mainly of pairwise NCI contacts between host and guest in the complex. However, we generally observe a significant contribution of the ATM term of 8.0% on average, which indicates the importance of incorporating many-body dispersion effects by including the three-body ATM term. This is consistent with observations made in previous studies on smaller NCI complexes.[106] Consequently, we expect overbinding of methods which neglect the three-body dispersion. However, the inclusion of higher-order dispersion terms in the many-body D4-MBD approach is on average only 0.5% different to the D4-ATM

values with a maximum of 1.4% for BrCap. This demonstrates that the D4 correction sufficiently includes many-body dispersion terms by the inclusion of the ATM term, even for large systems found in the LNCI16 set.

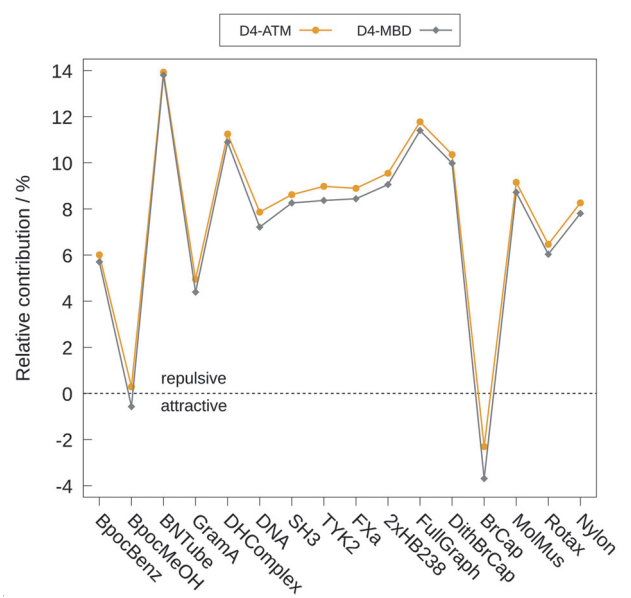


Figure 2 Relative contributions of the ATM/MBD-terms to the respective overall D4-ATM/MBD dispersion energies. The D4 correction with only pairwise terms is given as a reference (total dispersion energies are calculated as a sum of the pairwise and higher-order contributions). The lines connecting the data points are given for better visibility.

For the system sizes covered in the LNCI16 set, DFT calculations are only feasible with rather small basis sets which, however, are subject to a significant BSSE, especially for NCI complexes. In this respect, the '3c'[85] or '-C'[29] composite DFT methods are particularly suited as they combine relatively small basis sets with an approximate BSSE correction or by absorbing it in the basis set and a D4 refit (see Section 4). The importance of such correction schemes can be exemplified for the B97M-V functional using the def2-SVPD basis set, for which a relMAD of 58.9% without any BSSE correction is obtained for the LNCI16 set. The relMAD can be clearly reduced to 25.1% upon applying the DFT-C correction (see the Supporting Information for more details). Hence, this correction scheme seems to work effectively, even for the very large NCI complexes evaluated in the present study, which can also be seen in the overall good accuracy of B97M-D3-ATM-C (relMAD: 8.8%).

Figure 3 shows the relative deviations from the reference values for all the tested composite QM methods. Since the SCF iterations of the evaluated (meta-)GGA functionals did not converge for FXa, 2xHB238, and FullGraph, only the remaining 13 systems are included in the respective relMADs of these methods. In contrast, PBEh-3c and HF-3c could be converged for all systems of the LNCI16 set, stressing the importance of Fock exchange for the robust treatment of large NCI complexes in the gas phase.

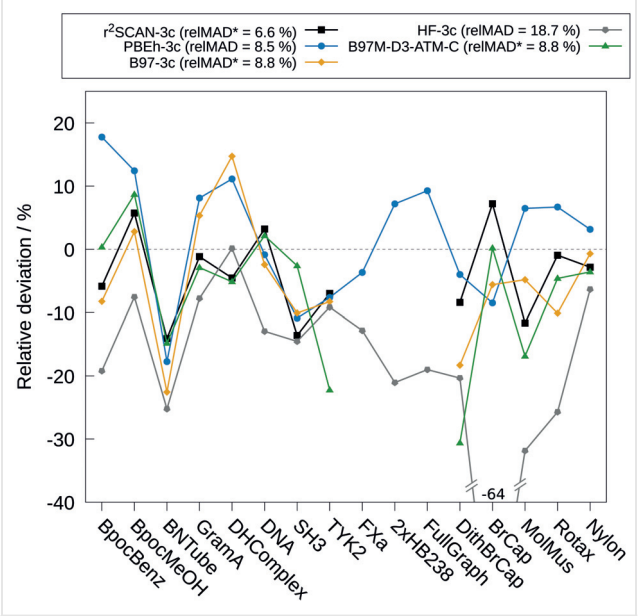


Figure 3 Relative deviations (given in %) to ωB97X-3c of the tested composite QM methods. The lines connecting the data points are given for better visibility. * Only systems **1–8** and **12–16**, for which the SCF converged, were taken into consideration for determination of the relMADs.

The HF-3c method, however, systematically overestimates the interaction energies of the investigated complexes, which is in line with the reported behavior of HF-3c for the S12L supramolecular NCI benchmark set.[86] With an overestimation of more than 60% for BrCap, HF-3c is not even able to describe this system qualitatively correctly. Therefore, this method can only be recommended to a very limited extent for the calculation of the interaction energies of very large NCI complexes.

B97-3c and B97M-D3-ATM-C yield interaction energies of comparable accuracy but also show a tendency to overestimate the interaction energies, presumably due to some residual BSSE. The smallest relative MAD of 6.6% is obtained by r²SCAN-3c. Thus, the method comes close to the chemical accuracy for NCIs of ca. 5% for the LNCI16 set, which further confirms its generally good performance for non-covalently bound systems.[85,99] In terms of accuracy, PBEh-3c performs second best among all the tested methods, and more importantly, it converges for all systems of the LNCI16 set. For the subset where the (meta-)GGA DFAs also converged, PBEh-3c achieved a relMAD of 8.9%. Moreover, this composite hybrid DFT method also yields the smallest relative MD of –1.8% (also the smallest relMD for the mentioned subset), therefore systematic errors can be largely excluded. Overall, PBEh-3c provides the best compromise

between accuracy and robustness among all the composite QM methods tested and once again underlines the effectiveness of the '3c' approach.

By employing much cruder approximations such as, among others, the use of very small basis sets, SQM methods are even able to compute very large systems with more than a thousand atoms, often with acceptable accuracy.[32,107] The relative deviations with respect to the LNCI16 reference values for the best-performing SQM methods tested in this work are shown in Figure 4. The observed tendency of PM6-D3H4X to overestimate the interaction energies (relMD: 16.8%) is in agreement with previous studies on smaller NCI complexes.[86] To a similar extent, GFN1-xTB also shows this tendency (relMD: 16.6%). Closely followed by PM6-D3H4X, DFTB(Quasinano) is the least accurate among the evaluated SQM methods. DFTB(Quasinano) underestimates the interaction energies of most of the hydrogen-bonded systems in the test set (BpocMeOH, GramA, DNA, and Nylon). With a relative MAD of only 11.1%, which is outstanding for an SQM method, GFN2-xTB provides by far the best accuracy within this class of methods. It does not show systematic under- or overbinding. Although this was to some extent expected, since GFN2-xTB was specifically parameterized to accurately describe non-covalent interactions; this also holds true for the diverse and very large NCI complexes of the LNCI16 set. In contrast, GFN1-xTB, although also parameterized with a focus on non-covalent interactions, is clearly outperformed by its successor, suggesting that the parameterization of GFN2-xTB in combination with its modified Hamiltonian also works better for very large NCI complexes, consistent with the performance for the S30L set.[62] A similar overbinding tendency of GFN1-xTB was observed for the ACONFL[108] set, which consists of conformers with long alkane chains, whereas this was not observed for GFN2-xTB. This can be explained by the larger basis set for hydrogen in GFN1-xTB, which leads to an underestimation of the repulsive NCI contacts. In addition, the higher multipole terms in the GFN2-xTB Hamiltonian improve the description of hydrogen bonding.[62]

Without self-consistent charge iterations, the GFN0-xTB method saves computation time compared to GFN1-/GFN2-xTB (cf. Figure 6),[64] but at the price of significantly larger errors (relMAD: 26.3%). Contrary to the reasonable accuracy of PM6-D3H4X, PM7 drastically overestimates most interaction energies (relMAD: 70.1%, relMD: –68%). Besides the already known poor performance of PM7 for dispersion bound complexes,[109] it also showed bad results for other non-covalent interaction types (e.g., H bonds) in this study. One reason for a systematic overbinding of the method might be the use of the D2[110] dispersion correction, which generally performs worse than the more sophisticated D3 or D4 methods.[111]

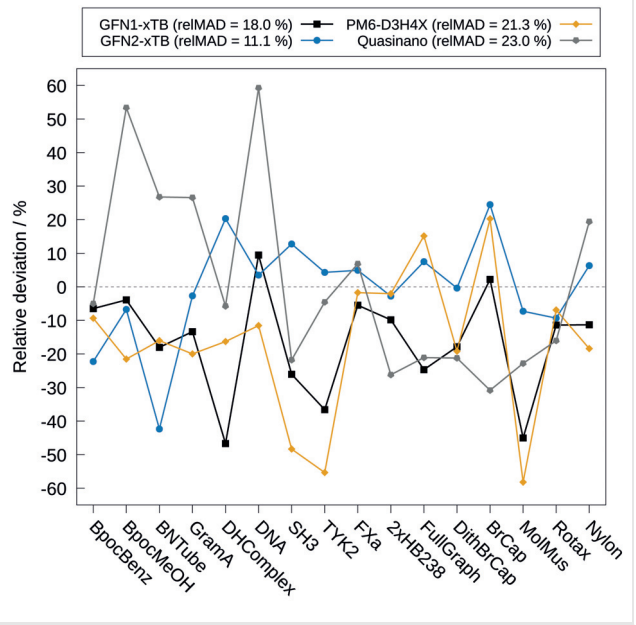


Figure 4 Relative deviations (given in %) from the ωB97X-3c reference values presented for the evaluated SQM methods that performed best (the PM7 and GFN0-xTB results can be found in the Supporting Information). The lines connecting the data points are given for better visibility.

Figure 5 shows the four best-performing force fields assessed for this study. Detailed results for MMFF94 and Ghemical can be found in the Supporting Information and their overall performance is discussed below. GFN1-xTB charges were used for the UFF calculations, while GFN2-xTB charges were employed for the GAFF calculations, which significantly improved the results of both force fields (see the Supporting Information for more details). Notably, the UFF force field yields very inaccurate interaction energies, predicting BpocMeOH, GramA, DNA, and Nylon to be unbound complexes (relative deviations larger than +100%). All these complexes have in common that hydrogen bonding plays a significant role, which indicates that UFF may not be able to describe this type of interaction qualitatively correctly. The xTB-IFF force field generally predicts quite accurate interaction energies, except for TYK2 and FXa, which are the only two charged complexes in the benchmark set. Problems of the xTB-IFF force field in describing charged systems were already reported in the original publication.[67] Excluding these two systems from the statistical evaluation, results in a small relMAD of only 23.0%. In fact, xTB-IFF is by far the most accurate force field for neutral systems of the LNCI16 set among all the tested FF methods. Considering the complete benchmark set, GFN-FF yields the smallest relMAD of 34.7% with a tendency to overbind (relMD: –21.5%). The GAFF force field is only slightly less accurate (relMAD: 37.0%) compared to GFN-FF and shows

only a moderate underbinding with a relMD of 11.0%. GAFF does not include parameters for boron and hence, hydrogen parameters are used instead.

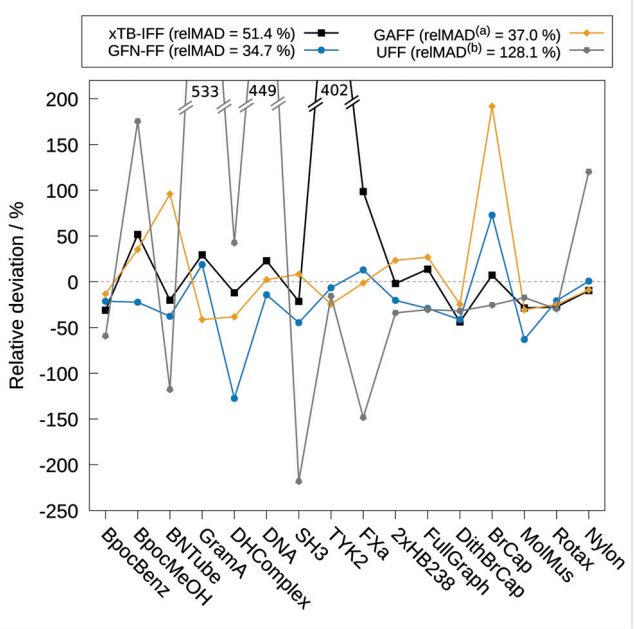


Figure 5 Relative deviations (given in %) from the ωB97X-3c reference values presented for the evaluated FF methods that performed best (MMFF94 and Ghemical results can be found in the Supporting Information). The lines connecting the data points are given for better visibility. [a] Using GFN2-xTB charges. [b] Using GFN1-xTB charges.

Surprisingly, this rather crude workaround yields quite accurate results for systems **1** and **2**. However, the large deviation for system **3** may be attributed to the missing genuine parametrization for boron. The halogen capsule (BrCap) is predicted to be unbound, verifying the known inaccurate description of halogen bonding with GAFF.[112] The MMFF94 force field is not parameterized for boron and since it was not possible to use standard parameters in Open Babel for these cases, we could not obtain results for the boron-containing complexes **1**–**3**. Hence, these systems are not considered in the respective relMAD. The force field systematically underbinds every system of the LNCI16 set resulting in a relMAD (and relMD) of 93.0%, and is therefore not recommended for the modeling of large NCI complexes. Due to the incorrectly assigned global charges for systems **8** and **9** by the Ghemical FF, both systems were identified as outliers and are thus excluded from the relMAD. The force field underbinds almost every system (relMD: 112.9%) and shows an extremely poor performance that is comparable to that of UFF (relMAD: 115.8%), also incorrectly predicting systems like the DNA double-helix to be unbound.

Finally, the computational costs of the assessed methods were compared. The wall times were determined by summing up the total wall times needed for the Nylon complex, host, and guest calculations and multiplying them by the number of cores that were used. Please note that the actual wall times were smaller because the calculations were run in parallel on multiple processes. However, since the number of CPU processes used in each case was different for the tested methods, we have normalized the wall times to one process for a comparison that is as unbiased as possible. The wall times obtained in this way for the Nylon interaction energy as well as the respective relMADs for the whole benchmark set of all the methods used are shown in Figure 6. The methods are divided into three groups, namely composite QM, SQM, and FF.

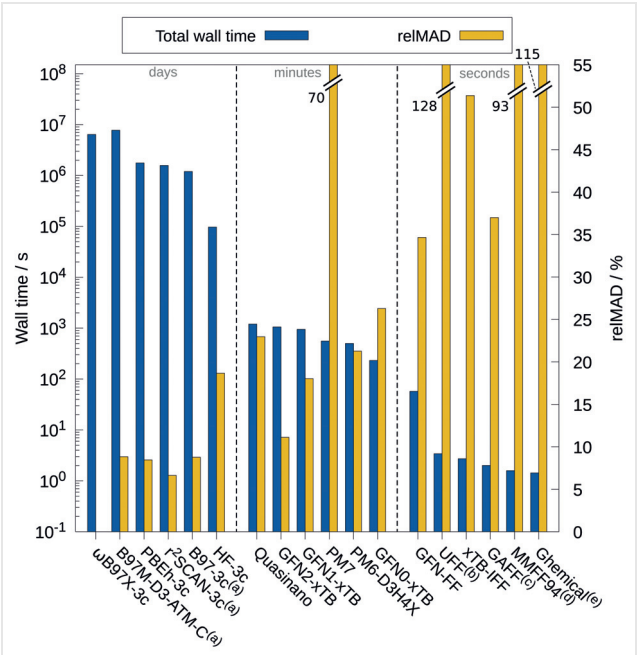


Figure 6 Computational wall times (Intel® Xeon® E5-2660 v4 @ 2.00 GHz CPU) for the calculation of the interaction energies of the largest system in the LNCI16 set (Nylon), together with the respective relMADs for the complete benchmark set: (a) only converged systems were taken into account for the relMADs, (b) using GFN1-xTB charges, (c) using GFN2-xTB charges, (d) systems including boron were excluded from the statistical analysis, and (e) charged systems could not be calculated and were not considered in the respective relMAD. (Total wall times needed for the Nylon complex, host, and guest calculations were summed and multiplied by the number of cores used.)

The composite QM methods require weeks of computation time and are probably too expensive to be used routinely for systems of this size. A special case is the HF-3c method, which has a theoretical wall time of roughly one day for the Nylon interaction energy (on one CPU core), but yields even less accurate results than the best-performing SQM methods, which are still more than 90 times faster. Therefore, we do not recommend the use of HF-3c for the description of very large NCI complexes. The best-performing composite QM method is PBEh-3c, providing a good compromise of computation time, accuracy, and robustness. Although the r^2-SCAN-3c method required a slightly

smaller wall time (10% less than PBEh-3c) and also yielded a smaller relMAD compared to PBEh-3c, it is considerably less robust than the latter, as already discussed above. Hence, we consider PBEh-3c as the best performer for the LNCI16 set among all the composite QM methods tested in this work.

The discussed SQM methods typically required a computation time of several minutes and are therefore well suited for the presented system size, at least if not too many systems need to be computed (too slow for MD simulations or large-scale screening applications for the system sizes). Although the GFN2-xTB method requires the second largest wall time among all the tested SQM methods, it can still be considered as the best performer within this class of methods due to its exceptional accuracy. As intended, the GFN0-xTB method requires less computation time than the other GFN methods, but the significantly lower accuracy makes the application of GFN0-xTB unattractive for very large NCI systems. The PM7 method required roughly the same computation time as PM6-D3H4X, but considering the very large errors for the LNCI16 set, we cannot recommend the PM7 method for the treatment of very large NCI complexes.

Lastly, the performance of the tested force fields has been evaluated. These methods are designed for rapid calculations of systems significantly exceeding the sizes presented in this study. Hence, all the evaluated force fields only require a few seconds to compute the interaction energy of the Nylon complex. The GFN-FF force field obtained the best accuracy, but is a factor of 30 slower than the GAFF force field with only slightly worse accuracy. In addition, it is, to the best of our knowledge, the only FF besides UFF that is parameterized for the entire periodic table up to Rn. The computation time needed for the GFN2-xTB charges used for the GAFF calculations is not considered in the timings. Similarly, to obtain interaction energies with xTB-IFF, GFN1-xTB or GFN2-xTB, calculations of the monomers have to be performed first, which we have excluded from the discussion. However, since FF methods are usually used for the screening of many different NCI binding motifs of a given molecule, these costs can be considered negligible.

6 Conclusions

For the theoretical description of huge supramolecular systems, the development of efficient yet accurate computational methods to describe NCIs, most prominently London dispersion, represents a big challenge. For the evaluation of these methods, we compiled a new benchmark set called LNCI16, consisting of very large supramolecular complexes. We used the newly introduced ωB97X-3c efficient composite DFT method to calculate accurate reference energies for complexes with up to 2000 atoms in less than a month of computation time. We chose to calculate the interaction energies in the gas phase to allow a meaningful comparison of all methods, since there is no single solvation model that is implemented for all the methods tested.

By assessing various composite QM, SQM, and force field methods against high-level DFT data we were able to show that the majority of the investigated methods can describe these interactions with sufficient accuracy. Although the r^{2}SCAN-3c method achieves the best results, it shows SCF convergence problems for large systems with small HOMO–LUMO gaps. For these cases, we recommend the more robust and nearly as accurate PBEh-3c low-cost DFT method. This method is suitable for calculating limited accurate single-point energies, e.g., in the last step of a screening workflow. Of the SQM methods, GFN2-xTB in particular has an excellent accuracy-to-cost ratio with a relMAD of 11.1% and a computation time of only 18 minutes for roughly 2000 atoms on a typical desktop computer. Therefore, GFN2-xTB can be used in screening applications with a limited number of structures, e.g., in a refinement step. Among the FF methods, GFN-FF yields the highest accuracy, whereas GAFF has the best accuracy-to-cost ratio. In applications of screening large structural databases or MD simulations, both are the method of choice. For uncharged systems, xTB-IFF is an excellent choice. Moreover, all the recommended methods can be calculated with freely available software (see the Supporting Information) and may be helpful in supramolecular synthesis planning. This is especially true for GFN-FF, as this force field is both robust and generally applicable, and provides interaction energies of large NCI complexes with reasonable accuracy and low computational resources.

Some popular methods show large deviations from the reference energies. The commonly used UFF force field predicts an unbound DNA helix, for example, and is therefore not recommended for computing large NCI complexes. The PM7 method is also unsuitable for this purpose as it drastically overbinds almost all complexes in the benchmark set.

The LNCI16 benchmark set may also serve as a fit or validation set for the parameterization of new SQM, FF, or machine-learning methods as well as (many-body) dispersion corrections.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors thank M. Müller, T. Gasevic, and S. Ehlert for technical help regarding the calculations. We also thank Prof. Dr. S. Höger, Dr. O. Hollóczki, and Dr. M. de Wergifosse for providing structures.

Supporting Information

Supporting information for this article is available online at https://doi.org/10.1055/s-0042-1753141.

References

(1) Kolesnichenko, I. V.; Anslyn, E. V. *Chem. Soc. Rev.* **2017**, *46*, 2385.
(2) Bom, A.; Bradley, M.; Cameron, K.; Clark, J. K.; van Egmond, J.; Feilden, H.; MacLean, E. J.; Muir, A. W.; Palin, R.; Rees, D. C.; Zhang, M.-Q. *Angew. Chem. Int. Ed.* **2002**, *41*, 265.
(3) Suresh, K.; López-Mejías, V.; Roy, S.; Camacho, D. F.; Matzger, A. J. *Synlett* **2020**, *31*, 1573.
(4) Kassem, S.; Leeuwen, T. V.; Lubbe, A. S.; Wilson, M. R.; Feringa, B. L.; Leigh, D. A. *Chem. Soc. Rev.* **2017**, *46*, 2592.
(5) Phipps, R. *Synlett* **2016**, *27*, 1024.
(6) Renzi, P.; Bella, M. *Synlett* **2016**, *28*, 306.
(7) Stone, A. *The Theory of Intermolecular Forces*; Oxford University Press: Oxford, **2013**.
(8) Riley, K. E.; Hobza, P. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 3.
(9) Song, Q.; Cheng, Z.; Kariuki, M.; Hall, S. C. L.; Hill, S. K.; Rho, J. Y.; Perrier, S. *Chem. Rev.* **2021**, *121*, 13936.
(10) Piskorz, T. K.; Martí-Centelles, V.; Young, T. A.; Lusby, P. J.; Duarte, F. *ACS Catal.* **2022**, *12*, 5806.
(11) Grimme, S.; Hansen, A.; Brandenburg, J. G.; Bannwarth, C. *Chem. Rev.* **2016**, *116*, 5105.
(12) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104.
(13) Grimme, S.; Ehrlich, S.; Goerigk, L. *J. Comput. Chem.* **2011**, *32*, 1456.
(14) Caldeweyher, E.; Ehlert, S.; Hansen, A.; Neugebauer, H.; Spicher, S.; Bannwarth, C.; Grimme, S. *J. Chem. Phys.* **2019**, *150*, 154122.
(15) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2005**, *123*, 154101.
(16) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2005**, *122*, 154104.
(17) Vydrov, O. A.; Voorhis, T. V. *J. Chem. Phys.* **2010**, *133*, 244103.
(18) Sedlak, R.; Janowski, T.; Pitoňák, M.; Řezáč, J.; Pulay, P.; Hobza, P. *J. Chem. Theory Comput.* **2013**, *9*, 3364.
(19) Sure, R.; Grimme, S. *J. Chem. Theory Comput.* **2015**, *11*, 3785.
(20) von Lilienfeld, O. A.; Tkatchenko, A. *J. Chem. Phys.* **2010**, *132*, 234109.
(21) Muto, Y. *J. Phys. Math. Soc. Jpn.* **1943**, 629.
(22) Axilrod, B. M.; Teller, E. *J. Chem. Phys.* **1943**, *11*, 299.
(23) Tkatchenko, A.; DiStasio, R. A.; Car, R.; Scheffler, M. *Phys. Rev. Lett.* **2012**, *108*, 236402.
(24) DiStasio, R. A.; Gobre, V. V.; Tkatchenko, A. *J. Phys.: Condens. Matter* **2014**, *26*, 213202.
(25) Risthaus, T.; Grimme, S. *J. Chem. Theory Comput.* **2013**, *9*, 1580.
(26) Maurer, R. J.; Ruiz, V. G.; Tkatchenko, A. *J. Chem. Phys.* **2015**, *143*, 102808.
(27) Boys, S.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553.
(28) Kruse, H.; Grimme, S. *J. Chem. Phys.* **2012**, *136*, 154101.
(29) Witte, J.; Neaton, J. B.; Head-Gordon, M. *J. Chem. Phys.* **2017**, *146*, 234105.
(30) Lever, G.; Cole, D. J.; Hine, N. D. M.; Haynes, P. D.; Payne, M. C. *J. Phys.: Condens. Matter* **2013**, *25*, 152101.
(31) Christensen, A. S.; Kubař, T.; Cui, Q.; Elstner, M. *Chem. Rev.* **2016**, *116*, 5301.
(32) Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2021**, *11*, e1493.
(33) Bohle, F.; Grimme, S. *J. Serb. Chem. Soc.* **2019**, *84*, 837.
(34) Kohn, J.; Spicher, S.; Bursch, M.; Grimme, S. *Chem. Commun.* **2022**, *58*, 258.
(35) Mackerell, A. D. *J. Comput. Chem.* **2004**, *25*, 1584.
(36) Řezáč, J.; Hobza, P. *Chem. Rev.* **2016**, *116*, 5038.
(37) Ni, Z.; Guo, Y.; Neese, F.; Li, W.; Li, S. *J. Chem. Theory Comput.* **2021**, *17*, 756.
(38) Wu, D.; Truhlar, D. G. *J. Chem. Theory Comput.* **2021**, *17*, 3967.
(39) Spicher, S.; Bursch, M.; Grimme, S. *J. Phys. Chem. C* **2020**, *124*, 27529.
(40) Wang, Z.; Liu, Y. F.; Yan, H.; Tong, H.; Mei, Z. *J. Phys. Chem. A* **2017**, *121*, 1833.
(41) Ketchem, R.; Hu, W.; Cross, T. *Science* **1993**, *261*, 1457.
(42) Wang, X.; Wicher, B.; Ferrand, Y.; Huc, I. *J. Am. Chem. Soc.* **2017**, *139*, 9350.
(43) Nguyen, J. T.; Turck, C. W.; Cohen, F. E.; Zuckermann, R. N.; Lim, W. A. *Science* **1998**, *282*, 2088.
(44) Ehrlich, S.; Göller, A. H.; Grimme, S. *ChemPhysChem* **2017**, *18*, 898.
(45) Raffaini, G.; Ganazzoli, F. *J. Appl. Biomater. Biomech.* **2010**, *8*, 135.
(46) Bürckstümmer, H.; Tulyakova, E. V.; Deppisch, M.; Lenze, M. R.; Kronenberg, N. M.; Gsänger, M.; Stolte, M.; Meerholz, K.; Würthner, F. *Angew. Chem. Int. Ed.* **2011**, *50*, 11628.
(47) Bahr J., Höger S., Jester S., Brandenburg J. G., Grimme S.; manuscript in preparation.
(48) Riley, K. E.; Hobza, P. *Phys. Chem. Chem. Phys.* **2013**, *15*, 17742.
(49) Kozuch, S.; Martin, J. M. L. *J. Chem. Theory Comput.* **2013**, *9*, 1918.
(50) Sure, R.; Grimme, S. *Chem. Commun.* **2016**, *52*, 9893.
(51) Schweez, C.; Shushkov, P.; Grimme, S.; Höger, S. *Angew. Chem. Int. Ed.* **2016**, 553328.
(52) Hollóczki, O. *Int. J. Quantum Chem.* **2021**, *121*, e26372.
(53) Hollóczki, O.; Gehrke, S. *Sci. Rep.* **2019**, *9*, 16013.
(54) Noble, R. E. *Sci. Total Environ.* **2000**, *262*, 1.
(55) Bahl, V.; Jacob, P.; Havel, C.; Schick, S. F.; Talbot, P. *PLoS ONE* **2014**, *9*, e108258.
(56) Müller M., Hansen A., Grimme S. *ωB97X-3c: A composite range-separated hybrid DFT method with a molecule-optimized polarized valence double-ζ basis set J. Chem. Phys.*, under review.
(57) Stewart, J. J. P. *J. Mol. Model.* **2007**, *13*, 1173.
(58) Řezáč, J.; Hobza, P. *J. Chem. Theory Comput.* **2012**, *8*, 141.
(59) Řezáč, J.; Hobza, P. *Chem. Phys. Lett.* **2011**, *506*, 286.
(60) Stewart, J. J. P. *J. Mol. Model.* **2013**, *19*, 1.
(61) Molecular Orbital PACkage (Version 19.179L); https://github.com/openmopac/mopac (accessed Nov 15, 2022)
(62) Bannwarth, C.; Ehlert, S.; Grimme, S. *J. Chem. Theory Comput.* **2019**, *15*, 1652.
(63) Grimme, S.; Bannwarth, C.; Shushkov, P. *J. Chem. Theory Comput.* **2017**, *13*, 1989.
(64) Pracht, P.; Caldeweyher, E.; Ehlert, S.; Grimme, S. A. *ChemRxiv* **2019**, preprint; DOI: 10.26434/chemrxiv.8326202.v1.
(65) Spicher, S.; Grimme, S. *Angew. Chem. Int. Ed.* **2020**, *59*, 15665.
(66) *Semiempirical Extended Tight-Binding Program Package*, *xtb*; https://github.com/grimme-lab/xtb (accessed Nov 15, 2022).
(67) Grimme, S.; Bannwarth, C.; Caldeweyher, E.; Pisarek, J.; Hansen, A. *J. Chem. Phys.* **2017**, *147*, 161708.
(68) General Intermolecular Force Field based on Tight-Binding Quantum Chemical Calculations (Version 1.1); https://github.com/grimme-lab/xtbiff (accessed Nov 15, 2022)
(69) Reimplementation of the DFT-D3 program (Version 0.5.0); https://github.com/dftd3/simple-dftd3/releases/tag/v0.5.0 (accessed Nov 15, 2022)

(70) Caldeweyher, E.; Bannwarth, C.; Grimme, S. *J. Chem. Phys.* **2017**, *147*, 034112.
(71) Generally Applicable Atomic-Charge Dependent London Dispersion Correction (Version 3.4.0); https://github.com/dftd4/dftd4/releases/tag/v3.4.0 (accessed Nov 15, 2022)
(72) Rüger, R.; Yakovlev, A.; Philipsen, P.; Borini, S.; Melix, P.; Oliveira, A.; Franchini, M.; van Vuren, T.; Soini, T.; de Reus, M.; Ghorbani Asl, M.; Teodoro, T. Q.; McCormack, D.; Patchkovskii, S.; Heine, T. *AMS DFTB 2022.1, SCM, Theoretical Chemistry*; Vrije Universiteit: Amsterdam, https://www.scm.com/product/dftb/ (accessed Nov 15, 2022).
(73) Wahiduzzaman, M.; Oliveira, A. F.; Philipsen, P.; Zhechkov, L.; van Lenthe, E.; Witek, H. A.; Heine, T. *J. Chem. Theory Comput.* **2013**, *9*, 4006.
(74) Oliveira, A. F.; Philipsen, P.; Heine, T. *J. Chem. Theory Comput.* **2015**, *11*, 5209.
(75) Rüger, R.; Franchini, M.; Trnka, T.; Yakovlev, A.; van Lenthe, E.; Philipsen, P.; van Vuren, T.; Klumpers, B.; Soini, T. *AMS 2022.1, SCM, Theoretical Chemistry*; Vrije Universiteit: Amsterdam, http://www.scm.com (accessed Nov 15, 2022).
(76) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. *J. Am. Chem. Soc.* **1992**, *114*, 10024.
(77) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *J. Cheminf.* **2011**, *3*, 33.
(78) Open Babel: The Open Source Chemistry Toolbox (Version 2.4.0); https://github.com/openbabel/openbabel/releases/tag/openbabel-2-4-0 (accessed Nov 15, 2022)
(79) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490.
(80) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 520.
(81) Hassinen, T.; Peräkylä, M. *J. Comput. Chem.* **2001**, *22*, 1229.
(82) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157.
(83) Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, *36*, 3219.
(84) TURBOMOLE Version 7.5.1 (2021), A development of the University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989–2007, TURBOMOLE GmbH, since 2007; TURBOMOLE GmbH: Karlsruhe, **2021**; https://www.turbomole.org (accessed Nov 15, 2022)
(85) Grimme, S.; Hansen, A.; Ehlert, S.; Mewes, J.-M. *J. Chem. Phys.* **2021**, *154*, 064103.
(86) Sure, R.; Grimme, S. *J. Comput. Chem.* **2013**, *34*, 1672.
(87) Brandenburg, J. G.; Bannwarth, C.; Hansen, A.; Grimme, S. *J. Chem. Phys.* **2018**, *148*, 064104.
(88) Grimme, S.; Brandenburg, J. G.; Bannwarth, C.; Hansen, A. *J. Chem. Phys.* **2015**, *143*, 054107.
(89) Neese, F.; Wennmohs, F.; Becker, U.; Riplinger, C. *J. Chem. Phys.* **2020**, *152*, 224108.
(90) Neese, F. *ORCA – An ab initio, density functional and semiempirical program package, Version 5.0.1*; Max-Planck-Institut für Kohlenforschung: Germany, **2021**.
(91) Mardirossian, N.; Head-Gordon, M. *J. Chem. Phys.* **2015**, *142*, 074111.
(92) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.
(93) Rappoport, D.; Furche, F. *J. Chem. Phys.* **2010**, *133*, 134105.
(94) Epifanovsky, E.; Gilbert, A. T. B.; Feng, X.; Lee, J.; Mao, Y.; Mardirossian, N.; Pokhilko, P.; White, A. F.; Coons, M. P.; Dempwolff, A. L.; Gan, Z.; Hait, D.; Horn, P. R.; Jacobson, L. D.; Kaliman, I.; Kussmann, J.; Lange, A. W.; Lao, K. U.; Levine, D. S.; Liu, J.; McKenzie, S. C.; Morrison, A. F.; Nanda, K. D.; Plasser, F.; Rehn, D. R.; Vidal, M. L.; You, Z.-Q.; Zhu, Y.; Alam, B.; Albrecht, B. J.; Aldossary, A.; Alguire, E.; Andersen, J. H.; Athavale, V.; Barton, D.; Begam, K.; Behn, A.; Bellonzi, N.; Bernard, Y. A.; Berquist, E. J.; Burton, H. G. A.; Carreras, A.; Carter-Fenk, K.; Chakraborty, R.; Chien, A. D.; Closser, K. D.; Cofer-Shabica, V.; Dasgupta, S.; de Wergifosse, M.; Deng, J.; Diedenhofen, M.; Do, H.; Ehlert, S.; Fang, P.-T.; Fatehi, S.; Feng, Q.; Friedhoff, T.; Gayvert, J.; Ge, Q.; Gidofalvi, G.; Goldey, M.; Gomes, J.; González-Espinoza, C. E.; Gulania, S.; Gunina, A. O.; Hanson-Heine, M. W. D.; Harbach, P. H. P.; Hauser, A.; Herbst, M. F.; Vera, M. H.; Hodecker, M.; Holden, Z. C.; Houck, S.; Huang, X.; Hui, K.; Huynh, B. C.; Ivanov, M.; Jász, Á.; Ji, H.; Jiang, H.; Kaduk, B.; Kähler, S.; Khistyaev, K.; Kim, J.; Kis, G.; Klunzinger, P.; Koczor-Benda, Z.; Koh, J. H.; Kosenkov, D.; Koulias, L.; Kowalczyk, T.; Krauter, C. M.; Kue, K.; Kunitsa, A.; Kus, T.; Ladjánszki, I.; Landau, A.; Lawler, K. V.; Lefrancois, D.; Lehtola, S.; Li, R. R.; Li, Y.-P.; Liang, J.; Liebenthal, M.; Lin, H.-H.; Lin, Y.-S.; LiuF, ; Liu, K.-Y.; Loipersberger, M.; Luenser, A.; Manjanath, A.; Manohar, P.; Mansoor, E.; Manzer, S. F.; Mao, S.-P.; Marenich, A. V.; Markovich, T.; Mason, S.; Maurer, S. A.; McLaughlin, P. F.; Menger, M. F. S. J.; Mewes, J.-M.; Mewes, S. A.; Morgante, P.; Mullinax, J. W.; Oosterbaan, K. J.; Paran, G.; Paul, A. C.; Paul, S. K.; Pavošević, F.; Pei, Z.; Prager, S.; Proynov, E. I.; Rák, Á.; Ramos-Cordoba, E.; Rana, B.; Rask, A. E.; Rettig, A.; Richard, R. M.; Rob, F.; Rossomme, E.; Scheele, T.; Scheurer, M.; Schneider, M.; Sergueev, N.; Sharada, S. M.; Skomorowski, W.; Small, D. W.; Stein, C. J.; Su, Y.-C.; Sundstrom, E. J.; Tao, Z.; Thirman, J.; Tornai, G. J.; Tsuchimochi, T.; Tubman, N. M.; Veccham, S. P.; Vydrov, O.; Wenzel, J.; Witte, J.; Yamada, A.; Yao, K.; Yeganeh, S.; Yost, S. R.; Zech, A.; Zhang, I. Y.; Zhang, X.; Zhang, Y.; Zuev, D.; Aspuru-Guzik, A.; Bell, A. T.; Besley, N. A.; Bravaya, K. B.; Brooks, B. R.; Casanova, D.; Chai, J.-D.; Coriani, S.; Cramer, C. J.; Cserey, G.; DePrince, A. E. III.; DiStasio, R. A. Jr.; Dreuw, A.; Dunietz, B. D.; Furlani, T. R.; Goddard, W. A. III.; Hammes-Schiffer, S.; Head-Gordon, T.; Hehre, W. J.; Hsu, C.-P.; Jagau, T.-C.; Jung, Y.; Klamt, A.; Kong, J.; Lambrecht, D. S.; Liang, W.; Mayhall, N. J.; McCurdy, W.; Neaton, J. B.; Ochsenfeld, C.; Parkhill, J. A.; Peverati, R.; Rassolov, V. A.; Shao, Y.; Slipchenko, L. V.; Stauch, T.; Steele, R. P.; Subotnik, J. E.; Thom, A. J. W.; Tkatchenko, A.; Truhlar, D. G.; Van Voorhis, T.; Wesolowski, T. A..; Whaley, K B..; Woodcock, H. L. III.; Zimmerman, P. M.; Faraji, S.; Gill, P. M. W.; Head-Gordon, M.; Herbert, J. M.; Krylov, A. I. *J. Chem. Phys.* **2021**, *155*, 084801.
(95) Villot, C.; Ballesteros, F.; Wang, D.; Lao, K. U. *J. Phys. Chem. A* **2022**, *126*, 4326.
(96) Mardirossian, N.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2014**, *16*, 9904.
(97) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
(98) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
(99) Spicher, S.; Caldeweyher, E.; Hansen, A.; Grimme, S. *Phys. Chem. Chem. Phys.* **2021**, *23*, 11635.
(100) Al-Hamdani, Y. S.; Nagy, P. R.; Zen, A.; Barton, D.; Kállay, M.; Brandenburg, J. G.; Tkatchenko, A. *Nat. Commun.* **2021**, *12*, 3927.
(101) Řezáč, J.; Riley, K. E.; Hobza, P. *J. Chem. Theory Comput.* **2011**, *7*, 2427.
(102) Miriyala, V. M.; Řezáč, J. *J. Phys. Chem. A* **2018**, *122*, 2801.
(103) Řezáč, J. *J. Chem. Theory Comput.* **2020**, *16*, 2355.
(104) Teale, A.; Helgaker, T.; Savin, A.; Adamo, C.; Aradi, B.; Arbuznikov, A.; Ayers, P.; Baerends, E. J.; Barone, V.; Calaminici, P.; Cancès, E.; Carter, E. A.; Chattaraj, P. K.; Chermette, H.; Ciofini, I.; Crawford, T. D.; De Proft, F.; Dobson, J. F.; Draxl, C.; Frauenheim, T.; Fromager, E.; Fuentealba, P.; Gagliardi, L.; Galli, G.; Gao, J.; Geerlings, P.; Gidopoulos, N.; Gill, P. M. W.; Gori-Giorgi, P.; Görling, A.; Gould, T.; Grimme, S.; Gritsenko, O.; Jensen, H. J. A.; Johnson, E. R.; Jones, R. O.; Kaupp, M.; Köster, A. M.; Kronik, L.; Krylov, A. I.; Kvaal, S.; Laestadius, A.; Levy, M.; Lewin, M.; Liu, S.; Loos, P.; Maitra, N. T.; Neese, F.; Perdew, J. P.;

Pernal, K.; Pernot, P.; Piecuch, P. E..; Rebolini, E.; Reining, L.; Romaniello, P.; Ruzsinszky, A.; Salahub, D. R.; Scheffler, M.; Schwerdtfeger, P.; Staroverov, V. N.; Sun, J.; Tellgren, E.; Tozer, D. J.; Trickey, S. B.; Ullrich, C. A.; Vela, A.; Vignale, G.; Wesolowski, T. A.; Xu, X.; Yang, W. *Phys. Chem. Phys. Chem.* **2022**, *24*, Advance Article.

(105) Bursch, M.; Caldeweyher, E.; Hansen, A.; Neugebauer, H.; Ehlert, S.; Grimme, S. *Acc. Chem. Res.* **2019**, *52*, 258.

(106) Grimme, S. *Chem. Eur. J.* **2012**, *18*, 9955.

(107) Thiel, W. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2014**, *4*, 145.

(108) Ehlert, S.; Grimme, S.; Hansen, A. *J. Phys. Chem. A* **2022**, *126*, 3521.

(109) Hostaš, J.; Řezáč, J.; Hobza, P. *Chem. Phys. Lett.* **2013**, *568-569*, 161.

(110) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787.

(111) Tsuzuki, S.; Uchimaru, T. *Phys. Chem. Chem. Phys.* **2020**, *22*, 22508.

(112) Kolář, M.; Hobza, P. *J. Chem. Theory Comput.* **2012**, *8*, 1325.

APPENDIX B

Appendix: Reliable Prediction of Association (Free) Energies of Supramolecular Complexes with Heavy Main Group Elements – the HS13L Benchmark Set

Johannes Gorges,[†] Stefan Grimme[†] Andreas Hansen[†]

Received: 31 August 2022
Published online: 18 November 2022

Reproduced in Appendix B from Ref. J. Gorges, S. Grimme, and A. Hansen, *Reliable prediction of association (free) energies of supramolecular complexes with heavy main group elements - the HS13L benchmark set*, Phys. Chem. Chem. Phys. **24**.47 (2022) 28831, DOI: 10.1039/d2cp04049b with permission from the Royal Society of Chemistry.

Own contributions

- Compiling the HS13L benchmark set
- Performing all density functional theory, semiempirical quantum mechanical, and force-field calculations
- Interpretation of the results
- Writing the manuscript

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

PCCP

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d2cp04049b

Reliable prediction of association (free) energies of supramolecular complexes with heavy main group elements – the HS13L benchmark set†

Johannes Gorges, Stefan Grimme and Andreas Hansen *

We introduce a set of 13 supramolecular complexes featuring diverse non-covalent interactions with heavy main group elements (Zn, As, Se, Te, Br, I), high charges (−2 up to +4), and large systems with up to 266 atoms (HS13L). The experimental Gibbs free energies of association cover the typical range (−1.9 to −9.2 kcal mol^{-1}). An efficient automated multilevel theoretical workflow is applied for the determination of the respective minimum structures in solution by conformer ensemble generation with the ***CREST*** program at the semiempirical GFN2-xTB level. Subsequent refinement is performed with the r^{2}SCAN-3c composite DFT method including thermostatistical corrections at the GFN2-xTB level and solvation contributions by COSMO-RS using the ***CENSO*** free energy ranking algorithm. Various density functional approximations in combination with three London dispersion correction schemes are assessed against "back-corrected" experimental association energies as well as accurate local coupled cluster reference values. Our protocol predicts association free energies with a mean absolute deviation of only 2 kcal mol^{-1} from the measured values. Thus, it is well suited to generate reference association free energies for assessing theoretical methods on realistically sized supramolecular complexes or to support experimental chemists. For specifically evaluating methods for calculating gas-phase association energies, we recommend using the provided accurate coupled cluster reference values. We propose to use this set as an extension of the S30L benchmark set [Sure *et al.*, *J. Chem. Theory Comput.*, 2015, **11**, 3785–3801] with a special focus on the challenging computation of non-covalent interactions of heavy main group elements.

Received 31st August 2022,
Accepted 18th November 2022

DOI: 10.1039/d2cp04049b

rsc.li/pccp

1 Introduction

Non-covalently bound host–guest complexes represent an important research field with many practical applications.[1] They are used as reaction containers, for molecular recognition, in template-directed synthesis, biomimetics, and self-assembly.[2–7] Due to their unique coordination preferences and electronic properties, heavy main group elements are of special interest to prepare novel structures with new and interesting supramolecular properties.[8] Their characteristic interactions, such as halogen bonding, chalcogen bonding, pnictogen bonding, and tetrel bonding are valuable in many areas of chemistry.[8]

For many applications, the stability of supramolecular complexes is decisive, which is directly linked to the Gibbs free energy of association. Thus, it is important to obtain accurate experimental values for this quantity. Among various experimental techniques, Isothermal Titration Calorimetry (ITC) stands out as the most universal one[9] and is the method of choice for measuring experimental binding thermodynamics of ligand binding.[10] However, the interpretation of the measured energies in terms of specific molecular processes can be difficult if, for example, the stoichiometry of the individual components in the formed complex is not clear.[9] Here, a reliable computational protocol is useful to reproduce experimental data for the assumed association mechanism or to predict alternative ones.

The calculation of free energies for the formation of larger supramolecular complexes still poses a challenge to computational chemistry. Since most experimentally synthesized complexes consist of about 100 atoms or more, the computational costs of highly accurate quantum chemical methods are often

Mulliken Center for Theoretical Chemistry, Clausius-Institute for Physical and Theoretical Chemistry, University of Bonn, Beringstr. 4, 53115 Bonn, Germany. E-mail: hansen@thch.uni-bonn.de

† Electronic supplementary information (ESI) available: Calculated relative energy contributions for all employed methods (HS13Lenergies.xlsx) as well as optimized geometries for the HS13L and HS13L-CI benchmark sets (HS13L.zip) together with the coupled cluster reference association energies are provided. Additional statistical evaluations are provided in the file SI.pdf. See DOI: **https://doi.org/10.1039/d2cp04049b**

too large for treating these systems. Furthermore, solvation and entropic effects have to be considered in the binding process, which are generally difficult to predict accurately by computational methods.[11] One successful approach was proposed by one of us, in which different *ab initio* and semiempirical quantum chemical (SQM) methods were combined. For the so-called S12L benchmark set, this procedure yields accurate association free energies with an average deviation from the experimental values of only 2 kcal mol^{-1}.[12] The S12L Benchmark set was later extended to 30 complexes (S30L benchmark set), which already covers a broad spectrum of different non-covalent interactions, such as London dispersion (LD), π–π stacking, ion-π interactions, or hydrogen and halogen bonding.[13] However, S30L includes heavy main group elements only to a small extent. Benchmark sets focusing on characteristic main group non-covalent interactions (NCI), as the CHAL336[14] for chalcogen bonding, or the ATLAS bechmark sets by Řezác for hydrogen bonding (HB300SPXx10),[15] σ-hole interactions (SH250x10),[16] and LD interactions (D1200 and D442x10)[17] contain only small systems, for which canonical coupled cluster reference values can be computed. To the best of our knowledge, no comparable benchmark studies for large supramolecular complexes with heavy elements as significant component exist. Since reliable reference data for such systems are also important, especially for the development of new efficient computational methods, this issue is addressed here with a new benchmark set.

To emphasize the focus on heavy elements it is named ''heavy S13L'' (HS13L) and covers elements of groups 12 to 17. To demonstrate the accuracy of our approach in direct comparison to experimentally accessible reference values, complexes with available experimental association free energies were selected. Systems with small HOMO–LUMO gaps were consciously not included in HS13L. In metallic-like systems, large electron density fluctuations occur (so-called ''type C non-additivity''), which lead to a slower decay of the LD interaction than described by the additive pair-wise approach and thus require special treatment.[18] As NCI complexes generally feature many possible binding sites depending on the number and nature of the respective functional groups, it is essential for reliable modeling to determine the most favorable conformer. Therefore, we applied the ***CREST***[19] and ***CENSO*** workflow[20] to screen the large conformational space of the investigated complexes with SQM methods and determine the minimum structure with subsequent refinement at the DFT level of theory. For an accurate calculation of binding thermodynamics of non-rigid systems in solution, it is necessary to consider conformers.[20] In this work, we benchmark this workflow for the first time systematically on large supramolecular complexes.

We aim to provide a reliable protocol without any empirical adjustments to predict or validate experimental association free energies of supramolecular complexes including heavy main group elements. First, a short overview of the underlying theory for our approach for the calculation of association free energies in solution is given. After a description of the test set, the computational details are given. Further, we present and discuss the results for the benchmark set. Computer timings are compared for the most accurate methods and evaluated with respect to their cost-accuracy ratio. Finally, we draw general conclusions concerning the proposed workflow and give method recommendations for the computation of association free energies of realistic, experimentally observable, supramolecular complexes.

2 Theory

The association free energy in solution is calculated by

$$\Delta G_{\mathrm{a}} = \Delta E + \Delta G^{T}_{\mathrm{mRRHO}} + \Delta\delta G^{T}_{\mathrm{solv}}(X), \quad (1)$$

where ΔE is the gas-phase association energy, $\Delta G^{T}_{\mathrm{mRRHO}}$ the thermostatistical corrections to the free energy, and the solvation free energy in solvent X, both at temperature T. In the supermolecular approach, ΔE is calculated as

$$\Delta E = E(\mathrm{complex}) - E(\mathrm{host}) - E(\mathrm{guest}),$$

where E is the gas-phase electronic energy of the respective species. Hence, the so-called ''relaxation energy'' upon complexation is included. The missing LD contribution to the electronic energy in the framework of density functional theory (DFT) is computed by the semi-classical DFT-D3[21,22] method with Becke–Johnson (BJ) damping[23,24] and its successor DFT-D4[25] including charge-dependent polarizabilities. For both, beyond the pair-wise contributions $\Delta E^{(2)}_{\mathrm{disp}}$ also the three-body Axilrod–Teller–Muto (ATM)[26,27] term is applied consistently, which is especially important for large systems:[12]

$$\Delta E = \Delta E^{\mathrm{DFT}}_{\mathrm{el}} + \Delta E^{(2)}_{\mathrm{disp}} + \Delta E^{(\mathrm{ATM})}_{\mathrm{disp}}. \quad (2)$$

Alternatively, the exchange-hole dipole moment (XDM)[28,29] approach, the many-body dispersion model,[30,31] and the non-local VV10 correction,[32] also called DFT-NL,[33] could be applied to compute the LD contribution. For comparison, we also assess the latter model here, which includes only pairwise contributions. We employ the different dispersion models in combination with various Kohn–Sham density functional approximations (DFA).

For the thermostatistical corrections to the free energy, the modified rigid-rotor-harmonic-oscillator (mRRHO) approach[12] is employed here. This approach treats vibrational modes below 50 cm^{-1}, which are notoriously problematic in the harmonic approximation in entropy calculations, as hindered rotations with smooth interpolation to the standard harmonic approach. Due to the large system size in the HS13L, vibrational frequencies are calculated with SQM or force-field (FF) methods. Since the minimum geometry at the SQM level may be distorted with respect to the DFT-optimized structure, the single-point hessian (SPH)[34] approach is applied here to compute the mRRHO contribution effectively on DFT geometries. The so-called ''conformational'' entropy[35] is not computed explicitly here, as we do not expect its contribution to the free energy to be significant. Most systems in the HS13L are relatively rigid and do not lose significant conformational freedom upon complexation considering other sources of errors

in the calculation of the individual contributions to the free energy. Furthermore, the computational costs for the systems' size in the HS13L for the conformational entropy become unfeasible for larger systems (above 100 atoms) even when using force-field methods in combination with implicit solvation models.[36]

We calculate the solvation free energy with the continuum solvation models COSMO-RS[37,38] and SMD.[39] The alternative explicit consideration of solvent molecules, *e.g.*, in our recent so-called quantum cluster growth (QCG) model[40] would be too computationally too demanding for this system size. Other physical aspects, such as the pH value and the ionic strength of the reaction solution can be relevant in special cases.[41] However, for the complexes included in the HS13L, these effects are expected to be smaller than 0.5 kcal mol^{-1} and therefore relatively small compared to other sources of errors. These effects are neglected in our approach to retain a fully-automated and generally applicable workflow. For a more throughout discussion of the mentioned and other less important factors contributing to binding free energies, we refer to ref. 41.

3 Description of the HS13L test set

In the following, we provide a short description of the investigated complexes. Fig. 1 depicts the optimized geometries of all complexes included in the HS13L. In Table 1 all complexes are given with their number in this set, their name, their charge, and the experimental conditions at which the association free energy was determined. For an easier interpretation of the results, complexes are sorted according to the most prominent type of interaction and charged complexes are grouped, as it is recommended for NCI complexes by Řezác and Hobza.[42]

Complex **1** comprises the guest diiodine and the host cucurbit[6]uril (CB[6]). In the crystal structure, halogen bonding was observed,[43] whereas in (implicit) aqueous solution this interaction is quenched according to the optimized geometry and was therefore classified as mainly bound by LD. CB[6] is a representative of the cucurbit[*n*]urils which are important excipients in medical formulations for improving drug delivery.[55] **2** is a complex of 1,2,4,5-tetracyanobenzene (TCB) bound *via* π–π stacking to a macrocyclic boronic ester connecting two tellurophenes, which exhibit advantageous optoelectronic properties[44] due to the "heavy-atom effect" of tellurium[56] The host of complex **3** is a Zn(II) complex of 2,6-bis(porphyrin)-substituted 3,5-dimethylpyrazine bound to the fullerene C_{70}. The binding motifs are LDs associated with π–π interactions between the electron-rich porphyrin nitrogen atoms and C_{70}.[45]

Complex **4** is the largest in HS13L with 266 atoms. Two conformers have to be considered as the guest iodocyclohexane

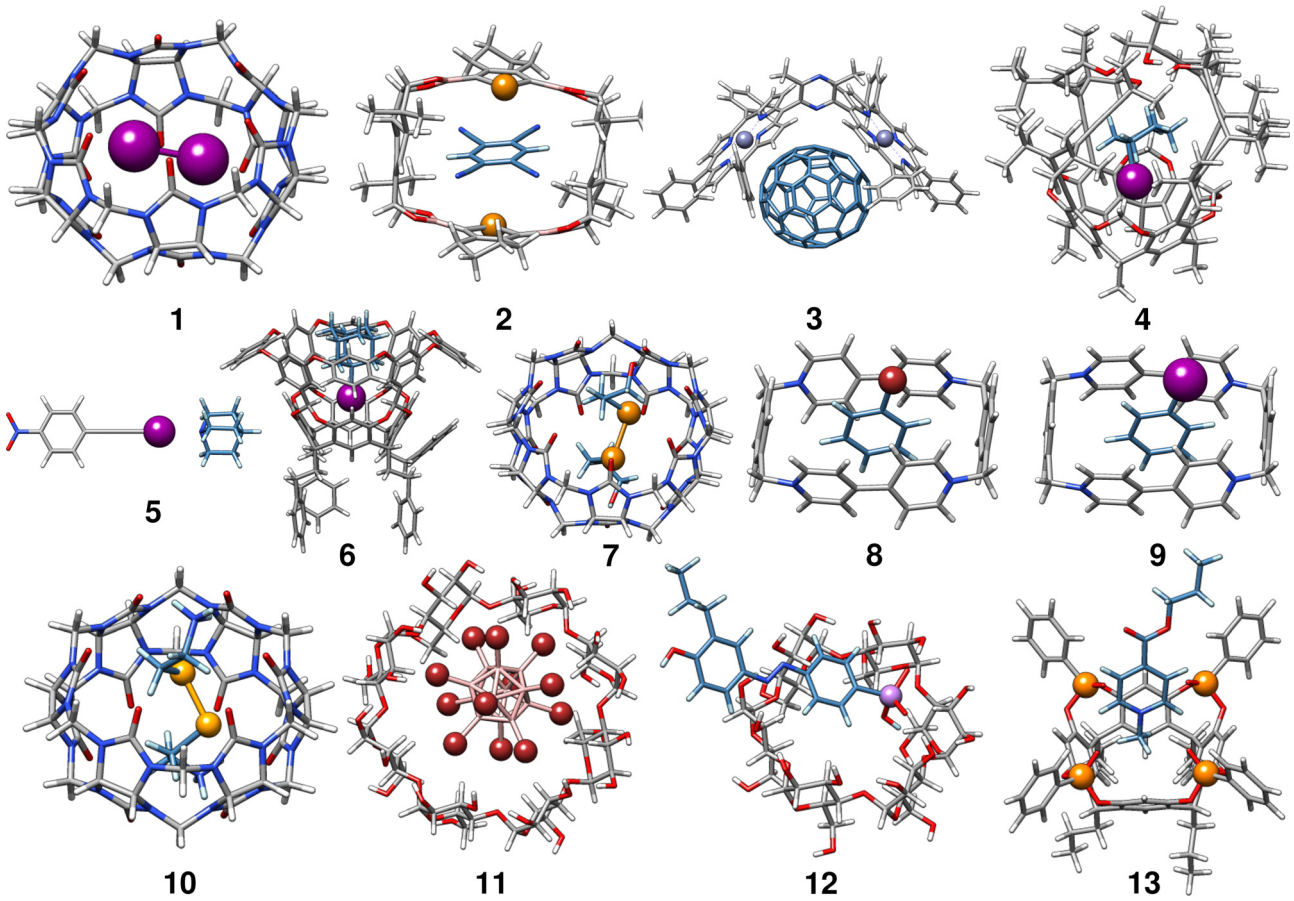


Fig. 1 r^2SCAN-3c[SMD] optimized geometries of the 13 host–guest complexes of the HS13L benchmark set. For better visibility, the C and H atoms of the guest molecules are colored in blue and light blue, respectively.

Table 1 Complexes in the HS13L set with charge and free energies of association ΔG_{exp} in kcal mol^{-1} measured at the given temperature T in the respective solvent

Entry	Complex	Charge	Solvent	T	ΔG_{exp}
1	I_2@CB[6][43]	0	H_2O	298	−8.2
2	tcb@tellurophene[44]	0	$CHCl_3$	298	−5.8
3	C_{70}@bisZnporphy[45]	0	Toluene	296	−4.8
4	Icy@$(P)_4$-AAC[46]	0	*n*-Octane	293	−5.7
5	Iethynyl@quinucli[47]	0	Benzene	298	−1.9
6	Iad@cav[48]	0	DMSO	298	−6.8
7	$(HOC3Te)_2$@CB[7][49]	0	H_2O	298	−4.7
8	Brbenz@CBPQT^{4+}[50]	4	H_2O	303	−5.0
9	Ibenz@CBPQT^{4+}[50]	4	H_2O	303	−5.5
10	SeCy@CB[6][51]	2	H_2O	298	−9.2
11	$B_{12}Br_{12}^{2-}$@γ-CD[52]	−2	H_2O	298	−8.1
12	Asdiazo@α-CD[53]	−1	H_2O	298	−5.2
13	$C_{10}H_{14}O_2$ N@Tiiii[54]	1	DCE	303	−7.6

can exist in an axial and equatorial conformation, both separately and in the complex. For the isolated guest, the equatorial conformation is the most stable, whereas in the complex the axial conformation is preferred. The binding motif with the host, an enantiopure alleno-acetylenic cage (AAC) with a resorcin[4]arene scaffold, is dominated by dispersion interaction and halogen-bonding.[46] With 37 atoms, complex **5** is the smallest one in HS13L. Its binding motif is a single halogen bond between the bond donor iodoethynylbenzene and the bond acceptor quinuclidine. **6** contains a deep cavity host providing four hydrogen atoms pointing into the cavity. It forms unusual hydrogen bonds (C–H···X–R) to the guest 1-iodoadamantane.[48] **7** is a ditelluride $(HOC3Te)_2$ guest interacting *via* chalcogen bonding to CB[7].[49] Complexes **8** to **13** are charged systems. **8** and **9** both have the +4 charged cyclobis-(paraquat-*p*-phenylene) host, which is the highest charge considered. This host can form stable complexes through steric complementarity and (assumed) charge transfer mechanisms with the volatile substances bromobenzene and iodobenzene.[50] With a value of −9.2 kcal mol^{-1} for the association free energy, **10** forms the strongest NCI bonds in HS13L. Its bonding motif consists mainly of chalcogen bonds between the selenium atoms of the guest selenocystamine and the carbonyl oxygens of the host CB[6].[51] The formation of the double negatively charged dodecaborate boron cluster with γ-cyclodextrin (**11**) is driven by the so-called chaotropic effect.[52] Complexes of cyclodextrin are often used for drug delivery and are therefore of great practical importance.[57] **12** consists of a diazonium compound with an arsenate group which is bound to α-cyclodextrin[53] mainly by hydrogen bonding. Last is a tetraphosphonate cavitand, which binds a methylpyridinium cation *via* cation-dipole and CH_3–π interactions.[54]

In summary, despite the limited number of systems, this benchmark set exhibits a broad range of different non-covalent binding motifs of (heavy) main group elements, such as hydrogen bonding, chalcogen bonding, halogen bonding, π–π stacking, and dispersion interaction. It involves polar as well as non-polar solvents, which are tabulated in Table 1. Furthermore, the complexes are of realistic size (37 up to 266 atoms) and contain a large variety of unusual main group elements (Se, Te, P, As, and Zn).

4 Computational details

Conformer search for the host, guest, and complex structures was performed with ***CREST***[19] Version 2.11[58] at the GFN2-xTB[59] level with the implicit solvation model ALPB[60] in ***xTB*** version 6.4.0.[61] To save computation time, we used GFN-FF[62] [ALPB] for larger systems (**2**, **4**, **6**, **7**, **10**, and **11**) instead if the optimized structures appeared to be reasonable and showed no deformations compared to the GFN2-xTB geometries. Likewise, we employed the special NCI mode of ***CREST*** for some of the larger complexes (**2**, **3**,**6**,**7**, and **11**) to reduce the number of generated conformers. For rigid systems (**1**, **12**, **13**), we utilized the rigid docking mode of the intramolecular force-field xtb-IFF.[63]

Subsequent refinement of the conformer ensembles generated as described above was performed with ***CENSO***[64] version 1.2.0[65] using the default thresholds for *Part0* to *Part2*, as described in ref. 64 and, with stronger focus on NCI complexes, in ref. 66. First, conformers that are more than 4 kcal mol^{-1} higher in energy than the lowest conformer at the B97-D3(0)[67]/def2-SV(P)[68]+gCP[69] level of theory were excluded. Solvation effects in *Part0* are captured by ALPB(GFN2-xTB). In *Part1*, we removed conformers above the free energy threshold of 3.5 kcal mol^{-1} calculated on geometries optimized on the GFN2-xTB[ALPB] level of theory. We performed the free energy ranking in this part already with the r^{2}SCAN-3c composite DFT method including thermostatistical corrections at the GFN2-xTB level in the single-point hessian (SPH)[70] approach and solvation contributions with COSMO-RS(16) normal (based on the r^{2}SCAN-3c electron density), which we will refer to in the following as default level. In the last part, we conduct the final geometry optimization at the r^{2}SCAN-3c[71] level using the SMD[39] continuum solvation model. The geometry optimization was performed with ***ORCA*** 5.0.3[72] employing *DefGrid2*. Using the default settings in ***CENSO***, *i.e.* DCOSMO-RS,[73] led to severe convergence problems, especially for the larger complexes, and was therefore not applied for this automated workflow. For some of the charged complexes (**10**, **12**, and **13**), large deviations from the experimental values were observed. In order to diminish the electrostatic contribution in the solvation free energy and the electronic energies, counterions were added to the charged complexes **10–13** and the resulting set is called HS13L-CI. Chloride counterions were added to the cations and sodium ions to the anions for neutralization with the docking algorithm of the intramolecular force-field xTB-IFF and the lowest found structure was re-optimized at the r^{2}SCAN-3c[SMD] level. We denote these structures with counterions by adding "_CI" to their respective name or number. For the charged complexes **8** and **9**, this was not done, as the experimental value was already well reproduced with the standard procedure and the addition of four counterions resulted in massive convergence problems in the SCF iterations, which would be problematic for a benchmark set.

Calculating the ensemble average by Boltzmann weighting all found conformers is not important here compared to other sources of errors and the association free energy can accurately be described by only one distinct minimum structure. Therefore, we investigated the effect of considering also higher conformers only for the default level of theory and applied the other computational methods only for the minimum conformer.

All single-point calculations were performed with ***ORCA*** 5.0.3 using the *DefGrid3*, the *TightSCF* convergence criteria, and the RIJCOSX[74,75] approximation. Ahlrichs def2-QZVP[68] basis set with corresponding default ECPs and auxiliary basis sets[76] were employed. For the double hybrids, we applied the frozen core approximation and the def2-QZVPP correlation auxiliary basis sets[77,78] in the RI-MP2 part. The nonlocal VV10 correction was computed non-self-consistently. The D3[21,22] and D4[25] correction were consistently applied with inclusion of the ATM term using the s-dftd3[79] dft-d4[80] standalone programs. Becke-Johnson damping[81] was applied for all DFAs except for the Minnesota functionals M06L and M06-2x, for which the zero-damping variant was used. Table 2 lists all tested DFAs and dispersion correction combinations. The DFT-C[82] basis set superposition error (BSSE) correction was computed for B97M-V/def2-SVPD[83] with the ***Q-Chem*** 5.4 program package.[84] In the following, this combination of small basis set and basis set correction is denoted with a ''-C''.

Harmonic frequencies for the thermostatistical contributions were calculated on the minimum structure of the respective method. The rotational symmetry numbers of the complexes were obtained with a *DESY* threshold of 0.1 in ***TURBOMOLE***(V. 7.5.1).[108] and used for the calculation of the rotational entropy, see ESI† (Table S7). For the geometry optimization as well as the frequency calculation, an implicit solvation model was consistently applied. GFN2-xTB, GFN1-xTB, and GFN-FF frequencies were calculated with ***xTB*** and the ALPB implicit solvation model. PM6-D3H4X and PM7 frequencies were computed with the COSMO[109] solvation model with ***MOPAC2016*** (version 19.179L)[110] using ***xtb*** as driver. Gas-phase single-point calculations with both PM methods were also conducted with the same program combination.

The solvation free energy was calculated with ***COSMOtherm19***.[111,112] The default procedure of one single-point calculation in gas-phase and one in continuum solution was performed using r^{2}SCAN-3c with the *m4* grid in ***TURBOMOLE*** (V. 7.5.1).[108] Additionally, we calculated solvation free energies at the default level of theory (BP86[113,114]/def2-TZVP) and BP86[113,114]/def2-TZVPD, respectively, for the fine parametrization as the parameters were fitted for this level. The respective solvation free energies with SMD and CPCM[115] were calculated with ***ORCA*** applying the same procedure.

Furthermore, we generated local coupled-cluster reference association energies for the HS13L and the HS13L-CI set. Due to the large size of the NCI complexes composed in these sets, the ''gold-standard'' reference level CCSD(T) at the approximate basis set limit[116] is computationally impossible without introducing further approximations. Specifically, we applied the domain based, local pair natural orbital coupled cluster method[117] in its ***ORCA*** 5.0.2[72] closed-shell, sparse maps non-iterative[118] or iterative triples[119] implementation (DLPNO-CCSD(T) and DLPNO-CCSD(T1), respectively) together with default *TightPNO* or special *VeryTightPNO* threshold settings (*i.e.*, *TCutMKN*, *TCutPNO*, and *TCutPairs* tightened to 10^{-4}, 10^{-8}, and 10^{-6}, respectively). We employed ***ORCA*** 5.0.2 *TightSCF* convergence criteria and default frozen core settings as well as Ahlrich's-type basis sets of different sizes (def2-SVP, def2-TZVPP, def2-QZVPP) together with the corresponding auxiliary basis sets. We used a specially developed correction scheme to minimize the local truncation errors and focal-point analysis[120,121] to reduce the BSSE and basis set incompleteness (BSIE) errors (see Section 5.5.1 for details). These so-called ''DLPNO-CCSD(T1)/CBS'' reference level was computationally unfeasible for complex **3**. Therefore, the respective PWPB95-D4/def2-QZVP association energy is used as reference values instead. Based on our experience,[122] this double-hybrid represents a very good approximation for coupled cluster association energy in the gas phase (see discussion below).

Table 2 Overview of all DFAs and applied dispersion corrections assessed in this work. The D3 and D4 correction consistently include the three-body ATM term

Functional	D3(BJ)	D4	VV10/NL	Ref.
GGA				
PBE	x	x	x	85
RPBE	x	x	x	86
***meta*-GGA**				
r^{2}SCAN	x	x	x	87
B97M-V	x	x	x	88–90
M06L[a]	x	x	x	91 and 92
Hybrid				
PW6B95	x	x	x	93
PBE0	x	x	x	94
B3LYP	x	x	x	95 and 96
M06-2X[a]	x	x	x	97
RS-hybrid				
ωB97M-V	x	x	x	89, 90 and 98
ωB97X-V[b]	x	x	x	89, 99 and 100
Double-hybrid				
PWPB95	x	x	x	101 and 102
revDSD-PBEP86[c]		x		103 and 104
Composite (''3c'')				
B97-3c	x			105
r^{2}SCAN-3c		x		106
PBEh-3c	x			107

[a] Zero-damping was used in the dispersion correction instead of BJ-damping. [b] Revised D4 parameter taken from ref. 99 were employed (see ESI for details). [c] 2019 parametrization with unscaled ATM term[103] as well as with downscaled ATM term ($s9 = 0.5132$)[104] was employed.

5 Results and discussion

In this section, the performance of all tested methods is presented and discussed with respect to the experimental association free energies. We compare the calculated association energies of the discussed methods by ''back-correcting'' the experimental values, *i.e.*, subtracting the respective two

remaining contributions to the free energy with the default level of theory: r^{2}SCAN-3c energies, GFN2-xTB[ALPB]-SPH thermostatistical contributions, and COSMO-RS(16)-normal-(r^{2}SCAN-3c) solvation free energies. For example, "back-corrected" experimental gas-phase association energies are obtained as $\Delta E_{exp} = \Delta G_{exp} - \Delta G^{T}_{mRRHO}$(GFN2-xTB[ALPB]-SPH) $- \Delta\delta G^{T}_{solv}$ (X)(COSMO-RS(16)-normal(r^{2}SCAN-3c)) computed at the respective temperature T in solvent X. We calculated the individual contributions to the free energy for the structures of the HS13L and HS13L-CI only for the lowest conformer. Not taking higher-lying conformers into account results in a maximum error of about 0.7 kcal mol^{-1} and on average 0.2 kcal mol^{-1} for the HS13L (see ESI†). Details for the statistical measures used, namely mean deviation (MD), mean absolute deviation (MAD) and standard deviation (SD), are given in the ESI.† We conduct the statistical evaluation for the HS13L-CI set, *i.e.*, with counterions, as the deviations to the experimental values are generally smaller than for the HS13L set without counterions (see Section 5.2) and the respective statistics for HS13L can be found in the ESI.† In Section 5.1 the effect of using different methods for calculating the contributions to the free energies are investigated for the HS13L-CI. The use of different solvation models and the addition of counterions is discussed in more detail in Section 5.2. After an evaluation of computational timings (Section 5.3) we discuss the individual contributions to the free energy for each complex in Section 5.4. Furthermore, we evaluate the performance of the DFAs used for the calculation of the gas-phase association energies with respect to the "DLPNO-CCSD(T1)/CBS" reference values in Section 5.5. The computed energies of all assessed methods as well as the conformer ensembles are provided in the ESI.† All DFAs were applied with the quadruple-ζ size basis set def2-QZVP, which is usually sufficient to ensure a diminishing basis set superposition error (BSSE).[123] Only for the double hybrids, a significant BSSE of up to 1 kcal mol^{-1} is expected (for results of the respective counterpoise calculations for the example of **1** see Table S4 in the ESI†).

5.1 Finding the best workflow for the calculation of free energies

The resulting change of the calculated association free energy upon using other methods than the default theory level (r^{2}SCAN-3c + GFN2-xTB-SPH + COSMO-RS(16)-normal(r^{2}SCAN-3c)) is shown in Fig. 2 for HS13L-CI. Statistical measures are discussed in comparison to the experimental association free energies, whereby only the discussed method is varied and evaluated in combination with the two other components to the association free energy computed at the default level of theory. The default level of theory has a mean absolute deviation (MAD) of only 2.0 kcal mol^{-1} which is remarkable considering the

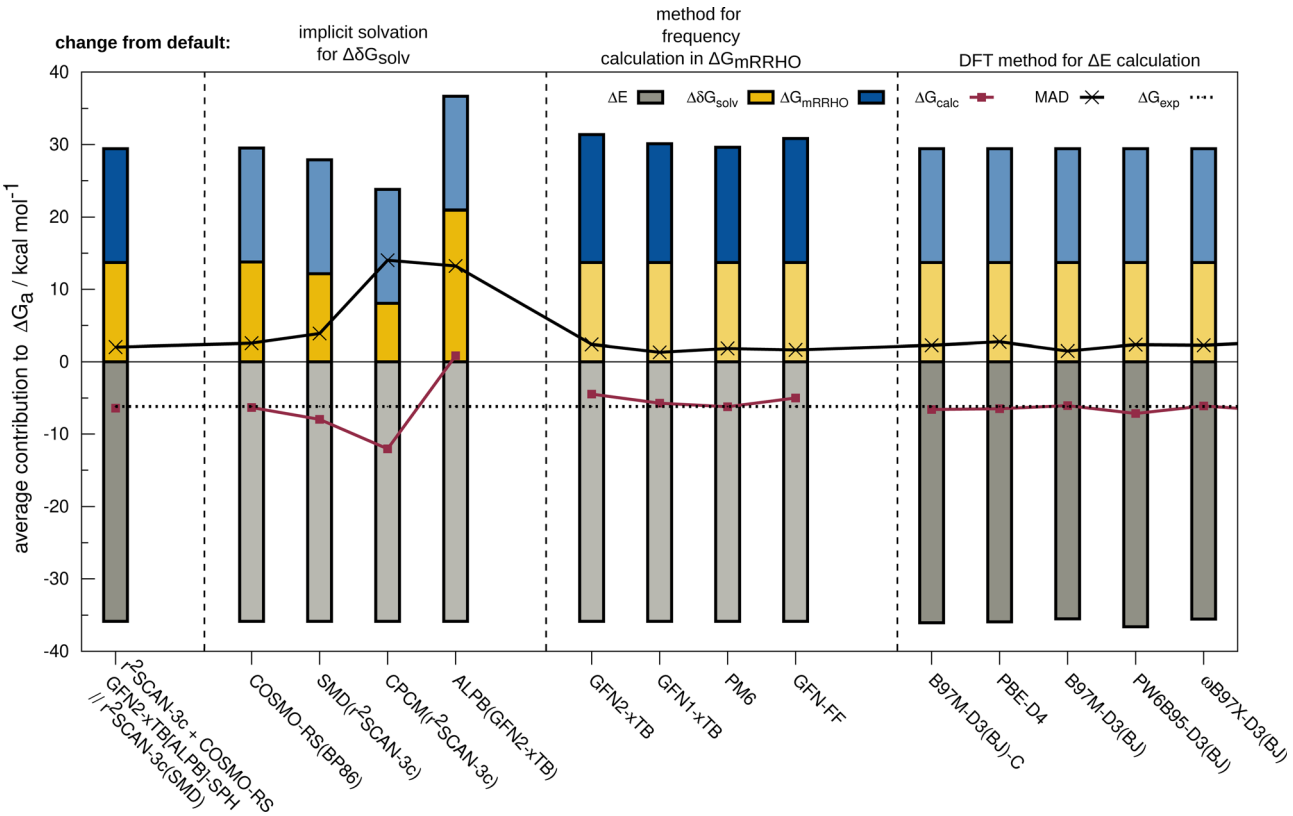


Fig. 2 Contributions to the calculated ΔG_{calc} averaged over all complexes of the HS13L-CI set. The leftmost bar illustrates the default level of theory used in this work, while the others illustrate the effect of using a different model or level of theory for the calculation of the solvation contribution, the thermostatistical correction, and the electronic energy. Contributions that are not affected by these variations are depicted in brighter colors. The MAD to the experimental association free energies is also given. COSMO-RS refers to the normal parametrization of 2016.

complexity of the considered property and systems. For the solvation free energy, COSMO-RS 16 with normal parameters yields very similar results with the default BP86/def-TZVP density compared to the r^{2}SCAN-3c density.

The deviations when using SMD as solvation model are larger with a MAD of 3.9 kcal mol^{-1} for HS13L-CI. As expected, the purely electrostatic CPCM solvation model (14 kcal mol^{-1} MAD) and the semiempirical ALPB(GFN2-xTB) model (13.2 kcal mol^{-1}) show the largest deviations from the assessed solvent models. We investigated the effect of using different solvent models in the geometry optimization on the overall association free energy for the example of complex **8** (see ESI,† for details). SMD and DCOSMO-RS both yield very good geometries for this complex with essentially no deviation for the free energy, validating the choice for the technically more robust SMD model in our workflow. For the G_{mRRHO} contribution, the MAD increases from 2.0 to 2.4 kcal mol^{-1} when calculating the GFN2-xTB[ALPB] frequencies for the relaxed geometries instead of using the SPH approach. Notably, the MAD is smaller when using GFN1-xTB[ALPB] (1.3 kcal mol^{-1}), PM6[COSMO] (1.8 kcal mol^{-1}), and GFN-FF[ALPB] (1.6 kcal mol^{-1}) indicating some error cancellation between the individual contributions. The comparison to r^{2}SCAN-3c frequencies for a subset composed of the five smallest complexes of HS13L (see ESI†) shows that GFN1-xTB[ALPB] (0.6 kcal mol^{-1} MAD) and GFN2-xTB[ALPB]-SPH (0.8 kcal mol^{-1} MAD) give the best results, whereas with PM6[COSMO] (1.1 kcal mol^{-1}), PM7[COSMO] (1.6 kcal mol^{-1}) and GFN-FF[ALPB](1.5 kcal mol^{-1}) the deviations are slightly larger. Since the differences in G_{mRRHO} are for all methods tested small compared to the errors in ΔE and G_{solv}, we tentatively conclude that the thermostatistical contribution is not the largest source of error in the workflow. Using a higher level of theory for the computation of frequencies, *e.g.*, DFT, is therefore in most cases not worth the computational costs. For the electronic energy, the best DFAs in each class of functionals and the BSSE corrected B97M-D3(BJ)/def2-SVPD DFA (denoted with B97M-D3(BJ)-C) are shown. Only the *meta*-GGA B97M-D3(BJ) with a MAD of 1.5 kcal mol^{-1} yields a better MAD than r^{2}SCAN-3c making it the overall best DFA on HS13L-CI with respect to the "back-corrected" experimental values.

However, as the error of the "back-corrected" experimental values is difficult to estimate, we cannot clearly say here which of both methods is better considering that only 13 systems are statistically evaluated. Even the double-hybrid PWPB95-D3(BJ) shows slightly larger deviations than r^{2}SCAN-3c demonstrating the excellent accuracy of this efficient composite method on this set and validating its use as the default.

5.2 Free energy of solvation and influence of counterions

In this section, we discuss the evaluated implicit solvent models for the HS13L and the HS13L-CI benchmark sets. Deviations with respect to the "back-corrected" experimental solvation free energies are shown in Fig. 3. For the HS13L, *i.e.*, without counterions, the fine parametrization of COSMO-RS only performs better for some complexes (**2**, **5**, **12**, and **13**) than

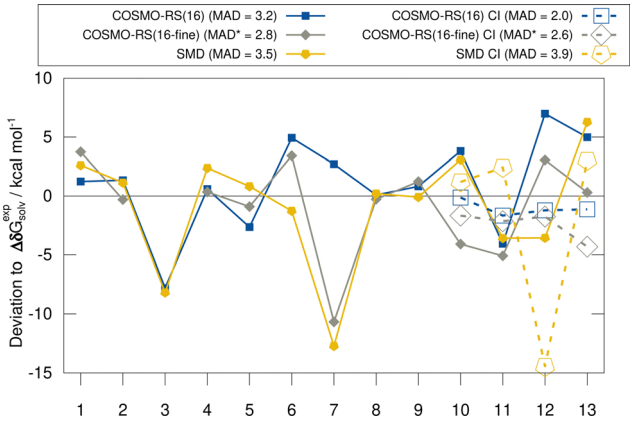


Fig. 3 Deviations to the "back-corrected" experimental solvation free energies for the HS13L with different assessed solvation models. The dashed lines show the results for the complexes with counterions (HS13L-CI). MADs are given in kcal mol^{-1}. The single-point calculations for complex **3** did not converge with BP86/def2-TZVPD for the fine parametrization and were omitted for the respective MAD of the method indicated by the * symbol.

the normal parametrization. However, for complex **3** the SCF did not converge for BP86/def2-SVPD. Removing this outlier, the normal parametrization yields the same MAD of 2.8 kcal mol^{-1} as COSMO-RS-fine. The addition of counterions reduces the deviations to the experimental values of the COSMO-RS results significantly, especially for the normal parametrization (from 3.2 kcal mol^{-1} to 2.0 kcal mol^{-1}) making it the overall best solvation model. In contrast, for SMD the errors increase upon the inclusion of counterions. This is consistent with previous observations made for the S30L[13] but not a good sign in our opinion regarding the quality of the model itself. However, this is due to the large deviation for complex **12**, which may be due to inaccurate atomic radii for arsenic in the SMD model. After excluding this outlier, the MAD for SMD also decreases from 3.5 kcal mol^{-1} to 3.0 kcal mol^{-1}. As expected, the discrepancies between solvent models are larger for polar solvents than for nonpolar solvents, since polar solvents are generally more challenging for implicit solvation models.[38,124]

5.3 Timing comparison

Next, the computational timings are put into perspective. Fig. 4 shows the computational timings for the calculation of the gas-phase association energy for complex **6** including host, guest, and complex scaled down to one CPU core. The most accurate method of each DFA class is shown. For a better discussion of the performance, the MAD to the experimental values of the respective method for HS13L-CI is also given. An upper bound of the serial computation time needed for the complete coupled-cluster based reference value protocol for all complexes (see Section 4) is estimated at 3.3 years, showing how difficult the generation of high-level wave-function theory (WFT) reference values is for such large systems. In practice, the most expensive calculation of the protocol, the coupled-cluster calculation with the def2-TZVPP basis set for complex **4** (265 atoms, 5902 basis functions), took about 18 days wall time

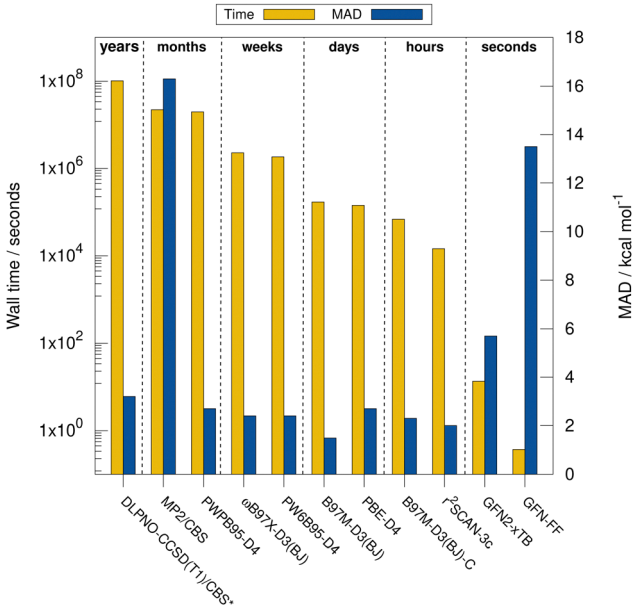


Fig. 4 Computational wall times (Intel® Xeon® E5-2660 v4 @ 2.00 GHz CPU) for the best performing methods of the assessed levels of theory for the calculation of the association energy of complex **6** in combination with the respective MADs to the "back-corrected" experimental values for the HS13L-CI. Complex **3** is not included in the MAD of "DLPNO-CCSD(T1)/CBS" indicated by the *.

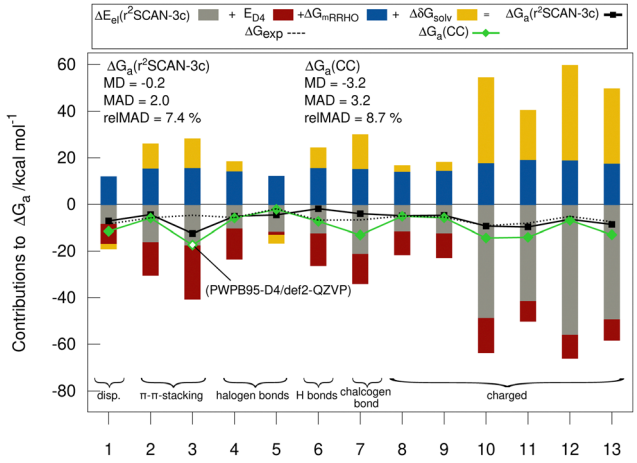


Fig. 5 Contributions to the association free energy ΔG_a for each complex of the HS13L-CI in kcal mol^{-1}. The statistical measures MD, MAD, relMAD, and for comparison, ΔG_a based on "DLPNO-CCSD(T1)/CBS" energies instead of r^{2}SCAN-3c are also given. For complex **3**, no coupled cluster values could be obtained and PWPB95-D4/def2-QZVP values are shown instead.

on 20 CPU cores. This demonstrates the importance of an efficient computational workflow to calculate "back-corrected" gas-phase association energies from experimental values of large systems for developing new efficient computational methods for this system size. MP2 and double hybrid DFAs require months of computation time. Additionally, a significant BSIE/BSSE even with the large quadruple-ζ size basis set used is indicated by the MAD of 2.7 kcal mol^{-1} for PWPB95-D3(BJ), which is larger than that of the best *meta*-GGA methods, such as r^{2}SCAN-3c or B97M-D3(BJ). Hybrid DFAs are already quite expensive with weeks of computation times and are only recommended for cross-checking in difficult cases, *i.e.*, when self-interaction error (artificial charge-transfer effects) is expected to be critical. The BSSE-corrected B97M-D3(BJ)-C method in combination with the def2-SVPD basis set is about 2.5 times faster than B97M-D3(BJ)/def2-QZVP, which comes at the cost of an increase of the MAD by 0.8 kcal mol^{-1}. r^{2}SCAN-3c shows a remarkable performance and gives very accurate results within hours of computation time. GFN2-xTB is very fast providing results in only seconds of computation time. However, considering its relatively large MAD it is only recommended for initial screening steps and for generating accurate geometries. The strength of the very efficient GFN-FF lies in the very low computation time and parametrization for heavier elements which yields reasonable geometries in seconds.

5.4 Contributions to the association free energy

Fig. 5 shows the individual contributions of the default level of theory to the calculated free energies. It is remarkable how this diverse set with very different sizes of contributions ranging from −69 to +40 kcal mol^{-1} depending on the binding motif of the complex lead to very similar and comparably small association energies in the range from −1.9 to −9.2 kcal mol^{-1}. This demonstrates the difficulty of calculating ΔG_a for large supramolecular complexes and renders the MAD of our default workflow of only 2 kcal mol^{-1} even more impressive. The r^{2}SCAN-3c values are in most cases in good agreement with the CC reference values indicating, that the approach mostly gives the right answer for the right physical reasons. Larger differences of over 4 kcal mol^{-1} between the "DLPNO-CCSD(T1)/CBS" values and the r^{2}SCAN-3c values are obtained for complexes **1**, **6**, **7**, **10**, **11**, and **13**. For **1**, **7**, **10**, **11**, and **13** the association free energy obtained with r^{2}SCAN-3c is closer to the experimental value than the coupled cluster result indicating some favorable error compensation in our workflow for these complexes.

5.5 Theory benchmark for HS13L-CI

Because of the potential error compensation between the individual energy contributions to the free association energy, we recommend that methods for calculating gas-phase association energies should be evaluated using the provided coupled cluster reference values (see below), for which at least one error range could be estimated, rather than the "back-corrected" experimental association energies. Hence, to further investigate the accuracy of the employed DFAs for the calculation of gas-phase association energies, we compare the calculated DFT to "DLPNO-CCSD(T1)/CBS" reference values for the HS13L-CI. The reference association energies for HS13L-CI range from −10.9 kcal mol^{-1} to −68.5 kcal mol^{-1} with an average of −38.9 kcal mol^{-1} and an estimated error of approximately 2 kcal mol^{-1} or 5%, respectively (see below). Geometries and all computed energies including the reference values for the

HS13L and HS13L-CI are provided in the ESI.† Because of the limited number of systems composed in HS13L-CI and their difficulty, which is reflected in the large error spread of all tested DFAs, we recommend to use this benchmark set as a challenging extension of S30L.[13]

5.5.1 Coupled cluster reference association energies. To evaluate the performance of the DFAs and SQM methods for their ability to reliably predict association energies for the HS13L-CI complexes, accurate reference values are needed. The systems composed in the HS13L-CI set have a size of 37–266 atoms, which is common for supramolecular complexes. To enable coupled cluster calculations for such large systems, local approximations (or other approaches such as FN-QMC[125,126]) must be applied, which introduces an additional error. To keep this as small as possible, tight threshold values must be used, which in turn makes these calculations quite computationally expensive. In addition, some of the complexes in HS13L-CI have a rather delocalized electronic structure with large correlation energies of up to $-30E_h$. This leads to further limitations of the protocol for the calculation of the reference association energies. The coupled cluster calculations were performed at the DLPNO-CCSD(T)/*TightPNO*/def2-TZVPP level (***ORCA*** 5.0.2 settings), and the resulting errors compared to CCSD(T) at the complete basis set limit (CBS) were approximately corrected as follows: to reduce the BSSE and BSIE, we performed a basis set extrapolation using focal point analysis according to Marshall *et al.*:[121]

$$\begin{aligned}\delta\mathrm{CBS} &= E(\mathrm{MP2/CBS(def2\text{-}TZVPP/def2\text{-}QZVPP)}) \\ &\quad + E_c(\mathrm{DLPNO\text{-}CCSD(T)/def2\text{-}TZVPP}) \\ &\quad - E_c(\mathrm{MP2/def2\text{-}TZVPP}),\end{aligned} \tag{3}$$

where E_c refers to the correlation energy fraction of the total energy E. To keep the local truncation errors as small as possible, we estimated the effect of an even tighter threshold by performing calculations with the smaller def2-SVP basis and added the difference between the correlation energy obtained with *VeryTightPNO* (see Section 4 for details) and *TightPNO* settings as a correction to the coupled cluster correlation energy. Analogously, we estimated the error due to the semilocal triples approximation by performing the corresponding calculations with iterative triples (DLPNO-CCSD(T1)) also at def2-SVP level. The estimation of these two error sources could not be performed with a larger basis set due to the 5–10 times higher computational cost of the T1 and *VeryTightPNO* calculations, respectively. Because the two corrections are small compared with the association energies in HS13L-CI (typically <0.5 kcal mol^{-1} for the semilocal triples and <1 kcal mol^{-1} for the difference between *VeryTightPNO* and *TightPNO*), the additional error due to the smaller basis should play a minor role in the overall error. This protocol presented here will be called "DLPNO-CCSD(T1)/CBS" in the following.

For the smallest complex from HS13L-CI (system **5**), we were also able to calculate reference values without further approximations, supporting the assumptions described above. The comparison of the DLPNO-CCSD(T1)/CBS(def2-TZVPP/def2-QZVPP)/*VeryTightPNO* association energy of -10.95 kcal mol^{-1} (*i.e.*, with genuine basis set extrapolation,[127,128] iterative triples, and *VeryTightPNO* threshold settings also for the extended basis sets) with the corresponding "DLPNO-CCSD(T1)/CBS" value of -10.94 kcal mol^{-1} shows that the additional approximations do not have a large impact on the accuracy of the reference values. Also, the comparison of the CCSD(T)/def2-TZVPP association energy (-10.75 kcal mol^{-1}) with the corresponding "DLPNO-CCSD(T1)/*VeryTightPNO*/def2-TZVPP" value (-10.48 kcal mol^{-1}) shows that for these conservative threshold settings, the local truncation errors are small compared to the association energies of the complexes in HS13L-CI.

The maximum error in the HS13L-CI reference association energies resulting from the local DLPNO approximations, the BSSE and BSIE, and the additional error from the focal point analysis is therefore estimated to be about $\approx 5\%$, which translates to about ± 2 kcal mol^{-1} for a mean association energy of -38.9 kcal mol^{-1}. Nevertheless, this uncertainty in the reference values is largely averaged in the analysis of the statistical descriptors for the entire HS13L-CI set. The square root of the sum of the squares of the estimated maximum error divided by the number of individual association energies, which for HS13L-CI gives ≈ 1.95 kcal mol^{-1}, can be used as an estimate of statistically distinguishable values of the analyzed descriptors (see 5.5.2). Thus, with the given accuracy of the reference values, we are able to distinguish statistically significant errors of each method above 0.5 kcal mol^{-1} in the respective MADs.

5.5.2 Electronic energies. Table 3 lists MDs, MADs, and SDs with respect to the DLPNO-CCSD(T1)/CBS reference values for all DFAs assessed in combination with their respective best-performing dispersion correction. Overall, a tendency for underbinding with respect to the reference is obvious for all DFAs assessed. With a MAD of only 2.5 kcal mol^{-1} B97M-V-C gives remarkable accurate results. Although r^{2}SCAN-3c shows systematically underbinding (3.0 kcal mol^{-1} MD) to the experimental values the MAD of 3.4 kcal mol^{-1} can be still regarded as good for this efficient composite method. The GGAs PBE-NL and RPBE-NL both yield good results. B97M-V(1.9 kcal mol^{-1}) is the most accurate *meta*-GGA, which was also the result of a recent benchmark study conducted by Villot *et al.* on interaction energies of large NCI complexes,[126] whereas M06L-D3 has a large MAD of 3.7 kcal mol^{-1}. Surprisingly, no systematic improvement is observed upon inclusion of Fock exchange. Well performing hybrid DFAs are B3LYP-NL, PW6B95-NL, PBE0-D4, and ωB97X-V (MADs of 2.2–2.6 kcal mol^{-1}), whereas ωB97M-V (3.2 kcal mol^{-1} MAD) and M06-2X-D3 (3.0 kcal mol^{-1} MAD) show larger deviations. This also holds for the double hybrid DFAs PWPB95-D3(BJ) (2.9 kcal mol^{-1} MAD) and rev-DSDPBEP86-D4 (3.8 kcal mol^{-1} MAD), which we tentatively attribute to the bad performance of MP2 on this set (MD of -13.1 kcal mol^{-1}) and a remaining BSIE/BSSE for this class of DFAs. Downscaling the ATM term of rev-DSDPBEP86-D4 reduces the underbinding and yields a MAD of 2.7 kcal mol^{-1}.

Fig. 6 shows the deviation between the best DFA of each class and the DLPNO-CCSD(T1)/CBS reference values in detail

Table 3 MDs, MADs, and SDs in kcal mol^{-1} to the "DLPNO-CCSD(T1)/CBS" reference values of all assessed DFAs with the def2-QZVP basis which are used in combination with their best dispersion correction and composite methods for the HS13L-CI. For complex **3**, PWPB95-D4/def2-QZVP values were used as reference, as no "DLPNO-CCSD(T1)/CBS" reference value could be obtained for this complex

	MD	MAD	SD
r^{2}SCAN-3c	3.0	3.4	3.0
PBEh-3c	6.2	8.0	9.4
B97-3c	3.8	5.3	9.9
B97M-V-C	-0.7	2.5	3.2
PBE-NL	0.6	2.7	3.3
RPBE-NL	-1.3	3.1	3.9
r^{2}SCAN-D4	2.4	2.8	2.9
M06L-D3(0)	-1.1	3.7	4.9
B97M-V	-0.2	1.9	2.7
PW6B95-NL	0.5	2.5	3.5
B3LYP-NL	-0.7	2.2	3.1
M06-2X-D3(0)	1.6	3.0	4.1
PBE0-D4	1.9	2.3	3.2
ωB97X-V	-1.2	2.6	3.2
ωB97M-V	-2.2	3.2	3.2
PWPB95-D3 (BJ)	0.7	2.9	3.6
rev-DSDPBEP86-D4	2.9	3.8	3.8
rev-DSDPBEP86-D4[a]	0.5	2.7	4.1
MP2-CBS	-13.1	13.1	12.4

[a] Downscaled *s*9.

for each complex for the HS13L-CI. For the large halogen-bonded complex **4**, all methods except r^{2}SCAN-3c show strong overbinding of over 5 kcal mol^{-1}. Systematic overbinding of all methods is observed for the tellurium containing complex **7** and less pronounced for complex **1** containing I_2.

5.5.3 Dispersion corrections. In the following, the accuracy of different LD corrections is assessed for the assessed DFAs.

MADs from the "DLPNO-CCSD(T1)/CBS" reference values for the HS13L-CI obtained with the assessed DFAs in combination with D3, D4, and the VV10 correction are shown in Fig. 7. For most DFAs, the VV10 correction yields smaller deviations than the D3 or D4 correction, although higher-order dispersion terms are missing in this model. The NL corrected DFAs tend to overbinding, whereas the D3 and D4 corrected DFAs show the opposite with respect to the "DLPNO-CCSD(T1)/CBS" reference, which can be attributed to the mostly repulsive ATM term in the latter two (see ESI,† for details).

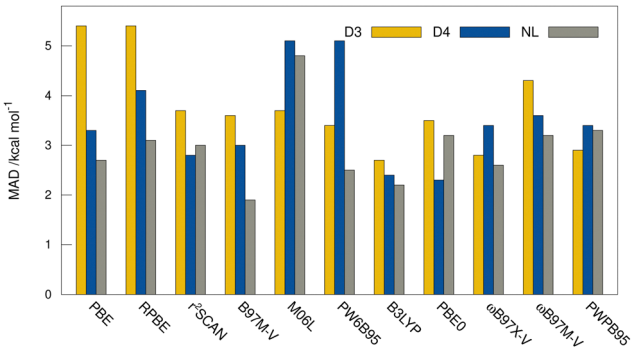


Fig. 7 MADs in kcal mol^{-1} to the DLPNO-CCSD(T1)/CBS reference values of the assessed DFAs for the HS13L-CI in combination with D3, D4, and the nonlocal VV10 dispersion corrections.

5.5.4 Performance of semiempirical methods. Due to the large system sizes and the large number of possible conformers, efficient methods are needed for screening applications of supramolecular complexes, *e.g.*, in drug development. Therefore, we also evaluate the accuracy of efficient SQM and FF methods for energies and geometries. Table 4 shows the deviations of the assessed SQM methods with respect to the "DLPNO-CCSD(T1)/CBS" reference values for the HS13L-CI.

The large error range of all assessed methods is notable. GFN1-xTB is the best method with a MAD of 6.1 kcal mol^{-1}. GFN2-xTB (MAD of 6.6 kcal mol^{-1}) also yields reasonably good results. PM6 (7.8 kcal mol^{-1}) shows larger deviations than both methods. PM7 (11.9 kcal mol^{-1}) and the force-field GFN-FF (13.6 kcal mol^{-1}) yield a similar accuracy. In summary, this shows that even the most accurate SQM methods tested, namely GFN1- and GFN2-xTB, are only useful in the early stage of screening procedures but with relatively large energy windows for structure selection.

The mean heavy atom $\overline{\text{RMSD}}$ of the optimized geometries in solution are compared to the r^{2}SCAN-3c[SMD] optimized geometries of HS13L, shown in Table 4. GFN2-xTB[ALPB] structures are remarkably accurate with a deviation of only

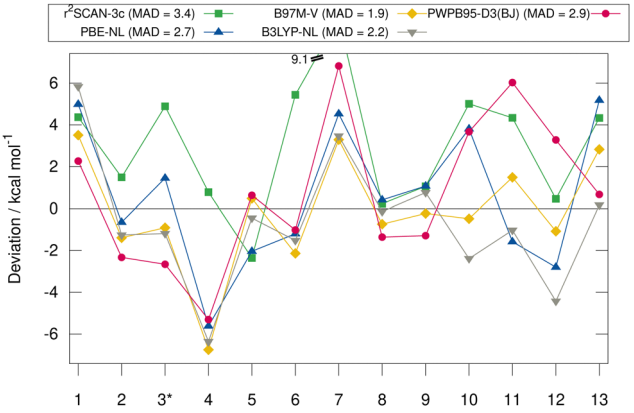


Fig. 6 Deviations in kcal mol^{-1} to the DLPNO-CCSD(T1)/CBS reference values of best-performing methods of each class for HS13L-CI. A negative deviation indicates overbinding, a positive underbinding. For complex **3** PWPB95-D4/def2-QZVP is used as reference, as no DLPNO-CCSD(T1)/CBS reference value could be obtained for this complex indicated by the *.

Table 4 MD, MAD, and SD from the "DLPNO-CCSD(T1)/CBS" reference values of the tested semiempirical methods for the HS13L-CI in kcal mol^{-1}. The average root-mean square deviation of the heavy atom positions ($\overline{\text{RMSD}}$) between the geometries optimized with SQM/FF methods and r^{2}SCAN-3c[SMD] for the HS13L is given in Å

	MD	MAD	SD	$\overline{\text{RMSD}}$
GFN1-xTB	3.2	6.1	6.7	0.25
GFN2-xTB	6.5	6.6	5.9	0.17
GFN-FF	8.0	13.6	20.7	0.22
PM6	0.4	7.8	11.2	0.30
PM7	-4.6	11.9	14.4	0.29

0.17 Å. Also GFN-FF[ALPB] yields outstanding geometries with an $\overline{\text{RMSD}}$ of 0.22 Å, which is better than the SQM methods GFN1-xTB[ALPB] (0.25 Å), PM6-D3H4X [COSMO] (0.30 Å), and PM7[COSMO] (0.29 Å).

6 Conclusions

The reliable prediction of association free energies of supramolecular complexes containing (heavy) main group elements is an important yet challenging task for computational methods. Especially for large systems, for which highly accurate reference energies are difficult to obtain with WFT methods, an efficient workflow for the calculation of experimentally accessible association free energies is needed. We assessed the ***CREST*** and ***CENSO*** workflow of conformer generation with SQM methods and subsequent free energy ranking at DFT level for the first time systematically on large, heavy atom containing supramolecular complexes in direct comparison to experimental values. We introduced a new benchmark set of 13 supramolecular complexes (HS13L or HS13L-CI with counterions added, respectively). By comparison to experimental association free energies, we showed that our protocol reliably predicts this property with a MAD of only 2.0 kcal mol^{-1}.

The comparison between various dispersion corrected DFAs and accurate local coupled cluster values shows that in special cases calculations profit from error compensation between the electronic energy and the solvation free energy between the individual contributions to the free energy. Therefore, we recommend to benchmark methods against the ''DLPNO-CCSD(T1)/CBS'' reference values instead of the ''back-corrected'' experimental gas-phase association energies. Large deviations of the DFA gas-phase association energy of over 4 kcal mol^{-1} from the respective coupled cluster values were observed for some complexes of HS13L-CI. Of the assessed methods, the *meta*-GGA composite method r^2SCAN-3c proved to be robust and showed the best cost-accuracy ratio, outperforming some very popular hybrid and double-hybrid DFAs. Also, B97M-D3(BJ) showed a remarkable accuracy. The assessed SQM methods yield large deviations from the ''DLPNO-CCSD(T1)/CBS'' reference values and are therefore only recommended for initial screening steps. However, GFN2-xTB[ALPB] gives accurate geometries for the HS13L and proves to be a viable tool for the conformer generation of challenging supramolecular complexes. The default structure thresholds in the ***CENSO*** workflow are also suitable for these systems. From the assessed solvation models, COSMO-RS is the most accurate for solvation free energies, whereas SMD is the more robust alternative to DCOSMO-RS for the geometry optimization. We propose the HS13L-CI set as test set for the development of new solvation models. GFN2-xTB provides accurate thermostatistical contributions in the SPH approach and can be recommended also for heavy main group systems. We recommend the HS13L-CI benchmark as an extension to the well established S30L set specifically assessing new methods for their robustness and accuracy for systems with heavy main group elements.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors thank Fabian Bohle for technical help regarding the ***CENSO*** calculations, fruitful discussions, and helpful suggestions, as well as Thomas Gasevic for providing helpful corrections to the manuscript.

References

1 I. V. Kolesnichenko and E. V. Anslyn, *Chem. Soc. Rev.*, 2017, **46**, 2385–2390.
2 D. Phlip, *Adv. Mater.*, 1996, **8**, 866–868.
3 I. P. Parkin, *Appl. Organomet. Chem.*, 2001, **15**, 236.
4 J.-M. Lehn, *Angew. Chem., Int. Ed.*, 1988, **27**, 89–112.
5 D. J. Cram, *Angew. Chem., Int. Ed.*, 1988, **27**, 1009–1020.
6 J. M. Lehn, *Science*, 1993, **260**, 1762–1763.
7 Z. Zhang and P. R. Schreiner, *Chem. Soc. Rev.*, 2009, **38**, 1187.
8 M. A. Pitt and D. W. Johnson, *Chem. Soc. Rev.*, 2007, **36**, 1441–1453.
9 F. P. Schmidtchen, *Isothermal Titration Calorimetry in Supramolecular Chemistry*, John Wiley & Sons, Ltd, 2012.
10 G. A. Holdgate and W. H. Ward, *Drug Discovery Today*, 2005, **10**, 1543–1550.
11 S. Grimme and P. R. Schreiner, *Angew. Chem., Int. Ed.*, 2018, **57**, 4170–4176.
12 S. Grimme, *Chem. – Eur. J.*, 2012, **18**, 9955–9964.
13 R. Sure and S. Grimme, *J. Chem. Theory Comput.*, 2015, **11**, 3785–3801.
14 N. Mehta, T. Fellowes, J. M. White and L. Goerigk, *J. Chem. Theory Comput.*, 2021, **17**, 2783–2806.
15 J. Řezáč, *J. Chem. Theory Comput.*, 2020, **16**, 6305–6316.
16 K. Kříž and J. Řezáč, *Phys. Chem. Chem. Phys.*, 2022, **24**, 14794–14804.
17 J. Řezáč, *Phys. Chem. Chem. Phys.*, 2022, **24**, 14780–14793.
18 J. F. Dobson, *Int. J. Quantum Chem.*, 2014, **114**, 1157–1161.
19 P. Pracht, F. Bohle and S. Grimme, *Phys. Chem. Chem. Phys.*, 2020, **22**, 7169–7192.
20 S. Grimme, F. Bohle, A. Hansen, P. Pracht, S. Spicher and M. Stahn, *J. Phys. Chem. A*, 2021, **125**, 4039–4054.
21 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.
22 S. Grimme, S. Ehrlich and L. Goerigk, *J. Comput. Chem.*, 2011, **32**, 1456–1465.
23 A. D. Becke and E. R. Johnson, *J. Chem. Phys.*, 2007, **127**, 124108.
24 A. D. Becke and E. R. Johnson, *J. Chem. Phys.*, 2007, **127**, 154108.

25 E. Caldeweyher, S. Ehlert, A. Hansen, H. Neugebauer, S. Spicher, C. Bannwarth and S. Grimme, *J. Chem. Phys.*, 2019, **150**, 154122.
26 Y. Muto, *J. Phys.-Math. Soc. Jpn*, 1943, **17**, 629.
27 B. M. Axilrod and E. Teller, *J. Chem. Phys.*, 1943, **11**, 299–300.
28 A. D. Becke and E. R. Johnson, *J. Chem. Phys.*, 2005, **123**, 154101.
29 A. D. Becke and E. R. Johnson, *J. Chem. Phys.*, 2005, **122**, 154104.
30 A. Tkatchenko, R. A. DiStasio, R. Car and M. Scheffler, *Phys. Rev. Lett.*, 2012, **108**, 236402.
31 R. A. DiStasio, V. V. Gobre and A. Tkatchenko, *J. Phys.: Condens. Matter*, 2014, **26**, 213202.
32 O. A. Vydrov and T. V. Voorhis, *J. Chem. Phys.*, 2010, **133**, 244103.
33 W. Hujo and S. Grimme, *J. Chem. Theory Comput.*, 2011, **7**, 3866–3871.
34 S. Spicher and S. Grimme, *J. Chem. Theory Comput.*, 2021, **17**, 1701–1714.
35 P. Pracht and S. Grimme, *Chem. Sci.*, 2021, **12**, 6551–6568.
36 J. Gorges, S. Grimme, A. Hansen and P. Pracht, *Phys. Chem. Chem. Phys.*, 2022, **24**, 12249–12259.
37 A. Klamt, V. Jonas, T. Bürger and J. C. W. Lohrenz, *J. Phys. Chem. A*, 1998, **102**, 5074–5085.
38 A. Klamt, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2018, **8**, e1338.
39 A. V. Marenich, C. J. Cramer and D. G. Truhlar, *J. Chem. Theory Comput.*, 2013, **9**, 609–620.
40 S. Spicher, C. Plett, P. Pracht, A. Hansen and S. Grimme, *J. Chem. Theory Comput.*, 2022, **18**, 3174–3189.
41 J. H. Jensen, *Phys. Chem. Chem. Phys.*, 2015, **17**, 12441–12451.
42 J. Řezáč and P. Hobza, *Chem. Rev.*, 2016, **116**, 5038–5071.
43 H. S. El-Sheshtawy, B. S. Bassil, K. I. Assaf, U. Kortz and W. M. Nau, *J. Am. Chem. Soc.*, 2012, **134**, 19935–19941.
44 K. Takahashi, S. Shimo, E. Hupf, J. Ochiai, C. A. Braun, W. T. Delgado, L. Xu, G. He, E. Rivard and N. Iwasawa, *Chem. – Eur. J.*, 2019, **25**, 8479–8483.
45 Y. Eda, K. Itoh, Y. N. Ito, M. Fujitsuka, T. Majima and T. Kawato, *Supramol. Chem.*, 2010, **22**, 517–523.
46 C. Gropp, T. Husch, N. Trapp, M. Reiher and F. Diederich, *J. Am. Chem. Soc.*, 2017, **139**, 12190–12200.
47 O. Dumele, D. Wu, N. Trapp, N. Goroff and F. Diederich, *Org. Lett.*, 2014, **16**, 4722–4725.
48 H. Gan and B. C. Gibb, *Supramol. Chem.*, 2010, **22**, 808–814.
49 C. Liu, Z. Zhang, Z. Fan, C. He, Y. Tan and H. Xu, *Chem. – Asian J.*, 2020, **15**, 4321–4326.
50 P. I. Dron, S. Fourmentin, F. Cazier, D. Landy and G. Surpateanu, *Supramol. Chem.*, 2008, **20**, 473–477.
51 H. Ren, Z. Huang, H. Yang, H. Xu and X. Zhang, *ChemPhysChem*, 2015, **16**, 523–527.
52 K. I. Assaf, M. S. Ural, F. Pan, T. Georgiev, S. Simova, K. Rissanen, D. Gabel and W. M. Nau, *Angew. Chem., Int. Ed.*, 2015, **54**, 6852–6856.
53 N. Yoshida and K. Hayashi, *J. Chem. Soc., Perkin Trans. 2*, 1994, 1285–1290.
54 D. Menozzi, E. Biavardi, C. Massera, F.-P. Schmidtchen, A. Cornia and E. Dalcanale, *Supramol. Chem.*, 2010, **22**, 768–775.
55 N. J. Wheate and C. Limantoro, *Supramol. Chem.*, 2016, **28**, 849–856.
56 M. Zander and G. Kirsch, *Z. Naturforsch., A: Phys. Sci.*, 1989, **44**, 205–209.
57 M. V. Rekharsky and Y. Inoue, *Chem. Rev.*, 1998, **98**, 1875–1918.
58 Conformer-Rotamer Ensemble Sampling Tool based on the xtb Semiempirical Extended Tight-Binding Program Package crest, **https://github.com/crest-lab/crest**, Accessed: 2021-12-21.
59 C. Bannwarth, S. Ehlert and S. Grimme, *J. Chem. Theory Comput.*, 2019, **15**, 1652–1671.
60 S. Ehlert, M. Stahn, S. Spicher and S. Grimme, *J. Chem. Theory Comput.*, 2021, **17**, 4250–4261.
61 Semiempirical Extended Tight-Binding Program Package xtb, **https://github.com/grimme-lab/xtb**, Accessed: 2022-5-15.
62 S. Spicher and S. Grimme, *Angew. Chem., Int. Ed.*, 2020, **132**, 15795–15803.
63 S. Grimme, C. Bannwarth and P. Shushkov, *J. Chem. Theory Comput.*, 2017, **13**, 1989–2009.
64 S. Grimme, F. Bohle, A. Hansen, P. Pracht, S. Spicher and M. Stahn, *J. Phys. Chem. A*, 2021, **125**, 4039–4054.
65 Commandline ENergetic SOrting of Conformer Rotamer Ensembles censo, **https://github.com/grimme-lab/censo**, Accessed: 2022-2-11.
66 A. H. Markus Bursch, J.-M. Mewes and S. Grimme, *Chem. Rev.*, 2022, 1875–1918.
67 S. Grimme, *J. Comput. Chem.*, 2006, **27**, 1787–1799.
68 F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297–3305.
69 H. Kruse and S. Grimme, *J. Chem. Phys.*, 2012, **136**, 154101.
70 S. Spicher and S. Grimme, *J. Chem. Theory Comput.*, 2021, **17**, 1701–1714.
71 S. Grimme, A. Hansen, S. Ehlert and J.-M. Mewes, *J. Chem. Phys.*, 2021, **154**, 064103.
72 F. Neese, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2022, e1606.
73 A. Klamt and M. Diedenhofen, *J. Phys. Chem. A*, 2015, **119**, 5439–5445.
74 O. Vahtras, J. Almlöf and M. W. Feyereisen, *Chem. Phys. Lett.*, 1993, **213**, 514–518.
75 F. Neese, F. Wennmohs, A. Hansen and U. Becker, *Chem. Phys.*, 2009, **356**, 98–109.
76 F. Weigend, *Phys. Chem. Chem. Phys.*, 2006, **8**, 1057–1065.
77 F. Weigend, A. Köhn and C. Hättig, *J. Chem. Phys.*, 2002, **116**, 3175–3183.
78 A. Hellweg, C. Hättig, S. Höfener and W. Klopper, *Theor. Chem. Acc.*, 2007, **117**, 587–597.
79 Reimplementation of the DFT-D3 program s-DFTD3, **https://github.com/awvwgk/simple-dftd3**, Accessed: 2022-06-10.
80 Generally Applicable Atomic-Charge Dependent London Dispersion Correction DFTD4, **https://github.com/dftd4/dftd4**, Accessed: 2022-07-06.

81 E. R. Johnson and A. D. Becke, *J. Chem. Phys.*, 2005, **123**, 024101.
82 J. Witte, J. B. Neaton and M. Head-Gordon, *J. Chem. Phys.*, 2017, **146**, 234105.
83 D. Rappoport and F. Furche, *J. Chem. Phys.*, 2010, **133**, 134105.
84 E. Epifanovsky, A. T. Gilbert, X. Feng, J. Lee, Y. Mao, N. Mardirossian, P. Pokhilko, A. F. White, M. P. Coons and A. L. Dempwolff, *et al.*, *J. Chem. Phys.*, 2021, **155**, 084801.
85 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865.
86 Y. Zhang and W. Yang, *Phys. Rev. Lett.*, 1998, **80**, 890.
87 J. W. Furness, A. D. Kaplan, J. Ning, J. P. Perdew and J. Sun, *J. Phys. Chem. Lett.*, 2020, **11**, 8208–8215.
88 N. Mardirossian and M. Head-Gordon, *J. Chem. Phys.*, 2015, **142**, 074111.
89 A. Najibi and L. Goerigk, *J. Chem. Theory Comput.*, 2018, **14**, 5725–5738.
90 A. Najibi and L. Goerigk, *J. Comput. Chem.*, 2020, **41**, 2562–2572.
91 Y. Zhao and D. G. Truhlar, *J. Chem. Phys.*, 2006, **125**, 194101.
92 J. Brandenburg, J. Bates, J. Sun and J. Perdew, *Phys. Rev. B*, 2016, **94**, 115144.
93 Y. Zhao and D. G. Truhlar, *J. Phys. Chem. A*, 2005, **109**, 5656–5667.
94 C. Adamo and V. Barone, *J. Chem. Phys.*, 1999, **110**, 6158–6170.
95 A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 5648–5652.
96 C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1988, **37**, 785–789.
97 Y. Zhao and D. G. Truhlar, *Theor. Chem. Acc.*, 2008, **120**, 215–241.
98 N. Mardirossian and M. Head-Gordon, *J. Chem. Phys.*, 2016, **144**, 214110.
99 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.*, under review.
100 N. Mardirossian and M. Head-Gordon, *Phys. Chem. Chem. Phys.*, 2014, **16**, 9904–9924.
101 F. Yu, *J. Chem. Theory Comput.*, 2014, **10**, 4400–4407.
102 L. Goerigk and S. Grimme, *J. Chem. Theory Comput.*, 2011, **7**, 291–309.
103 G. Santra, N. Sylvetsky and J. M. Martin, *J. Phys. Chem. A*, 2019, **123**, 5129–5143.
104 G. Santra, M. Cho and J. M. Martin, *J. Phys. Chem. A*, 2021, **125**, 4614–4627.
105 J. G. Brandenburg, C. Bannwarth, A. Hansen and S. Grimme, *J. Chem. Phys.*, 2018, **148**, 064104.
106 S. Grimme, A. Hansen, S. Ehlert and J.-M. Mewes, *J. Chem. Phys.*, 2021, **154**, 064103.
107 S. Grimme, J. G. Brandenburg, C. Bannwarth and A. Hansen, *J. Chem. Phys.*, 2015, **143**, 054107.
108 TURBOMOLE V7.5 2020, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007; available from **https://www.turbomole.com**.
109 A. Klamt and G. Schüürmann, *J. Chem. Soc., Perkin Trans. 2*, 1993, 799–805.
110 J. J. P. Stewart, *MOPAC2016*, Stewart Computational Chemistry, Colorado Springs, CO, USA, 2016.
111 F. Eckert and A. Klamt, *AIChE J.*, 2002, **48**, 369–385.
112 F. Eckert and A. Klamt, *COSMOtherm, version C3.0, release 19.01*, COSMOlogic GmbH & Co. KG, Leverkusen, Germany, 2019.
113 A. D. Becke, *Phys. Rev. A: At., Mol., Opt. Phys.*, 1988, **38**, 3098–3100.
114 J. P. Perdew, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1986, **33**, 8822–8824.
115 V. Barone and M. Cossi, *J. Phys. Chem. A*, 1998, **102**, 1995–2001.
116 A. Karton and J. M. Martin, *J. Chem. Phys.*, 2012, **136**, 124114.
117 C. Riplinger, B. Sandhoefer, A. Hansen and F. Neese, *J. Chem. Phys.*, 2013, **139**, 134101.
118 C. Riplinger, P. Pinski, U. Becker, E. F. Valeev and F. Neese, *J. Chem. Phys.*, 2016, **144**, 024109.
119 Y. Guo, C. Riplinger, U. Becker, D. G. Liakos, Y. Minenkov, L. Cavallo and F. Neese, *J. Chem. Phys.*, 2018, **148**, 011101.
120 A. G. Császár, W. D. Allen and H. F. Schaefer, *J. Chem. Phys.*, 1998, **108**, 9751–9764.
121 M. S. Marshall, L. A. Burns and C. D. Sherrill, *J. Chem. Phys.*, 2011, **135**, 194102.
122 S. Spicher, E. Caldeweyher, A. Hansen and S. Grimme, *Phys. Chem. Chem. Phys.*, 2021, **23**, 11635–11648.
123 T. Risthaus and S. Grimme, *J. Chem. Theory Comput.*, 2013, **9**, 1580–1591.
124 D. L. Mobley, A. E. Barber, C. J. Fennell and K. A. Dill, *J. Phys. Chem. B*, 2008, **112**, 2405–2414.
125 Y. S. Al-Hamdani, P. R. Nagy, A. Zen, D. Barton, M. Kállay, J. G. Brandenburg and A. Tkatchenko, *Nat. Commun.*, 2021, **12**, 1–12.
126 C. Villot, F. Ballesteros, D. Wang and K. U. Lao, *J. Phys. Chem. A*, 2022, **126**, 4326–4341.
127 T. Helgaker, W. Klopper, H. Koch and J. Noga, *J. Chem. Phys.*, 1997, **106**, 9639–9646.
128 A. Halkier, T. Helgaker, P. Jørgensen, W. Klopper, H. Koch, J. Olsen and A. K. Wilson, *Chem. Phys. Lett.*, 1998, **286**, 243–252.

APPENDIX C

Appendix: QCxMS2 – a Program for the Calculation of Electron Ionization Mass Spectra via Automated Reaction Network Discovery

Johannes Gorges[†], Stefan Grimme[†]

Received: 23 January 2025
Published online: 03 March 2025

Reproduced in Appendix C from
J. Gorges and S. Grimme, *QCxMS2 - a program for the calculation of electron ionization mass spectra via automated reaction network discovery*, Phys. Chem. Chem. Phys. **27**.14 (2025) 6899, DOI: 10.1039/d5cp00316d with permission from the Royal Society of Chemistry.

Own contributions

- Development of the QCxMS2 software
- Performing calculations with QCxMS2 using various program parameters and quantum chemical methods
- Interpretation of the results
- Writing the manuscript

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

PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2025, **27**, 6899

QCxMS2 – a program for the calculation of electron ionization mass spectra *via* automated reaction network discovery†

Johannes Gorges and Stefan Grimme *

We present a new fully-automated computational workflow for the calculation of electron ionization mass spectra by automated reaction network discovery, transition state theory and Monte-Carlo simulations. Compared to its predecessor QCxMS [S. Grimme, *Angew. Chem., Int. Ed.*, **52**, 6306–6312] based on extensive molecular dynamics (MD) simulations, QCxMS2's more efficient approach of using stationary points on the potential energy surface (PES) enables the usage of accurate quantum chemical methods. Fragment geometries and reaction paths are optimized with fast semi-empirical quantum mechanical (SQM) methods and reaction barriers are refined at the density functional theory (DFT) level. This composite approach using GFN2-xTB geometries in combination with energies at the ωB97X-3c level proved to be an efficient combination. On a small but diverse test set of 16 organic and inorganic molecules, QCxMS2 spectra are more accurate than ones from QCxMS yielding on average a higher mass spectral matching of 0.700 compared to QCxMS with 0.622, and is more robust with a minimal matching of 0.498 *versus* 0.100. Further improvements were observed when both geometries and energies were computed at the ωB97X-3c level, yielding an average matching score of 0.730 and a minimal score of 0.527. Due to its higher accuracy and robustness while maintaining computational efficiency, we propose QCxMS2 as a complementary, more reliable and systematically improvable successor to QCxMS for elucidating fragmentation pathways and predicting electron ionization mass spectra of unknown chemical substances, *e.g.*, in analytical chemistry applications. If coupled to currently developed improved SQM methods, QCxMS2 opens an efficient route to accurate, and routine mass spectra predictions. The QCxMS2 program suite is freely available on GitHub.

Received 23rd January 2025,
Accepted 28th February 2025

DOI: 10.1039/d5cp00316d

rsc.li/pccp

1 Introduction

Mass spectrometry (MS) is due to its high sensitivity and high-throughput capability an indispensable tool for structure elucidation in many areas of chemistry, such as drug discovery,[1] metabolomics[2] or forensics.[3] However, assigning a spectrum to an unknown substance is a challenging task and often proves unsuccessful.[2] For example, in recent metabolomics studies approximately 70% of the target metabolites remained unidentified despite extensive efforts.[4,5] Despite its great importance, reliable theoretical prediction of mass spectra routinely remains a challenge for chemical theory, and structure annotations in common *in silico* generated MS libraries are frequently found to be incorrect.[6] Data-driven machine-learning approaches, such as NEIMS[7] for electron ionization mass spectra (EI-MS), and GrAFF-MS,[8] CFM-ID,[9] and the recent ICEBERG model[10] for electrospray ionization/collision-induced dissociation mass spectra (ESI/CID-MS) show remarkable accuracy but are dependent on known data and are therefore unreliable for the prediction of unknown, unusual fragmentation pathways.[11]

To this end, our group has developed some years ago the QCEIMS program for the automatic calculation of standard 70 eV electron ionization mass spectra. It is based on Born–Oppenheimer molecular dynamics (BO-MD) using efficient quantum mechanical (QM) methods to simulate the fragmentation processes of molecules.[12] Due to the computational costs, the BO-MD simulations are mostly restricted to semi-empirical quantum mechanical (SQM) methods. The method was later extended to enable the simulation of electrospray ionization/collision-induced dissociation mass spectrometry (ESI/CID MS) and its name was changed to QCxMS (x = CID,

Mulliken Center for Theoretical Chemistry, Clausius-Institute for Physical and Theoretical Chemistry, University of Bonn, Beringstr. 4, 53115 Bonn, Germany. E-mail: grimme@thch.uni-bonn.de

† Electronic supplementary information (ESI) available: [Geometries in *xyz* format for all structures, as well as the computed spectra for the test set, can be found here: **https://github.com/grimme-lab/QCxMS2-data/**]. Additional details on the implementation, tests of different technical parameters, and computed spectra, which are not in the manuscript are provided in the file SI.pdf. See DOI: **https://doi.org/10.1039/d5cp00316d**

EI) to account for the new functionality of the program.[13] For the calculation of CID spectra, other quantum chemistry (QC)-based methods that use MD simulations, such as CIDMD[14] and the VENUS program package,[15,16] are also available. However, their accuracy has not yet been tested on a broad range of compounds, nor have they been applied to EI-MS.

QCxMS' good accuracy for a large variety of molecules was proven in several studies by our group[17,18] and others.[19,20] In several applications, it showed great success in elucidating unknown fragmentation pathways, *e.g.*, for environmental pollutants[21,22] or chemical warfare agents.[23] However, for challenging molecules or if complicated fragmentation pathways are involved, in some cases, significant deviations from the experimentally measured spectra are observed with QCxMS. In a recent study on a large number of diverse organic environmentally relevant molecules, QCxMS spectra at the GFN1-xTB[24] level were found to be too inaccurate for the application in spectral matching workflows. In particular, flexible molecules and molecules containing heteroatoms other than H, C, N, and O were found to be difficult for QCxMS.[25] Additionally, a separate study found that the spectra of organic oxygen compounds are often inaccurate.[20]

We concluded that many failures can be attributed to two fundamental limitations of the approach of simulating the fragmentation by MD simulations:

1. To keep computationally feasible, the time scale of the computations (by default 5 ps for a single reaction trajectory) is orders of magnitude shorter than the real time scale of slower fragmentations, which may occur on the ns up to the μs timescale. Consequently, the corresponding peaks can be completely missing in the computed spectra.

2. Already for medium-sized molecules (30–50 atoms), the level of theory for the underlying potential energy surface (PES) in the MD simulations is limited to rather approximate SQM methods. The corresponding errors for reaction energies and barrier heights directly (and in exponentially weighted form) influence the computed reaction probabilities (spectral intensities). Reducing the errors due to the SQM methods by performing the MD simulations at a higher density functional theory (DFT) level is impossible with typical computational resources.

An alternative, completely different route to the BO-MD-based approach is Rice–Ramsperger–Kassel–Marcus[26–28] quasi-equilibrium theory[29] (RRKM/QET). In this approach, relative intensities are calculated from reaction rates derived from barrier heights in the reaction network and the resulting "master equations". Drahos and Ve′key expanded this theory to non-equilibrium situations and implemented it in the program "Mass Kinetics".[30] RRKM/QET was applied in several studies concerning EI or CID mass spectrometry.[31–35] For more examples, we refer to ref. 36, where an overview of some important applications is given. However, these examples concern only small molecules, where a manual setup of all relevant reaction pathways is feasible. None of these approaches has been used routinely in a black-box type procedure for automated spectra prediction.

Here, we introduce a new program, QCxMS2, which enables the fully automated computation of mass spectra based on automated reaction network discovery. Herein, a forward open-end exploration approach[37] is followed, which focuses exclusively on unimolecular reactions happening in MS experiments, in contrast to more general exploration software, such as Chemoton,[38] Nanoreactor[39] or AutoMekn2021.[40] In QCxMS2, the well-established RRKM/QET approach is integrated with automated fragment/product generation and an efficient workflow utilizing QM methods to calculate reaction barriers. Previously well working parts in QCxMS like the assignment and treatment of fractional charge, the cascading reaction concept or the internal energy distribution model are kept.

This initial work focuses on the calculation of electron-ionization mass spectra (EI-MS) but the approach can be easily extended to CID. We begin by providing a brief overview of the theoretical background of the new approach. Next, we describe the implementation of the workflow and the computational details of the software. To assess the accuracy of the new QCxMS2 method, we apply it to a benchmark set of 16 organic and inorganic main-group molecules with diverse structural motifs and typical fragmentation patterns. We compare the resulting spectra to those computed by QCxMS, which, to the best of our knowledge, is the only comparable first-principles method for the QM-based calculation of EI-MS. After discussing computational timings, we present general conclusions on the accuracy and limitations of QCxMS2 and recommend potential use cases.

2 Theory

The basic assumption of QET is that fragmentation reactions in a mass spectrometer occur from thermally excited but quasi-equilibrated ions.[29] According to RRKM theory,[26–28] the rate constants of unimolecular decompositions is a function of the internal energy, E of an isolated ion

$$k(E) = \frac{\sigma N^{\ddagger}(E - E_a)}{h\rho(E)}, \quad (1)$$

where σ is the reaction path degeneracy, h is Planck's constant, $N^{\ddagger}(E - E_a)$ is the transition state sum of states, and $\rho(E)$ is the density of states, for which often only the vibrational states are considered.[41] Since accurate vibrational frequencies are required for the computation of $\rho(E)$ and $N^{\ddagger}(E - E_a)$, which have to be calculated on a fully geometry optimized transition state as even small imaginary frequencies would distort the result, it is challenging to compute them accurately in an automated workflow. Furthermore, barrierless reactions without clear transition state are often observed in the fragmentation reactions, for which a description by phase space theory[42–44] or variational transition state theory[45–48] would be needed.[49]

In preliminary studies, we found the advanced treatments mentioned above are impractical to use in an automated workflow as the uncertainty for the depending variables caused by errors of the employed underlying QM method or the overall workflow led to too large errors. Therefore, we decided to employ the Eyring equation from conventional transition state

theory[50] as a more robust but less exact theoretical description of the reaction rates within the QCxMS2 workflow to avoid inconsistent treatment of the differently occurring reaction types in the generated reaction network. It reads

$$k(T) = \kappa \frac{k_B T}{h} \cdot e^{\Delta G_a / k_B T}, \quad (2)$$

where, k_B is the Boltzmann constant, ΔG_a denotes the free energy of activation, and κ is the transmission coefficient, which is assumed to be unity for all reactions. The errors introduced by ignoring κ are expected to be negligible under the high-temperature conditions, typically several thousand Kelvin. We compared the rate constants obtained using both the RRKM and Eyring approaches for some examples and found good agreement between the two within the QCxMS2 workflow (see ESI,† Section S2 for details). The temperature of the isolated fragment, denoted by T, is estimated using the following approximation

$$T = \frac{E_{int}}{n_{vib} \cdot k_B}, \quad (3)$$

where n_{vib} is the number of harmonic oscillators of the molecule, and E_{int} is its internal energy.[51] For the initial molecule, E_{int} is the impact excess energy (IEE) in the molecule after the ionization process. The energy distribution for the IEE is approximated with a Poisson-type function as in QCxMS

$$P(E) = \frac{\exp[cE(1 + \ln(b/cE)) - b]}{\sqrt{(cE + 1)}}, \quad (4)$$

where $P(E)$ is the probability to have an IEE equal to E. The parameters a, b, and c are given as $\approx$0.2 eV, 1.0 and $\frac{1}{aN_{el}}$, respectively. The maximum value of the IEE is equal to $E_{impact} - \varepsilon_{HOMO}$, where E_{impact} is an input parameter and represents the kinetic energy of the free electron, before impact. The energy of the HOMO, denoted as ε_{HOMO}, is computed by a QM calculation (usually DFT). In this study, E_{impact} amounts to 70 eV in analogy to standard EI experiments, and the distribution is set to an average of about 0.8 eV per atom of the input molecule. This energy distribution was determined through extensive testing in the development of QCxMS and showed to be a good approximation for the usually unknown energy distribution in the experiment.[12] During the development and evaluation of QCxMS2, we found that the energy distributed uniformly over all atoms overestimates the rate constants for reactions involving hydrogen dissociations. Apart from potential errors related to the chosen QM method, this may suggest an inhomogeneous energy distribution at the timescale of these reactions. To address this systematic error, we applied a simple linear scaling factor to adjust the energy distribution specifically for these reactions (see ESI,† Section S5 for details).

For subsequent fragmentations, the internal energy is corrected by the energy loss of a fragment upon dissociation

$$E(\text{fragment}) = (E_0 - \text{KER} - \Delta H) \times n_{at}(\text{frag})/n_{at}(\text{prec}), \quad (5)$$

which consists of the reaction enthalpy ΔH, *i.e.*, including the zero point vibrational energy, and the kinetic energy release (KER). The KER is computed from the respective reaction energy and barrier using empirical parameters derived from experimental studies[52,53] (see ESI,† Section S9 for details). The energy is partitioned between the fragments according to the ratio of the number of atoms in the respective fragment, n_{at}(frag), to the total number of atoms in the precursor ion, n_{at}(prec).

Ion-tracking is conducted as in QCxMS.[12] Molecular charges are distributed according to the ionization potential (IP) of the formed fragments, which are determined by self-consistent field (ΔSCF) calculations at a QM level (usually DFT). The statistical weight of each product is then given by

$$P_i = \frac{\exp\left(\frac{-\Delta E_{SCF,i}}{k_B T_{Av}}\right)}{\sum_j^M \exp\left(\frac{-\Delta E_{SCF,j}}{k_B T_{Av}}\right)}, \quad (6)$$

where M is the number of fragments and $\Delta E_{SCF,i}$ denotes the energy difference between the neutral and charged states of a specific fragment. Negatively or multiply charged species can in principle be described in the same way, as was investigated with QCxMS[13,54] but are not considered in this work. The average temperature of the ion denoted T_{Av}, is estimated using eqn (3) from its average internal energy. The survival yield of a fragment, defined as the ratio of the initial intensity I_0 to the final intensity I, follows the rate law for unimolecular (first-order) reactions

$$\frac{I}{I_0} = e^{-k(E)t}, \quad (7)$$

where $t \approx 50$ ms is the typical time of flight in the spectrometer.[55] For subsequent reactions, the time of flight is corrected by the sum of the half-life of the previous reactions. Eqn (7) holds under the reasonable assumption that the reverse reaction, *i.e.*, the recombination of two dissociated fragments, does not occur. However, for frequently occurring isomerization reactions, this reversibility has to be taken into account, see ESI,† Section S13.

Some fundamental limitations of the QCxMS2 approach remain. Direct bond cleavage, also called non-statistical or nonergodic processes, *i.e.*, reactions occurring at a rate faster than the intramolecular vibrational energy redistribution (IVR) cannot be accounted for. Although these are known to happen in a mass spectrometer[56] they are assumed here to be less important for the computation of a (for typical applications) sufficiently accurate spectrum, and the assumptions of QET hold for most reactions occurring in a mass spectrometer.[29] These reactions can be modeled through dynamical (MD)-based approaches, such as QCxMS, where atomic velocities are scaled non-uniformly to account for the period before the energy is fully equilibrated across the molecule.[57] Quantum tunneling through reaction barriers[58] may also occur but are also assumed to be less relevant, as they mostly happen for subsequent fragmentation on the ns to μs timescale.[56] These effects are expected to cause the largest increase in rate constants for hydrogen dissociations. However, as discussed in

Section 2, we tend to overestimate their rates. Therefore, theoretical models to describe this effect, such as those described in ref. 59, are not considered in QCxMS2, but can in principle be applied for critical cases in the future.

Electronically excited states may also affect the reaction barriers. A study using QCxMS reported improved spectra through the application of excited-state dynamics.[60] We investigated this for the static approach of QCxMS2 by applying time-dependent DFT for the calculation of the reaction barriers in excited states but no improvement for the spectra was observed, as most excited states were found to be hardly populated at the assumed temperatures (see ESI,† S8 for details).

For a more thorough discussion of the mentioned and other less important physical effects, we refer to the excellent review of Dantus[56] of the time-scales of different events in a mass spectrometer observed by time-resolved spectroscopy and Drahos' and Vékey's theoretical work on "Mass Kinetics".[30]

3 Implementation and computational details

The theoretical model described above is implemented in the QCxMS2 program available on GitHub.[61]

QCxMS2 is an advanced script that integrates several external QM codes to fully automate the calculation of an electron ionization mass spectrum. The procedure follows a workflow consisting of seven main steps as shown in Fig. 1, which are detailed in the following sections. Additional technical details can be found in the open-source software code.

3.1 Fragment generation

The input is a coordinate file of a molecule. First, possible fragments of the input molecule are generated with the MSREACT mode of CREST.[62] The critical aspect of this step is to generate a comprehensive set of possible fragments, which can then be ranked based on their relative barrier heights. Fragments with relative energies exceeding three times the average fragment energy are excluded at this stage to save computation time. Fragments that are not generated at this stage will not appear in the final spectrum (see Section 4.1), whereas incorrectly generated fragments typically do not contribute significantly due to their prohibitively high energy barriers. Furthermore, the desired fragment has to be a local minimum on the PES of the employed level of theory, as its geometry is optimized using the respective method, which can potentially lead to (unintended) atomic rearrangements or artifacts of the method. As the fragment generator is applied to each newly formed fragment with significant relative intensity, QCxMS2 calculations typically involve hundreds to thousands of geometry optimizations, and only efficient SQM methods can be applied here. After removing duplicates (see below), the number of fragments is significantly reduced, allowing for the use of more expensive QM methods, *e.g.*, DFT for reoptimization of the unique fragments.

In the fragment (product) generation step with CREST, harmonic repulsive potentials are applied for each atom pair separated by up to three covalent bonds, leading in geometry optimization with GFN2-xTB to typical fragmentation products.[63,64] Additionally, further optimizations are conducted with attractive potentials between hydrogen atoms and potential protonation sites within a default cutoff distance of 4 Å to obtain often observed products due to hydrogen rearrangements. Note that these bias potentials are exclusively employed in the generation step and are not utilized in the subsequent energy and barrier calculations. Next, each obtained product is subsequently optimized in a maximum of 15 cycles without constraints to generate reasonable fragments on the GFN2-xTB PES while avoiding the recombination of the dissociation products. Both optimization steps are conducted at a high finite electronic temperature of 5000 K to favor the generation of open-shell (poly)radicals typically occurring in (EI-)MS. Duplicated structures produced are identified with MolBar[65] and removed to avoid redundant calculations. Additional (random) shifting of atom positions can be employed to generate a greater number of potential products, however, this option is not activated in the default settings.

For more details on this structure generator, we refer to the original publication in ref. 62.

3.2 Transition state search

For each fragmentation or isomerization reaction, a minimum energy path search is performed with the nudged-elastic band (NEB) method[66] as implemented in ORCA 6.0.0.[67,68]

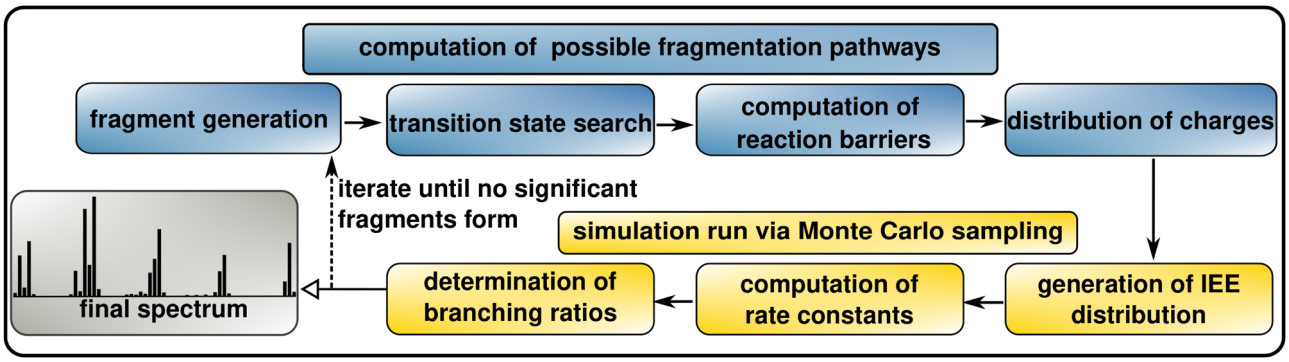


Fig. 1 Schematic representation of the workflow of QCxMS2 for the computation of EI-MS. For details on the computational protocol see Section 3.

Loose convergence criteria are chosen (*Tol_MaxFP_I* = 0.01 and *Tol_RMSFP_I* = 0.005), otherwise default settings (keyword ''NEB'') are used including energy-weighted spring forces. The initial path is generated with the image-dependent pair potential (IDPP) method.[69] Not converged NEB runs are restarted with a different guess for the initial path generated by the geodesic interpolation program.[70] For the transition state optimization, the Hessian of the structure with the highest energy on the minimum energy path is computed at the GFN2-xTB level and the intrinsic reaction coordinate (IRC) mode is identified by comparing the difference in the root-mean-square deviation (RMSD) of the atoms to start and end structure upon translation along each obtained imaginary frequency mode. The transition state optimization is performed along this mode at ''loose'' convergence settings in ORCA.

3.3 Computation of reaction energies and barriers

Reaction barriers and energies can be refined at a higher level of theory by single-point calculations on the previously optimized geometries. In this work, GFN2-xTB and ωB97X-3c were employed. Thermal corrections are accounted for by the single-point Hessian (SPH)[71] approach at the GFN2-xTB level as it is more robust than the conventional approach on not fully optimized structures often exhibiting other small imaginary modes beside the imaginary transition state mode, which may occur in the automated workflow. Low-frequency modes are described with the modified rigid-rotor harmonic oscillator (mRRHO) approximation.[72] Due to the high temperatures, the mRRHO rotor cutoff was set to 150 cm^{-1}. To ensure robustness, imaginary modes with an absolute value below 100 cm^{-1} were inverted. SCF calculations that do not converge with the default settings in ORCA 6.0.0, are restarted with Fermi smearing at elevated electronic temperature to account partially for the potential multireference character of the open-shell radical cations and for the correct dissociation behavior of two-electron bonds. The temperature is chosen according to the HOMO–LUMO gap of the respective QM method as described in ref. 57.

3.4 Distribution of charges

The IPs of the fragments are computed at the neutral optimized structures *via* a ΔSCF approach. By default, a composite level of GFN2-xTB IPs and refinement of close IPs (below 2 eV mol^{-1}) at the ωB97X-3c level of theory is employed.

3.5 Generation of IEE distribution

The energy distribution given by eqn (4) is sampled numerically in a Monte Carlo approach, using by default 10^5 sample points, that lead to sufficient convergence of the relative intensities according to our tests for this quantity. As default, the average energy is set to 0.8 eV times the number of atoms of the input molecule.

3.6 Computation of rate constants

For each energy in this distribution, a rate constant is calculated at the corresponding temperature. The thermal contributions to the reaction barriers are determined at each of these temperatures using the previously computed vibrational frequencies. For computational efficiency, these contributions are precomputed across the energy distribution in 200 discrete steps.

3.7 Determination of branching ratios

Finally, the branching ratios of the fragmentation reactions are calculated from the relative reaction rates. Relative intensities are determined based on the relative reaction rates and the survival yield of the precursor ion across the energy distribution using a Monte Carlo approach. This calculation is conducted separately for each fragmentation step, as the absolute rate constants for subsequent fragmentations are significantly slower than those of earlier steps due to energy loss upon fragmentation. This simplification is performed, as the branching ratios have to be computed for the entire energy distribution, which would be computationally very expensive to perform for a system of coupled differential equations.

The fragment intensities are multiplied by their respective statistical charge computed earlier. Normalization of all computed fragment intensities to the intensity of the largest signal as usual results in the final spectrum.

Note that steps 5–7 are negligible in terms of computational costs compared to steps 1–4, which require QM calculations. This has the advantage that the normally unknown energy distribution can be adapted to the experiment and only any new reaction paths that may arise at higher energies need to be calculated. This is a further advantage over the use of MD trajectories, which have to be completely recalculated for different internal energies.

The natural isotope ratios are introduced in a post-simulation treatment as in QCxMS.[12] Steps 1–7 are performed iteratively for each newly formed fragment with a relative intensity above a certain threshold, which is by default 1% of the initial intensity. Thus, subsequent fragmentations *via* cascade reactions are captured. For a more thorough discussion of the intensity threshold and the reproducibility of the workflow, see ESI,† Section S6.

3.8 Employed programs

The results discussed in Section 4 were computed with QCxMS2 version 1.0.0 with default settings.[61] As input, the minimum energy conformer of the radical cation of the molecule at the GFN2-xTB level found by CREST version 3.0.2[73] was used as a starting point. Fragments were generated with a development version of the CREST MSREACT mode and duplicates were identified with molbar 1.0.3.[74] ωB97X-3c calculations, NEB path searches, and transition state optimizations were performed with ORCA version 6.0.0.[67] The resolution of identity approximation[75] with matching auxiliary basis sets was applied for the Coulomb integrals,[76] whereas the exchange integrals were computed analytically, as it is faster than RIJCOSX[77] for the small system sizes investigated here. Geometry optimizations of equilibrium structures at the ωB97X-3c level were performed in ORCA with ''loose'' convergence settings. Initial reaction paths for restarted NEB calculations were generated

with geodesic-interpolation 1.0.0.[78] GFN2-xTB calculations were conducted with a development version of xTB 6.7.1 with default convergence settings.[79] QCxMS spectra for comparison were computed with QCxMS V5.2.1[13,80,81] with default settings at the GFN2-xTB level. Cosine similarity matching scores[82,83] were computed with matchms python package[84] and entropy similarity scores[85,86] with the msentropy python package.[87]

4 Results

In this section, standard 70 eV EI-MS spectra computed with QCxMS2 are shown for a set of 16 organic and inorganic main group molecules listed in Table 1. No system-specific adjustments were made in the calculation of spectra to evaluate QCxMS2's potential for cases with unknown experimental data. Additional investigations for the rotor-cutoff (ESI,† Section S7) and the average internal energy (ESI,† Section S4) parameters were made at the composite level ωB97X-3c//GFN2-xTB (see below) and can be found in the ESI.† For comparison, experimental spectra rounded to integer masses of all compounds are taken from the NIST database,[88] except for acibenzolar-*S*-methyl, for which a high-resolution spectrum was taken from ref. 25. With this selection of molecules we intend to discuss the strengths and weaknesses of the approach. The test set comprises a diverse range of organic and inorganic compounds, including the alkane *n*-octane, alkene 4-methyl-1-pentene, ether ethyl propyl ether, alcohol 1-butanol, aldehyde butanal, ketone 2-pentanone, carboxylic acid butanoic acid, ester methyl butyrate, amide butanamide, and N-heterocycles uracil, adenine, and caffeine. Additionally, main group inorganic substances such as tabun, tetramethylbiphosphine disulfide, acibenzolar-*S*-methyl, and dichloroethylaluminum are included. Lewis structures of all compounds can be found in the ESI,† Section S3. In principle, QCxMS2 can also compute molecules containing transition metals without special adjustments. However, due to their often rather special fragmentation patterns and generally more difficult electronic structure compared to the main group elements, they are omitted from this study and are planned for a later study.

To ease the assessment of the quality spectra, the spectral entropy matching score is used. It captures the presence of relevant peaks, as well as their relevant intensities compared to the experiment, and ranges from 0 (no agreement at all) to 1 (perfect agreement).[85] It was recently shown[85] that this score is more reliable than the commonly used cosine similarity score[82,83] and it is in our opinion a good metric to evaluate the accuracy of the spectra in this work. For comparison, the average values of the cosine score are also given in the discussion below. Herein, a score of at least 0.75 between experimental spectra was found to be a meaningful threshold for reliable structure identification[85] and should be aimed for with any theoretical procedure considering the uncertainty of the experiment. However, interpretation of this score is system-specific, *e.g.*, the most important peaks for substance identification may be present despite a comparatively low score.

Spectra were computed with three different combinations of QM methods, given in the short notation ''method used for reaction energies and barriers''//''method used for geometry optimizations and reaction path searches'', namely, GFN2-xTB//GFN2-xTB (''GFN2-xTB''), ωB97X-3c//GFN2-xTB (''composite''), and ωB97X-3c//ωB97X-3c (''ωB97X-3c''). IPs were calculated at the GFN2-xTB level and refined at the ωB97X-3c level as described above and for the DFT spectra only with ωB97X-3c calculated throughout. Harmonic vibrational frequencies were always computed with GFN2-xTB. The RSH ωB97X-3c was employed because it yields excellent barriers at low computational costs[89] and is considered by us as one of the best yet still affordable methods in our context. For comparison, we computed spectra with QCxMS at the GFN2-xTB level, as the refinement of energies and performing MDs at the ωB97X-3c level is computationally not feasible (see Section 4.3).

Table 1 Entropy similarity spectral match scores between experimental and theoretical spectra computed with QCxMS2 at the GFN2-xTB//GFN2-xTB, ''composite'' ωB97X-3c//GFN2-xTB, and ωB97X-3c//ωB97X-3c levels for all compounds of the test set. Values for spectra computed with QCxMS at the GFN2-xTB level are also given for comparison

Compound	GFN2-xTB	Composite	ωB97X-3c	QCxMS
n-Octane	0.686	0.703	0.841	0.840
4-Methyl-1-pentene	0.758	0.714	0.835	0.782
Ethyl propyl ether	0.762	0.869	0.813	0.697
1-Butanol	0.753	0.750	0.724	0.603
Butanal	0.852	0.807	0.807	0.803
2-Pentanone	0.781	0.718	0.818	0.743
Butanoic acid	0.683	0.751	0.761	0.558
Methyl butyrate	0.635	0.736	0.742	0.655
Butanamide	0.494	0.673	0.674	0.620
Uracil	0.644	0.498	0.659	0.769
Adenine	0.748	0.790	0.712	0.794
Caffeine	0.456	0.626	0.644	0.626
Tabun	0.637	0.508	0.655	0.649
Tetramethylbi-phosphine disulfide	0.691	0.796	0.782	0.269
Acibenzolar-*S*-methyl	0.389	0.599	0.527	0.100
Dichloroethyl-aluminium	0.752	0.667	0.679	0.438
Average	0.670	0.700	0.730	0.622
Minimum	0.389	0.599	0.527	0.100

4.1 Effect of the level of theory

First, we discuss the effect of the level of theory used for the spectra calculation. Entropy similarity match scores between experimental and theoretical spectra computed with QCxMS and QCxMS2 with the three method combinations described above for all 16 compounds of the test set are given in Table 1.

On average, the highest level of theory employed, *i.e.*, ωB97X-3c for geometries and energies, achieves a very good score of 0.73. Seven out of 16 compounds achieve the target accuracy of at least 0.75, while only four compounds, namely butanamide, uracil, tabun, and acibenzolar-*S*-methyl, exhibit a mediocre score below 0.7. As expected, using GFN2-xTB geometries instead of DFT geometries results in a slight decrease in accuracy with a still good average score of 0.7. When GFN2-xTB reaction barriers are used instead of ωB97X-3c, the accuracy drops to 0.67. The still good accuracy of GFN2-xTB is somewhat unexpected, considering its known limitations in accurately modeling radical cations and reaction barriers.[63]

Despite some outliers, for which GFN2-xTB or the composite level yields better results than ωB97X-3c presumably due to favorable error compensation, the trend is on average that a more accurate description of the PES leads to better spectra. This is an important observation and supports the underlying theoretical assumptions of the QCxMS2 approach.

For comparison, the commonly used cosine similarity score (see ESI,† S12 for scores for each compound), shows an even more pronounced trend with scores of 0.573 (GFN2-xTB//GFN2-xTB), 0.636 (ωB97X-3c//GFN2-xTB), and 0.711 (ωB97X-3c//ωB97X-3c).

The effect on the spectrum by refining the barriers at the ωB97X-3c level is exemplary shown in two examples. Fig. 2 depicts the spectra computed with QCxMS2 at the GFN2-xTB//GFN2-xTB and ωB97X-3c//GFN2-xTB levels for 2-pentanone and caffeine. For 2-pentanone, only very small differences between the spectra are visible and both show a good agreement with the experiment with matching scores of 0.781 and 0.718, respectively. Here, GFN2-xTB gives already a good description of the PES, and no refinement of the barriers is needed. The ωB97X-3c spectrum, shown in the ESI,† in Section S16, looks slightly better, as the peaks at *m*/*z* 57 and 29 are computed much smaller yielding excellent good matching score of 0.818. The spectrum of caffeine computed with GFN2-xTB agrees poorly with experiment. Although many relevant peaks are present, they have incorrect relative intensities which leads to a low matching score of only 0.456. By using ωB97X-3c reaction barriers, a substantial improvement to a score of 0.626 is obtained. However, the peak at *m*/*z* 109 has too low relative intensity, while the peaks at *m*/*z* 110 and *m*/*z* 111 are obtained with too high intensities. Additionally, the peak at *m*/*z* 55 is missing. Computing the spectrum using ωB97X-3c also for geometries further improves the agreement with experiment and yields a score of 0.644 (spectrum shown in the ESI,† in Section S16).

Next, we examine the effect on the spectra using ωB97X-3c geometries instead of GFN2-xTB geometries, as depicted in Fig. 3, with the examples of 4-methyl-1-pentene and uracil. The spectrum of 4-methyl-1-pentene computed with GFN2-xTB geometries generally shows good agreement with the experiment yielding a reasonable score of 0.714. However, several signal intensities are inaccurate, particularly the peaks at *m*/*z* 68, 57, and 53. When using ωB97X-3c geometries instead, the spectrum shows almost perfect agreement with the experiment with a matching score of 0.835, and the base peak is also correctly predicted to be at *m*/*z* 43. This suggests that the respective transition state geometries optimized at the GFN2-xTB level are insufficient for refinement at the ωB97X-3c level, and accurate relative barrier heights are only achieved when ωB97X-3c is also used for the geometry optimization.

An even more pronounced example for this observation is uracil. Here, the agreement with the experiment with GFN2-xTB optimized geometries is rather bad, as the peak *m*/*z* 84 is falsely predicted leading to a match score of only 0.498. This is due to a too "flat" PES of GFN2-xTB for the initial dissociation of CO leading to a wrong transition state structure too late at the reaction path and thus to an underestimated barrier for the peak at *m*/*z* 84. The apparently correct peak at *m*/*z* 41 stems from further dissociation of this fragment and is therefore predicted here only by chance but *via* a wrong pathway. As a result, the other correctly predicted peaks are consequently too low in intensity demonstrating the sensitivity of the approach, as one inaccurate barrier can potentially distort the whole spectrum. In contrast, the *m*/*z* 84 peak is virtually absent when using ωB97X-3c optimized reaction paths which gives the correct description for the CO dissociation leading to an overall much better agreement with the experimental spectrum (score of 0.659 *versus* 0.470). Overall, QCxMS2 demonstrates good accuracy, however, certain spectra exhibit low matching scores even at the ωB97X-3c level.

For example, in the ωB97X-3c spectrum of uracil, the signals at *m*/*z* 40, and 42 are missing and the peak at *m*/*z* 41 has a too low intensity as the competing fragmentation pathway to the peak at *m*/*z* 28 has a lower barrier. For *m*/*z* 42, we rationalized

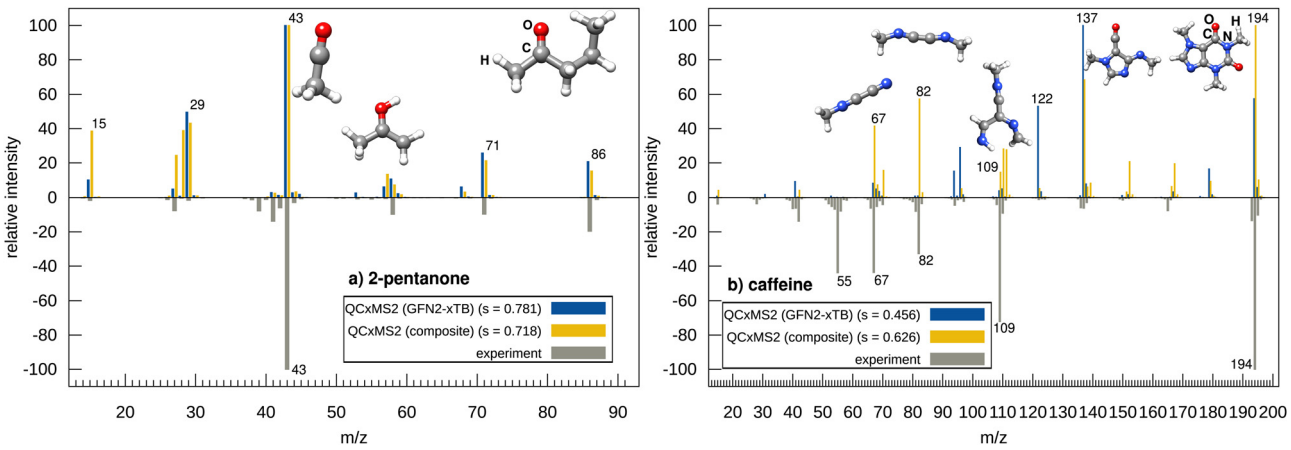


Fig. 2 Calculated spectra with QCxMS2 at the GFN2-xTB level and "composite" (ωB97X-3c//GFN2-xTB) levels of theory compared to the inverted experimental spectrum of (a) 2-pentanone, (b) caffeine. All spectra were rounded to integer masses, and peak positions in the theoretical spectra were shifted by 0.25 *m*/*z* units for better visibility. The entropy similarity score is denoted by s.

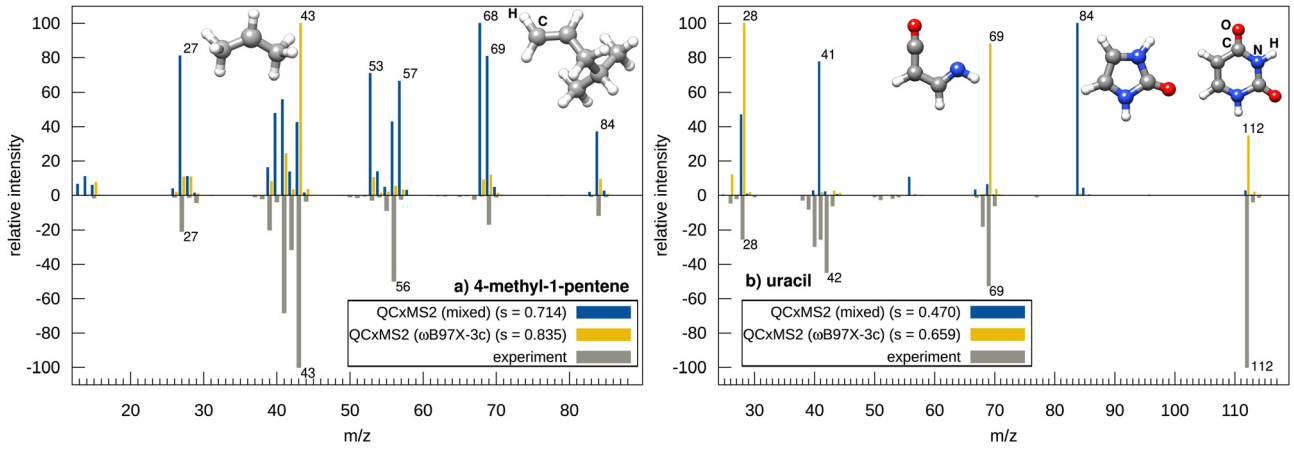


Fig. 3 Calculated spectra with QCxMS2 at the "composite" (ωB97X-3c//GFN2-xTB), and ωB97X-3c levels of theory compared to the inverted experimental spectrum of (a) 4-methyl-1-pentene and (b) uracil. All spectra were rounded to integer masses and masses of theoretical spectra were shifted by ±0.25 *m*/*z* units for better visibility. The entropy similarity score is denoted by s.

that the peak most likely stems from a hydrogen rearrangement of the fragment at *m*/*z* 69 to a ketene and subsequent loss of HCN. However, this fragmentation reaction is not predicted by the MSREACT fragment generator.

By modeling the reaction path of the H-rearrangement and the HCN loss manually we found a sufficiently low barrier for the fragment to be formed along the fragment at *m*/*z* 28.

The fragment with *m*/*z* 40 is also not formed by MSREACT and could not be identified manually. It may be not generated as it is very high in energy on the GFN2-xTB PES. The fact that the peak at *m*/*z* 40 is completely missing in the GFN2-xTB spectrum of QCxMS (shown in the ESI,† in Section S15) indicates that the fragment is not easily accessible on the GFN2-xTB PES.

To assess if MSREACT has in general problems in predicting the fragments of rigid ring systems, we computed additional spectra with ωB97X-3c//GFN2-xTB for the simplest representatives of this class, namely benzene and naphthalene and found also poor agreement with scores of 0.699 and 0.698, respectively. Notably, many peaks are missing in the computed spectra.

Using ωB97X-3c for geometry optimizations in QCxMS2 does not lead to better results for this class of compounds.

In addition, we computed the spectra at the ωB97X-3c//GFN2-xTB level with additional geometry optimizations after randomly shifting the atomic positions in the fragment generator. By applying these settings, more peaks observed in the experiment are correctly predicted but the overall accuracy of the spectrum does not increase as also more wrong intensities are predicted.

We conclude that the above described problems with conjugated π-ring systens is mainly due to the insufficient description of the PES of the formed fragments by GFN2-xTB. During the constrained optimizations in MSREACT, (unintended) atom rearrangements frequently occur, leading to the generation of numerous artificial structures. Refinement at the DFT level cannot resolve this issue, as the correct fragments are not generated at the GFN2-xTB stage. Employing a higher-level theory, such as ωB97X-3c, in MSREACT is computationally infeasible, as outlined in Section 3.1.

In contrast, QCxMS at the GFN2-xTB level produces reasonably accurate spectra, with scores of 0.895 and 0.826 for benzene and naphthalene. Here, the limited accuracy of GFN2-xTB does not appear to be as critical, as the presumed artifacts coincidentally align with the correct experimental masses. However, substituted benzenes and phenols also prove to be challenging for QCxMS.[19,90]

Another issue observed in the QCxMS2 spectra is that, for larger or more complex molecules, errors of ±1, *m*/*z* may occasionally occur. This indicates that a hydrogen atom is incorrectly assigned to the other fragment of the dissociation products in the respective fragmentation reaction compared to the experimental results. Such cases are observed, for instance, around the peak at *m*/*z* 109 for caffeine, as described above, or the missing peak at *m*/*z* 181 for acibenzolar-*S*-methyl (see below). Hydrogen dissociations are generally difficult to describe, as indicated by the scaling factor applied to the internal energy distribution for these reactions.

Another source of error are the ro-vibrational thermal contributions, which are computed only at the GFN2-xTB level, also due to computational costs. We investigated their effect for the spectrum of methyl-butyrate, for which we could achieve an improvement of the spectrum by using ωB97X-3c frequencies instead of GFN2-xTB frequencies (see ESI,† S10 for details).

Despite these problems mainly due to the limited accuracy of the (currently) feasible level of electronic structure theory, QCxMS2 shows overall good robustness, which is also reflected in the minimum score of 0.527 at the ωB97X-3c level for the test set. Taking into account that the set also contains complicated molecules with unusual fragmentation pathways, this is a good result.

4.2 Comparison to QCxMS

Next, we discuss the accuracy of QCxMS2 in comparison to its predecessor QCxMS. Overall on the test set, QCxMS2 at the ωB97X-3c level yields an average match score of 0.735 compared to 0.622 for the QCxMS spectra computed at the

GFN2-xTB level. Employing the cosine similarity score, the difference is even larger with values of 0.755 (QCxMS2) and 0.515 (QCxMS). Notably, the lowest score with QCxMS2 was 0.527 compared to only 0.100 of QCxMS, indicating that QCxMS2 is the more robust approach yielding less outliers, which is important for application in automated structure elucidation workflows. Already at GFN2-xTB//GFN2-xTB level QCxMS2 spectra are more accurate than QCxMS with an average score of 0.673. This indicates that the new "static" approach has intrinsic advantages over MD based QCxMS, which will be investigated in the following for four selected molecules that were identified as problematic for QCxMS.

Fig. 4 shows the computed spectra for ethyl propyl ether, butanoic acid, tetramethylbiphosphine disulfide, and acibenzolar-*S*-methyl using both QCxMS2 and QCxMS. For comparison, we take the best but still affordable level for QCxMS2, *i.e.*, ωB97X-3c. Since computing spectra at this level is unfeasible in the QCxMS (see Section 4.3), we take here the spectra computed with the GFN2-xTB level of theory in the comparison. For ethyl propyl ether, the base peak at *m*/*z* 31 is significantly computed only by QCxMS2 and is virtually absent in the QCxMS spectrum, which fails to predict this rearrangement reaction from the fragment with *m*/*z* 59. This is a typical reaction occurring for ethers and an important finding that this signal is obtained with QCxMS2. Consequently, the matching score with QCxMS2 is much better (0.813 *versus* 0.697). A similar cases is butanoic acid, for which the Mclafferty type rearrangement resulting in the peak at *m*/*z* 60, which is also the main peak in the experimental spectrum, occurs with a too low probability with QCxMS. Also here, QCxMS2 computed this fragment in good agreement with the intensity from the experiment, yielding also a much better score of 0.761 *versus* 0.558. Even when using GFN2-xTB with QCxMS2, improved scores compared to QCxMS are achieved for ethyl propyl ether (0.762) and butanoic acid (0.683). These two examples demonstrate, that fragments stemming from rearrangements are underestimated in QCxMS, probably due to the limited MD simulation time.

Examples (c) and (d) in Fig. 4 contain the inorganic main group elements P and S, which were also found to be problematic for QCxMS in a recent study.[25] The QCxMS computed spectrum for tetramethylbiphosphine disulfide shows poor agreement (score of 0.269), failing to predict the methyl dissociation to the peak at *m*/*z* 171 and the P–P bond breakage leading to the main peak at *m*/*z* 93 in the experimental

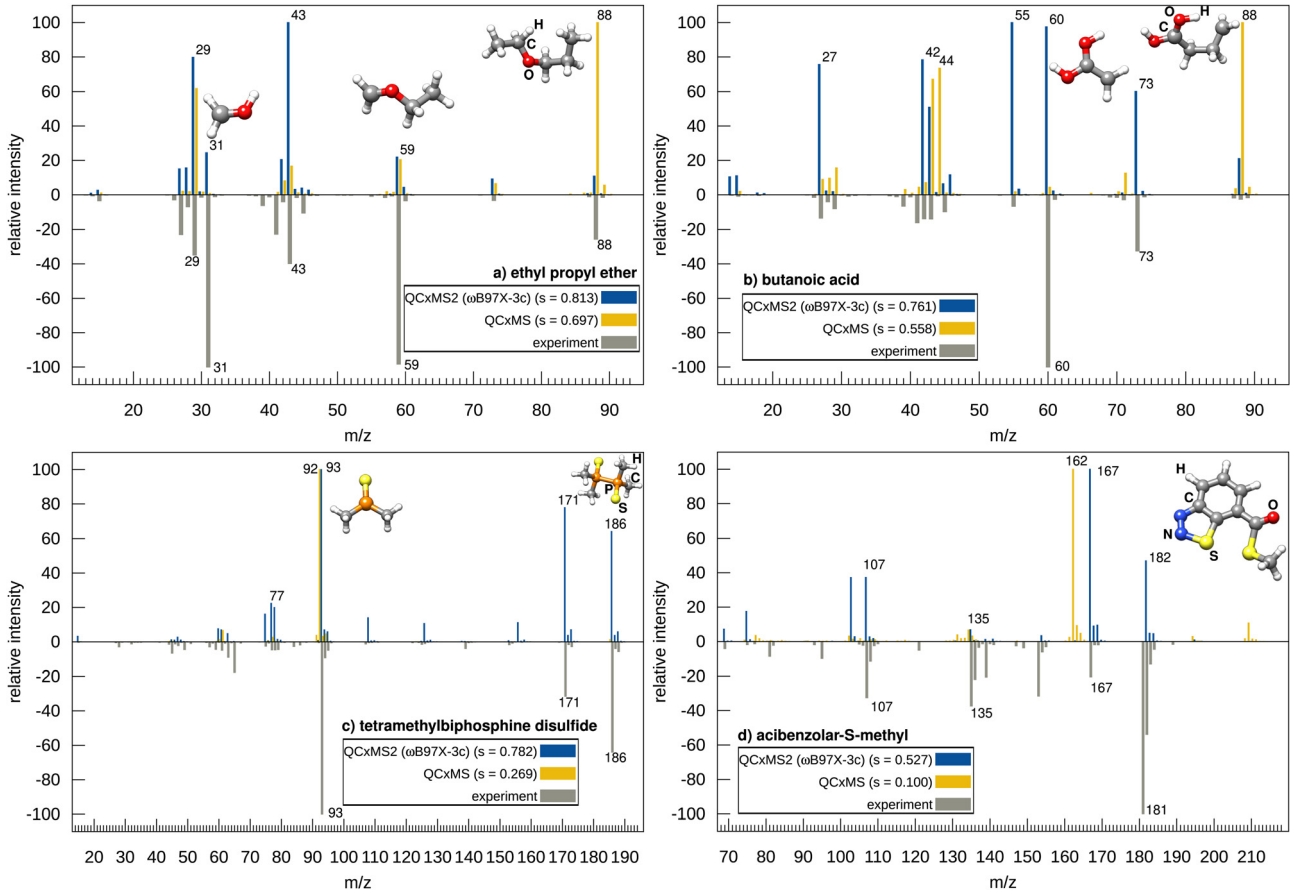


Fig. 4 Calculated spectra with QCxMS at the GFN2-xTB level and QCxMS2 at the ωB97X-3c//ωB97X-3c level, compared to inverted experimental spectrum of (a) ethyl propyl ether, (b) butanoic acid, (c) tetramethylbiphosphine disulfide, (d) acibenzolar-*S*-methyl. All spectra were rounded to integer *m*/*z* values and masses of theoretical spectra were shifted by ±0.25 *m*/*z* units for better visibility. The molecular structures shown are the fragments with the highest intensity of the respective *m*/*z* signal of the QCxMS2 spectrum. The entropy similarity score is denoted by s.

spectrum. Instead, QCxMS predicts an additional H-shift associated with the P–P bond breakage, resulting in an experimentally unobserved peak at m/z 92. Thus, the issue of missing peaks by ± 1 m/z unit, as described above for QCxMS2, also occurs with QCxMS. In contrast, QCxMS2 correctly predicts here the bond breakage and achieves a much higher score of 0.782. As the m/z 92 peak is also observed in the GFN2-xTB spectrum of QCxMS2 with a score of 0.691 (spectrum shown in the ESI,† in Section S16), the falsely predicted hydrogen shift in QCxMS is probably due to the MD approach and not due to the inaccuracy of GFN2-xTB.

For acibenzolar-*S*-methyl, the QCxMS spectrum shows almost no agreement to experiment, with a score of only 0.100. The peak at m/z 182, resulting from the loss of N_2, is not found, and instead, a false peak at m/z 162 is computed as the main peak. This peak arises from an α-cleavage, involving an H-shift to the sulfur atom and dissociation of methanethiol at the carbonyl C-atom. Interestingly, this peak is observed in the GFN2-xTB spectrum of QCxMS2 (shown in the ESI,† in Section S16) (however, without hydrogen shift, *i.e.*, resulting in a peak at m/z 163), indicating that the GFN2-xTB PES overestimates the stability of the thiadiazole ring.

Using QCxMS2 in conjunction with ωB97X-3c, the loss of N_2 to the peak at m/z 182 is correctly computed. However, the main peak at m/z 181 is also missing here. Due to its high intensity in the experimental spectrum, this stems most probably not from a hydrogen dissociation from the fragment of m/z 182 and has to occur *via* a different mechanism, as the computed barrier of the hydrogen dissociation is much too high (even without scaling of the IEE applied) compared to the methyl loss to the fragment of m/z 167. The fragment generator does not produce the correct fragment here, probably due to the insufficient accuracy of GFN2-xTB as discussed in Section 4.1. However, apart from the main experimental peak, the relevant signals are obtained and the score of 0.527 is still reasonable compared to QCxMS. Overall, the results for the test set demonstrate that QCxMS2 exhibits improved accuracy and robustness in comparison to QCxMS.

4.3 Computation time

Finally, the computational timings are discussed using the examples of 2-pentanone, a molecule with 16 atoms, and caffeine, a typical metabolite with 24 atoms and the largest molecule in the test set.

Fig. 5 shows the computational timings scaled to 16 Intel Xeon "Sapphire Rapids" v4 @ 2.10 GHz CPU cores for the spectra calculation with QCxMS2 with the three different theory levels employed here and timings with QCxMS with GFN2-xTB and ωB97X-3c in comparison.

QCxMS can in principle be perfectly parallelized as every (cascading) trajectory is obtained separately, whereas with QCxMS2 the parallelization efficiency depends on the number of fragments in a fragmentation step and, how long particular calculations, *e.g.*, a specific transition state search takes since some calculations have to be performed in a subsequent manner. The QCxMS2 calculations were performed with

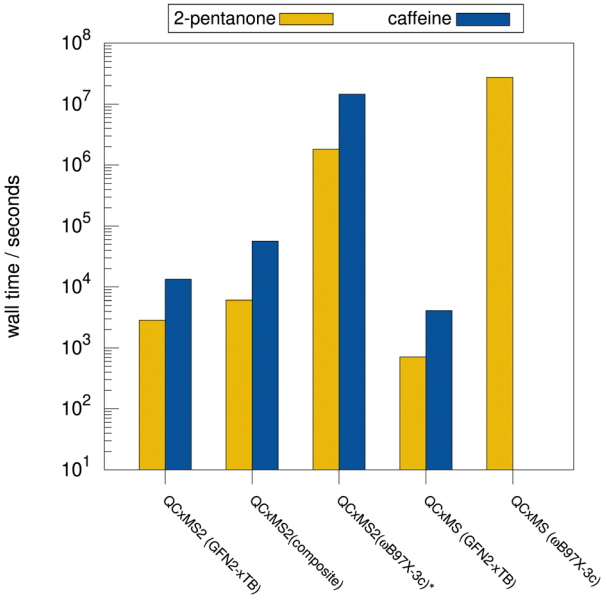


Fig. 5 Computational wall times on 16 Intel Xeon "Sapphire Rapids" v4 @ 2.10 GHz CPU cores for the calculation of 2-pentanone and caffeine with QCxMS2 using GFN2-xTB//GFN2-xTB, the "composite" level ωB97X-3c//GFN2-xTB, ωB97X-3c//ωB97X-3c, and QCxMS with GFN2-xTB.*: calculation performed on AMD EPYC 7763 CPUs.

16 CPU cores, with the exception of the expensive ωB97X-3c spectra calculations, which were performed on 96 cores for 2-pentanone and on an 128 AMD EPYC 7763 CPU for caffeine. The respective computational timings are scaled to 16 CPU cores. For the QCxMS calculations, the same number of cores was used as the number of trajectories. However, for a meaningful comparison in terms of the practical use of the program, the timings are scaled to the typical computational resources of 16 CPU cores. A QCxMS2 calculation for 2-pentanone at the GFN2-xTB level takes about an hour. Refining the barriers at the ωB97X-3c level takes only one hour more computation time. Computing the geometries and the reaction path search also at the ωB97X-3c level is very expensive and increases the computational costs massively to 502 hours. For caffeine, the computation time is as expected significantly larger as also more fragments have to be computed. Whereas in the calculation for 2-pentanone 79 isomers and 121 fragment pairs and hence 200 transition state searches and barrier calculations have to be performed, for caffeine 462 isomers and 292 fragment pairs were found, *i.e.*, 754 reaction barriers have to be computed. However, the calculation is still feasible, requiring 3.7 hours for the GFN2-xTB calculation and 15.7 hours if the barriers are refined at the ωB97X-3c level. Computing the geometries at the ωB97X-3c level for caffeine, however, becomes impractically expensive, requiring about 4050 hours.

In comparison, the QCxMS calculation for 2-pentanone at the GFN2-xTB level takes 30 minutes using the default number of 400 trajectories for this molecular size. For caffeine, 600 trajectories have to be computed leading to an overall wall time of about one hour. However, refinement of the geometries at a

higher level of theory as in QCxMS2 is not possible in this approach and to reach more accuracy all calculations need to be performed at the higher level of theory too. This is computationally very expensive, as demonstrated by the 2-pentanone calculation, which takes 7664.5 hours at the ωB97X-3c level, which is too expensive to be of practical use. Similarly, the corresponding calculation for caffeine is not feasible with our available resources and was therefore not performed. While QCxMS is computationally cheaper at the GFN2-xTB level, achieving better accuracy using a higher level of theory quickly becomes unfeasible. In contrast, refining barriers *via* DFT single-point calculations is possible with QCxMS2 and improves the accuracy (see Table 1). Even when using ωB97X-3c also for geometry optimizations, QCxMS2 remains computationally more efficient than QCxMS.

5 Conclusions

Computational tools for predicting mass spectra are of great importance for elucidating the chemical structure of unknown compounds. QCxMS, currently the only fully automated QM-based program for calculating EI-MS spectra, achieves reasonable accuracy but faces limitations in generating accurate spectra with less ''outliers'' needed for application in structure elucidation workflows. To this end, a new program termed QCxMS2 for the calculation of mass spectra based on automated reaction network discovery using QM methods was developed. In this work, we demonstrate that the approach of computing spectral intensities from relative reaction rate constants in an automated workflow is generally possible. We presented promising results for a diverse test set of 16 organic and inorganic compounds. Here, QCxMS2 yields a good entropy similarity match score compared to experiment of 0.67 and improves upon its predecessor QCxMS with 0.622 at the same GFN2-xTB level of theory. We recommend refining the barriers *via* single-point calculations with ωB97X-3c on the GFN2-xTB geometries, which yields an improved score of 0.7 at still feasible computational costs. Using ωB97X-3c also for geometries yields even better accuracy and robustness with an average score of 0.73 but at significantly higher computational costs.

We attribute the remaining deviations from the experimental data primarily to errors in the methods used to calculate the electronic barriers, the vibrational contributions, and the possible structures appearing in the reaction networks. Due to the large size of these networks, we are limited to using efficient DFT methods for the energy calculations and SQM methods for the frequency calculations.

The CREST MSREACT fragment generator found in most cases all relevant peaks, and only in a few instances, particularly involving complex unsaturated ring rearrangements, missing fragments are suspected as a source of error. This issue is likely due to the limited accuracy of GFN2-xTB used in this step, and we anticipate that employing improved SQM methods will significantly reduce this error. We plan to investigate the issue of missing peaks in more detail in future studies.

For flexible structures, particularly those containing heavy main-group elements, QCxMS2 demonstrated excellent accuracy on average, significantly outperforming QCxMS. Additionally, typical rearrangement reactions of common organic functional groups are better captured by QCxMS2 than with QCxMS.

The QCxMS2 program is open-source and freely available.[61] Note that all of the employed programs in the QCxMS2 workflow are open-source or at least free for academic use (ORCA) making QCxMS2 free to use for academia. Furthermore, QCxMS2 is systematically improvable and a more ''controlled'' approach than the MD-based QCxMS. We expect that QCxMS2 will benefit especially from newly developed QM methods for the computation of reaction pathways and barriers. Currently, efficient tight-binding methods are being developed in our lab and have already shown promising results close to the accuracy of DFT spectra at significantly reduced computation time. Initial tests with the unpublished g-xTB method currently developed in our lab employed for energies and geometries gave on average an excellent matching score of 0.736 close to the ωB97X-3c values at about the same computation time needed for the GFN2-xTB spectra.

Furthermore, an extension of QCxMS2 to describe negatively or multiply charged species, as well as to calculate the experimentally also very relevant ESI/CID-MS, is planned. Since the QCxMS2 approach can be systematically improved with more advanced methods, we view it as a promising pathway toward highly accurate and reliable mass spectrum predictions.

Author contributions

Johannes Gorges: conceptualization (supporting); data curation (lead); methodology (equal); software (lead); writing – original draft (lead); writing – review & editing (equal). Stefan Grimme: conceptualization (lead); methodology (equal); writing – original draft (supporting); writing – review & editing (equal).

Data availability

The employed software code is open-source and can be found here: **https://github.com/grimmelab/QCxMS2**. Geometries in XYZ format for all structures, as well as the computed spectra for the test set including the experimental reference data taken from the literature, can be found here: **https://github.com/grimme-lab/QCxMS2-data/**. Additional details on the implementation, tests of different technical parameters, and computed spectra not included in the manuscript are provided in the file SI.pdf.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported by the DFG grant no. 533949111, ''Quantum Chemical Calculation of Mass Spectrometry *via*

Automated Transition State Search''. We gratefully acknowledge the access to the Marvin cluster of the University of Bonn. We thank Dr Jeroen Koopman from FACCTs, Thomas Froitzheim, and Julia Kohn for fruitful discussions, and Jasmin Klotz for testing the program. Dr Hagen Neugebauer and Dr Andreas Hansen are thanked for proofreading the manuscript. We further thank Nils van Staalduinen from Aachen for providing a development version of MolBar.

Notes and references

1 G. L. Glish and R. W. Vachet, *Nat. Rev. Drug Discovery*, 2003, **2**, 140–150.
2 R. R. da Silva, P. C. Dorrestein and R. A. Quinn, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 12549–12550.
3 W. D. Hoffmann and G. P. Jackson, *Annu. Rev. Anal. Chem.*, 2015, **8**, 419–440.
4 L. O. Dragsted, Q. Gao, A. Scalbert, G. Vergères, M. Kolehmainen, C. Manach, L. Brennan, L. A. Afman, D. S. Wishart, C. Andres Lacueva, M. Garcia-Aloy, H. Verhagen, E. J. M. Feskens and G. Praticò, *Genes Nutr.*, 2018, **13**, 8.
5 X. Zhou, M. M. Ulaszewska, C. Cuparencu, C. De Gobba, N. Vaazquez-Manjarrez, G. Gürdeniz, J. Chen, F. Mattivi and L. O. Dragsted, *J. Agric. Food Chem.*, 2020, **68**, 6122–6131.
6 L. van Tetering, S. Spies, Q. D. Wildeman, K. J. Houthuijs, R. E. van Outersterp, J. Martens, R. A. Wevers, D. S. Wishart, G. Berden and J. Oomens, *Commun. Chem.*, 2024, **7**, 30.
7 J. N. Wei, D. Belanger, R. P. Adams and D. Sculley, *ACS Cent. Sci.*, 2019, **5**, 700–708.
8 M. Murphy, S. Jegelka, E. Fraenkel, T. Kind, D. Healey and T. Butler, *arXiv*, 2023, preprint, arXiv:2301.11419, DOI: 10.48550/arXiv.2301.11419.
9 F. Allen, R. Greiner and D. Wishart, *Metabolomics*, 2015, **11**, 98–110.
10 S. Goldman, J. Li and C. W. Coley, *Anal. Chem.*, 2024, **96**, 3419–3428.
11 P. L. Bremer, A. Vaniya, T. Kind, S. Wang and O. Fiehn, *J. Chem. Inf. Model.*, 2022, **62**, 4049–4056.
12 S. Grimme, *Angew. Chem., Int. Ed.*, 2013, **52**, 6306–6312.
13 J. Koopman and S. Grimme, *J. Am. Soc. Mass Spectrom.*, 2022, **33**, 2226–2242.
14 J. Lee, D. J. Tantillo, L.-P. Wang and O. Fiehn, *J. Chem. Inf. Model.*, 2024, **64**, 7470–7487.
15 X. Hu, W. L. Hase and T. Pirraglia, *J. Comput. Chem.*, 1991, **12**, 1014–1024.
16 U. Lourderaj, R. Sun, S. C. Kohale, G. L. Barnes, W. A. de Jong, T. L. Windus and W. L. Hase, *Comput. Phys. Commun.*, 2014, **185**, 1074–1080.
17 V. Ásgeirsson, C. A. Bauer and S. Grimme, *Chem. Sci.*, 2017, **8**, 4879–4895.
18 J. Koopman and S. Grimme, *ACS Omega*, 2019, **4**, 15120–15133.
19 P. R. Spackman, B. Bohman, A. Karton and D. Jayatilaka, *Int. J. Quantum Chem.*, 2018, **118**, e25460.
20 S. Wang, T. Kind, D. J. Tantillo and O. Fiehn, *J. Cheminf.*, 2020, **12**, 63.
21 S. A. Schreckenbach, J. S. Anderson, J. Koopman, S. Grimme, M. J. Simpson and K. J. Jobst, *J. Am. Soc. Mass Spectrom.*, 2021, **32**, 1508–1518.
22 R. Schnegotzki, J. Koopman, S. Grimme and R. D. Süssmuth, *Chem. – Eur. J.*, 2022, **28**, e202200318.
23 F. C. Chernicharo, L. Modesto-Costa and I. Borges Jr, *J. Mass Spectrom.*, 2020, **55**, e4513.
24 S. Grimme, C. Bannwarth and P. Shushkov, *J. Chem. Theory Comput.*, 2017, **13**, 1989–2009.
25 H. Hecht, W. Y. Rojas, Z. Ahmad, A. Krenek, J. Klánová and E. J. Price, *Anal. Chem.*, 2024, **96**, 13652–13662.
26 L. S. Kassel, *J. Phys. Chem.*, 1927, **32**, 225–242.
27 O. K. Rice and H. C. Ramsperger, *J. Am. Chem. Soc.*, 1927, **49**, 1617–1629.
28 R. A. Marcus, *J. Chem. Phys.*, 1952, **20**, 359–364.
29 H. M. Rosenstock, M. B. Wallenstein, A. L. Wahrhaftig and H. Eyring, *Proc. Natl. Acad. Sci. U. S. A.*, 1952, **38**, 667–678.
30 L. Drahos and K. Vékey, *J. Mass Spectrom.*, 2001, **36**, 237–263.
31 C. Lifshitz, *Acc. Chem. Res.*, 1994, **27**, 138–144.
32 D. Asakawa, *J. Am. Soc. Mass Spectrom.*, 2023, **34**, 435–440.
33 D. Asakawa, K. Todoroki and H. Mizuno, *J. Am. Soc. Mass Spectrom.*, 2022, **33**, 1716–1722.
34 D. Lesage, S. Mezzache, Y. Gimbert, H. Dossmann and J.-C. Tabet, *J. Am. Soc. Mass Spectrom.*, 2019, **25**, 219–228.
35 C. Chalet, D. Lesage, E. Darii, A. Perret, S. Alves, Y. Gimbert and J.-C. Tabet, *J. Am. Soc. Mass Spectrom.*, 2024, **35**, 456–465.
36 C. A. Bauer and S. Grimme, *J. Phys. Chem. A*, 2016, **120**, 3755–3766.
37 J. P. Unsleber and M. Reiher, *Annu. Rev. Phys. Chem.*, 2020, **71**, 121–142.
38 J. P. Unsleber, S. A. Grimmel and M. Reiher, *J. Chem. Theory Comput.*, 2022, **18**, 5393–5409.
39 L.-P. Wang, A. Titov, R. McGibbon, F. Liu, V. S. Pande and T. J. Martnez, *Nat. Chem.*, 2014, **6**, 1044–1048.
40 E. Martnez-Núñez, G. L. Barnes, D. R. Glowacki, S. Kopec, D. Peláez, A. Rodrguez, R. Rodrguez-Fernández, R. J. Shannon, J. J. Stewart and P. G. Tahoces, *et al.*, *J. Comput. Chem.*, 2021, **42**, 2036–2048.
41 T. Baercor and P. M. Mayerfn, *J. Am. Soc. Mass Spectrom.*, 1997, **8**, 103–115.
42 P. Pechukas and J. C. Light, *J. Chem. Phys.*, 1965, **42**, 3281–3291.
43 C. E. Klots, *Z. Naturforsch., A*, 1972, **27**, 553.
44 J. I. Steinfeld, J. S. Francisco and W. L. Hase, *Chemical kinetics and dynamics*, Prentice Hall, Upper Saddle River, NJ, 1999.
45 D. M. Wardlaw and R. Marcus, *Chem. Phys. Lett.*, 1984, **110**, 230–234.
46 W. L. Hase, *Acc. Chem. Res.*, 1983, **16**, 258–264.
47 R. G. Gilbert and S. C. Smith, *Theory of unimolecular and recombination reactions*, 1990.
48 J. L. Bao and D. G. Truhlar, *Chem. Soc. Rev.*, 2017, **46**, 7548–7596.
49 D. G. Truhlar, B. C. Garrett and S. J. Klippenstein, *J. Phys. Chem.*, 1996, **100**, 12771–12800.

50 H. Eyring, *J. Chem. Phys.*, 1935, **3**, 107–115.
51 M. Barbatti, *J. Chem. Phys.*, 2022, **156**, 204304.
52 M. A. Haney and J. Franklin, *J. Chem. Phys.*, 1968, **48**, 4093–4097.
53 K. Kim, J. Beynon and R. Cooks, *J. Chem. Phys.*, 1974, **61**, 1305–1314.
54 V. Ásgeirsson, C. A. Bauer and S. Grimme, *Phys. Chem. Chem. Phys.*, 2016, **18**, 31017–31026.
55 J. H. Gross, *Mass spectrometry: a textbook*, Springer Science & Business Media, 2006.
56 M. Dantus, *Acc. Chem. Res.*, 2024, 033003.
57 S. Grimme, *Angew. Chem., Int. Ed.*, 2013, **52**, 6306–6312.
58 J. Meisner and J. Kästner, *Angew. Chem., Int. Ed.*, 2016, **55**, 5400–5413.
59 J. Pu, J. Gao and D. G. Truhlar, *Chem. Rev.*, 2006, **106**, 3140–3169.
60 S. Wang, T. Kind, P. L. Bremer, D. J. Tantillo and O. Fiehn, *J. Chem. Inf. Model.*, 2022, **62**, 4403–4410.
61 *Program package for the quantum mechanical calculation of EI mass spectra using automated reaction network exploration qcxms2*, **https://github.com/grimme-lab/QCxMS2**, Accessed: 2025-1-20.
62 P. Pracht, S. Grimme, C. Bannwarth, F. Bohle, S. Ehlert, G. Feldmann, J. Gorges, M. Müller, T. Neudecker and C. Plett, *et al.*, *J. Chem. Phys.*, 2024, **160**, 114110.
63 C. Bannwarth, S. Ehlert and S. Grimme, *J. Chem. Theory Comput.*, 2019, **15**, 1652–1671.
64 C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher and S. Grimme, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2020, **11**, e01493.
65 N. van Staalduinen and C. Bannwarth, *Digital Discovery*, 2024, **3**, 2298–2319.
66 G. Henkelman, B. P. Uberuaga and H. Jónsson, *J. Chem. Phys.*, 2000, **113**, 9901–9904.
67 F. Neese, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2022, **12**, e1606.
68 V. Ásgeirsson, B. O. Birgisson, R. Bjornsson, U. Becker, F. Neese, C. Riplinger and H. Jónsson, *J. Chem. Theory Comput.*, 2021, **17**, 4929–4945.
69 S. Smidstrup, A. Pedersen, K. Stokbro and H. Jónsson, *J. Chem. Phys.*, 2014, **140**, 214106.
70 X. Zhu, K. C. Thompson and T. J. Martnez, *J. Chem. Phys.*, 2019, **150**, 164103.
71 S. Spicher and S. Grimme, *J. Chem. Theory Comput.*, 2021, **17**, 1701–1714.
72 S. Grimme, *Chem. – Eur. J.*, 2012, **18**, 9955–9964.
73 *CREST – A program for the automated exploration of low-energy molecular chemical space*, **https://github.com/crest-lab/crest**, Accessed: 2025-1-16.
74 *A Molecular Identifier for Inorganic and Organic Molecules with Full Support of Stereoisomerism*, **https://git.rwth-aachen.de/bannwarthlab/molbar**, Accessed: 2024-10-29.
75 O. Vahtras, J. Almlöf and M. W. Feyereisen, *Chem. Phys. Lett.*, 1993, **213**, 514–518.
76 F. Weigend, *Phys. Chem. Chem. Phys.*, 2006, **8**, 1057–1065.
77 F. Neese, F. Wennmohs, A. Hansen and U. Becker, *Chem. Phys.*, 2009, **356**, 98–109.
78 *Interpolation of molecular geometries through geodesics in redundant internal coordinate hyperspace for complex transformations*, **https://github.com/virtualzx-nad/geodesic-interpolate**, Accessed: 2024-10-29.
79 *Semiempirical Extended Tight-Binding Program Package xtb*, **https://github.com/grimme-lab/xtb**, Accessed: 2024-10-29.
80 J. Koopman and S. Grimme, *J. Am. Soc. Mass Spectrom.*, 2021, **32**, 1735–1751.
81 *Quantum mechanic mass spectrometry calculation program*, **https://github.com/qcxms**, Accessed: 2024-10-29.
82 F. Huber, S. Verhoeven, C. Meijer, H. Spreeuw, E. M. V. Castilla, C. Geng, J. J. J. van der Hooft, S. Rogers, A. Belloum, F. Diblen and J. H. Spaaks, *J. Open Source Software*, 2020, **5**, 2411.
83 N. F. de Jonge, H. Hecht, M. Strobel, M. Wang, J. J. van der Hooft and F. Huber, *J. Cheminf.*, 2024, **16**, 88.
84 *Python program package Matchms*, **https://github.com/matchms**, Accessed: 2024-10-29.
85 Y. Li, T. Kind, J. Folz, A. Vaniya, S. S. Mehta and O. Fiehn, *Nat. Methods*, 2021, **18**, 1524–1531.
86 Y. Li and O. Fiehn, *Nat. Methods*, 2023, **20**, 1475–1478.
87 *Spectral entropy for mass spectrometry data*, **https://github.com/YuanyueLi/MSEntropy**, Accessed: 2024-10-29.
88 W. E. Wallace, in *Mass Spectra*, ed. P. J. Linstrom and W. G. Mallard, National Institute of Standards and Technology, NIST Chemistry WebBook, Gaithersburg, MD, NIST Standard Reference Database Number 69, 2019. **https://webbook.nist.gov**, Accessed: October 24, 2024.
89 M. Müller, A. Hansen and S. Grimme, *J. Chem. Phys.*, 2023, **158**, 014103.
90 S. Devata, H. J. Cleaves, J. Dimandja, C. A. Heist and M. Meringer, *J. Am. Soc. Mass Spectrom.*, 2023, **34**, 1584–1592.

APPENDIX D

Appendix: Evaluation of the QCxMS2 Method for the Calculation of Collision-Induced-Dissociation Spectra via Automated Reaction Network Exploration

Johannes Gorges[†], Marianne Engeser[‡], Stefan Grimme[†]

Submitted to the Journal of the American Society of Mass Spectrometry: 10 July 2025 – Published online: 4 September 2025

Reprinted in Appendix D from
J. Gorges, M. Engeser, and S. Grimme, *Evaluation of the QCxMS2 method for the calculation of collision-induced-dissociation spectra via automated reaction network exploration*, ChemRxiv Prepr. (2025), DOI: 10.26434/chemrxiv-2025-gcws2 – License: CC BY-NC-ND 4.0.

Own contributions

- Implementation of the CID mode in the QCxMS2 software
- Performing calculations with QCxMS2 using various program parameters and quantum chemical methods
- Interpretation of the results
- Writing the manuscript

[†]Mulliken Center for Theoretical Chemistry, Universität Bonn, Beringstr. 4, D-53115 Bonn, Germany
[‡]Kekulé Institute for Organic Chemistry and Biochemistry, Gerhard-Domagk-Str. 1, 53121 Bonn, Germany

Evaluation of the QCxMS2 method for the calculation of collision-induced-dissociation spectra via automated reaction network exploration

Johannes Gorges,[†] Marianne Engeser,[*,‡] and Stefan Grimme[*,†]

†*Mulliken Center for Theoretical Chemistry, Clausius-Institute for Physical and Theoretical Chemistry, University of Bonn, Beringstr. 4, 53115 Bonn, Germany*
‡*Kekulé Institute for Organic Chemistry and Biochemistry, Gerhard-Domagk-Str. 1, 53121 Bonn, Germany*

E-mail: marianne.engeser@uni-bonn.de; grimme@thch.uni-bonn.de

Abstract

Collision-induced dissociation mass spectrometry (CID-MS) is an important tool in analytical chemistry for the structural elucidation of unknown compounds. The theoretical prediction of CID spectra plays a critical role in supporting and accelerating this process. To this end, we adapt the recently developed QCxMS2 program originally designed for the calculation of electron ionization (EI) spectra to enable the computation of CID-MS. To account for the fragmentation conditions characteristic of CID within the automated reaction network discovery approach of QCxMS2 we adapted the internal energy distribution to match the experimental conditions. This distribution can be adjusted via a single parameter to approximate various activation settings, thereby eliminating the need for explicit simulations of the collisional process. We evaluate our approach on a test set of 13 organic molecules with diverse functional groups, compiled specifically for this study. All reference spectra were recorded consistently under the same measurement conditions, including both CID and higher-energy collisional dissociation (HCD) modes. Overall, QCxMS2 achieves good average entropy similarity scores (ESS) of 0.687 for the HCD spectra and 0.773 for the CID spectra. The direct comparison to experimental data demonstrates that the QCxMS2 approach, even without explicit modeling of collisions, is generally capable of computing both CID and HCD spectra with reasonable accuracy and robustness. This highlights its potential as a valuable tool for integration into structure elucidation workflows in analytical mass spectrometry.

Introduction

Given its high sensitivity and compatibility with a broad range of compounds, mass spectrometry (MS) has become a central technique in analytical chemistry, and is widely used across numerous application areas.[1] Among the many different ionization techniques, the "soft" electrospray ionization (ESI) method is often employed because it can be applied for a vast variety of analytes, such as organic molecules, metabolites, inorganic ions, synthetic and biological polymers, nucleic acids, peptides, and proteins.[2–4] To obtain structural information about stable ions, tandem MS is often applied. In a collision cell, kinetically accelerated precursor ions undergo activation through collisions with inert gas molecules. This induces fragmentation in a process known as collision-induced dissociation (CID).[5] In a linear ion trap, ion ac-

tivation may not be high enough to induce fragmentation sufficient for structure elucidation. More recently, advances such as higher-energy collisional dissociation (HCD), implemented in Orbitrap hybrid instruments, have enabled the acquisition of more information-rich fragmentation spectra.[6]

In practice, compounds are typically identified by matching measured fragmentation spectra against those from reference libraries.[7] This approach, however, is limited because only a small fraction of the relevant chemical space is covered by library spectra, leaving the majority as unidentified "dark matter".[8,9] Consequently, peak annotation and structural identification remain highly challenging, especially for novel or unexpected compounds lacking reference data.

To address the limitations of experimental spectral libraries, synthetic spectra predicted by theoretical methods can be used to enhance coverage of reference data and support compound identification. However, the accurate prediction of ESI-/CID-MS continues to pose a significant challenge due to the complex and often molecule-specific nature of fragmentation processes.

In recent years, data-driven machine learning (ML) approaches have demonstrated promising results in predicting CID-MS spectra. Notable examples include GrAFF-MS,[10] CFM-ID,[11] ICEBERG,[12] FraGNNet,[13] and FIROA.[14] In parallel, the reverse task—predicting structures directly from spectra, known as *de-novo* generation—has also gained attention. For instance, the recent DiffMS model[15] achieved remarkably good results, yet on the large-scale MassSpecGym dataset[16] correctly identified the molecular structure in only 2.3% of cases. This underscores the intrinsic difficulty of mapping a spectrum to a structure and highlights current limitations of ML models in MS.

While ML-based methods can be highly effective when trained on large, representative datasets, their applicability is fundamentally constrained by the availability and coverage of experimental spectra. In contrast, quantum chemical (QC) approaches follow a more general paradigm and are, in principle, capable of describing arbitrary molecular structures and fragmentation pathways without relying on prior empirical data. Compared to the broad variety of existing ML-based approaches, considerably fewer QC methods are available for calculating CID-MS spectra.[17] Most QC-based programs rely on molecular dynamics (MD) simulations, such as the VENUS program package,[18,19] the recent CIDMD method,[20] and the QCxMS program,[21–23] developed over many years in our laboratory. CIDMD, which employs computationally demanding ab initio MD based on density functional theory (DFT), has demonstrated highly accurate results for a small benchmark set of twelve metabolites.[20] In contrast, QCxMS can also be used with more efficient semi-empirical quantum mechanical (SQM) methods, enabling the simulation of larger molecular systems with reasonable accuracy, as shown in several studies.[22–25]

Due to the high computational cost of MD-based approaches—which restricts the feasible system size and level of theory, as well as the limited simulation times (typically on the picosecond scale), which are several orders of magnitude shorter than many fragmentation processes (up to microseconds)—we recently developed the QCxMS2 program. QCxMS2 employs an automated reaction network discovery approach, which enables the use of more accurate electronic structure methods by combining efficient SQM methods for structure optimization and energetic refinement with higher-level DFT methods. This approach has demonstrated superior accuracy compared to QCxMS in the prediction of electron ionization mass spectra (EI-MS).[26]

In this work, we apply QCxMS2 to the calculation of CID-MS spectra. Previous studies with QCxMS have shown that its *temprun* mode, i.e., simulations without explicit modeling of individual collisions, can often yield sufficiently accurate CID spectra for practical applications.[22] When the internal energy distribution is appropriately chosen to reflect the experimental conditions, the important fragmentation pathways are typically captured correctly. Here, we evaluate the assumption that fragmentation in CID occurs after collision events

from thermally excited but quasi-equilibrated ions, as described by quasi-equilibrium theory (QET).[27] QET allows us to bypass the explicit modeling of the collisional process, which is inherently incompatible with the MD-free QCxMS2 approach. Instead, QCxMS2 employs a predefined internal energy distribution, assumed to arise from ionization and subsequent collisional activation, to compute unimolecular reaction rates of fragmentation events and thereby estimate the relative intensities of the resulting fragment ions.

For the evaluation of QCxMS2, we compiled a benchmark set of 13 organic compounds containing a diverse range of functional groups and measured both CID and HCD spectra for each substance at varied collision energies. To ensure consistency and minimize experimental uncertainties, particularly in signal intensities, all reference spectra were recorded under largely identical conditions on the same instrument. The experimental spectra are in very good accordance with respective spectra reported previously.[28–34] In this study, we focus on singly protonated precursor ions and evaluate whether the QCxMS2 approach is generally applicable for the theoretical prediction of CID spectra. Extensions to multiply charged or negatively charged species are possible using a similar approach as in QCxMS[23,35] and will be addressed in future work.

First, we provide a brief overview of the theoretical background of QCxMS2. Next, we summarize the experimental setup used for recording the reference measurements and describe the methods and adaptations used to compute CID-MS spectra. The accuracy of QCxMS2 is then assessed by comparing the computed spectra with the experimental data for the test set and by benchmarking its performance against its predecessor, QCxMS. Due to the high computational cost of CIDMD[20,36] and its reliance on the commercial TeraChem software package,[37] we excluded it from our study. Finally, we discuss computational timings, evaluate the overall performance and limitations of QCxMS2 for CID-MS applications, and propose potential use cases.

Theory

The maximum increase in internal energy during a collision is given by the center-of-mass energy E_{com} defined by

$$E_{\mathrm{com}} = \frac{m_{\mathrm{g}}}{m_{\mathrm{g}} + m_{\mathrm{p}}} E_{\mathrm{kin}}, \tag{1}$$

where E_{kin} is the kinetic energy of the precursor ion (in the so-called lab frame), and m_{g} and m_{p} denote the masses of the collision gas and the precursor ion, respectively.[38]

The proportion of E_{com} converted into internal energy ΔE_{int} is determined by the empirical collision inelasticity η:

$$\Delta E_{\mathrm{int}} = \eta E_{\mathrm{com}}. \tag{2}$$

Based on experimental studies, the value of η was determined to be ≈ 0.5[39] which yields good agreement with experimental data in QCxMS.[22]

Similar to QCxMS, QCxMS2 employs a normally distributed internal energy model to approximate the activation energies, arising from both the ESI process and subsequent collisions. The distribution is generated using the Box–Muller method[40] and follows the well-known normal distribution:

$$P(E) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(E - E_{\mathrm{avg}})^2}{2\sigma^2}\right), \tag{3}$$

where the width of the distribution is set to $\sigma = 0.2\,\mathrm{eV}$ and the average internal energy E_{avg} is adjusted via a single input parameter, *esiatom*, usually set to values between 0.2 eV and 0.6 eV per atom depending on the experimental conditions.

With this approach, non-statistical processes, i.e., direct bond cleavages occurring immediately upon collision, before full equilibration of internal energy, are not captured. According to experimental and theoretical studies,[41,41,42] such processes may play a role in certain fragmentation pathways, but are expected to be of less relevance for the main fragmentation pathways. The survival yield (SY) of the precursor

ion is defined as

$$SY = \frac{I_\mathrm{P}}{I_\mathrm{P} + \sum I_{\mathrm{F},i}}, \tag{4}$$

where I_P denotes the intensity of the precursor ion and $\sum I_{\mathrm{F},i}$ the sum of all fragment signal intensities. SY is an important quantity in experimental studies.[43,44] In QCxMS2, the ratio of the initial intensity I_{P0} and the remaining intensity I_P is computed via the rate law for unimolecular (first-order) reactions

$$\frac{I_\mathrm{P}}{I_{\mathrm{P0}}} = e^{-k(E)t}, \tag{5}$$

where t is the time of flight in the mass spectrometer. While for EI-MS, a value of $\approx 50\,\mu$s[45] is taken, the cycle time in ion trap CID instruments amounts from milliseconds up to a second. However, as the absolute rate constants computed in QCxMS2 are due to the uncertainty of the internal energy distribution anyway adjusted to the experiment, we also employ $50,\mu$s for the CID-MS calculations in this work. The choice of the correct value of t is effectively captured by adjusting the average internal energy, and using a longer time of flight has no significant impact on the resulting spectra (see SI†, S4).

The rate constants $k(E)$ of the reactions are derived from the reaction barriers via conventional transition state theory.[46] Molecular charges are assigned according to the ionization potentials (IP) of the resulting fragments, which are determined by self-consistent field (ΔSCF) calculations at the chosen QC level (typically DFT). The statistical weight of each ion is then given by Boltzmann weighting of the energy differences $\Delta E_{\mathrm{SCF},i}$ between the neutral and charged states of fragment i:

$$P_i = \frac{\exp\left(-\frac{\Delta E_{\mathrm{SCF},i}}{k_\mathrm{B} T_\mathrm{Av}}\right)}{\sum_{j=1}^{M} \exp\left(-\frac{\Delta E_{\mathrm{SCF},j}}{k_\mathrm{B} T_\mathrm{Av}}\right)}, \tag{6}$$

where M is the total number of fragments. For further details on the exact procedure implemented in QCxMS2, see the original publication.[26]

Methods

Experimental details

All mass spectra were recorded with an Orbitrap XL ETD instrument equipped with a HESI ion source. Samples were dissolved in acetonitrile or, when necessary, in mixtures of acetonitrile and water. Standard ESI source conditions were as follows: Spray Voltage 5 kV, Capillary Temperature 275 °C, Capillary Voltage 35 V, Tube Lens 130 V. Spectra were recorded using the Orbitrap analyzer with the resolution set to $R = 60\,000$. Resolution and mass accuracy are sufficiently high for unequivocal signal assignment to elemental compositions throughout this study (see SI†, S15, for a complete list of signals and assignments). Due to instrumental limitations, signals below m/z 50 were not detected.

Tandem mass spectra were acquired after mass selection of the monoisotopic peak of the precursor ion within the linear ion trap of the instrument. No isotope patterns are thus visible in the tandem mass spectra. Further, all fragmentations occurring in the ion source and transfer region, known as in-source (IS)-CID,[47,48] prior to mass selection do not contribute to the recorded spectra. Due to the known severe in-source fragmentation (IS-CID) of serotonin,[28] we employed softer ESI conditions (Capillary Voltage 5 V, Tube Lens 30 V) for this compound. Histamine is also known to be prone to IS-CID,[29] but a reasonable mass spectrum could still be obtained under standard ESI conditions.

The Orbitrap XL is a hybrid instrument that provides two options to induce ion fragmentation by collisions with an inert gas. On the one hand, CID can be performed in the helium-filled linear ion trap. On the other hand, a nitrogen-filled octopol at the end of the instrument enables higher-energy collisions (HCD). For both modes, series of spectra were recorded at varied collision energies. The HCD spectra presented herein were acquired at a collision energy of $E_\mathrm{lab} = 70,\mathrm{eV}$ for most compounds. For hexamethoxyphosphazine (HMP), leucine, aspartic acid, arginine, and glutamic acid, lower energies

were selected to retain some precursor ions and ensure a meaningful comparison with the computed spectra. Similarly, the CID spectra were measured at $E_{\mathrm{lab}} = 25, \mathrm{eV}$, except for the same compounds, for which 20 or 15 eV were chosen.

A summary of the test set comprising 13 compounds and their respective experimental conditions is provided in Table 1 for HCD, and in Table S2 of the SI†for CID.

Computational details

The results discussed in Section were computed with QCxMS2 version 1.2.0[49] using a normal-distributed energy distribution with different average internal energies specified by the parameter *esiatom* given in eV per atom. The intensity threshold for subsequent fragmentation was consistently set to 5%, as this provides the best compromise between accuracy and computational efficiency (see SI†, Section S12). Otherwise, default settings were used. For the calculation of ionization potentials (IPs), the default composite level of GFN2-xTB[50,51] was used, with refinement of close IPs (below 2 eV/mol) at the ωB97X-3c[52] level of theory.

Input structures of the protomers in the test set were generated with the protonation tool of CREST[53,54] at the GFN2-xTB level within a 60 kcal mol^{-1} energy window. Duplicate protomers were identified and removed using MolBar[55] version 1.1.3,[56] and artificially rearranged structures were manually detected and excluded. Relative protomer energies were computed with ωB97X[57]-D4[52,58]/def2-QZVP[59] on r^{2}SCAN-3c[60]-optimized geometries, including thermal contributions at the same level of theory, using the modified rigid-rotor harmonic oscillator (mRRHO) approximation[61] with a default cutoff of 50 cm^{-1} at a temperature of 548 K (matching the experimental capillary temperature). All protomers within 40 kcal mol^{-1} of the lowest-energy structure were considered in this study. Additionally, protomer energies were calculated with the SMD[62] solvation model, consistently using water as solvent. As input, the minimum-energy conformers of each protomer at the GFN2-xTB level, identified by CREST development version 3.0.2,,[63] were used.

Fragments were generated with CREST MSREACT, and duplicates were identified using MolBar. All DFT calculations, NEB path searches, and transition state optimizations were carried out with ORCA version 6.0.0.[64] Reaction barriers in QCxMS2 were refined at the ωB97X-3c level.[52] The resolution-of-identity approximation[65] with matching auxiliary basis sets[66] was applied for the Coulomb integrals, while exchange integrals were computed using the RIJCOSX approximation.[67] Single-point calculations with g-xTB (the successor of GFN2-xTB) were performed using a development version of the gxtb program.[68,69]

Initial reaction paths for restarted NEB calculations were generated with geodesic-interpolation version 1.0.0.[70] GFN2-xTB calculations were performed with a development version of xTB 6.7.2[71] using default convergence settings and the tblite implementation, as it includes two separate electronic spin channels and an unrestricted SCF procedure.[72] QCxMS spectra for comparison were computed with QCxMS V5.2.1[22,23,73] using default settings at the GFN2-xTB level in its tblite implementation. Spectra were generated with PlotMS version 6.2.1.[74] Entropy similarity scores (ESS)[75,76] were computed using the msentropy Python package.[77]

Results and discussion

In this section, we compare the QCxMS2 spectra with experimental spectra for 13 compounds, encompassing a total of 41 protomers. The test set covers a chemically diverse range of biologically relevant molecules, including the amino acids aspartic acid, glutamic acid, arginine, leucine, lysine, and hydroxyproline. It also features the biogenic amines and neurotransmitters histamine, serotonin, and the choline ester methacholine. N-heterocycles are represented by caffeine and nicotine. Additionally, the set includes the widely used analgesic paracetamol, exemplifying aromatic drug-like phenol derivatives. Finally, the phosphorus-containing heterocycle hexamethoxyphosp-

hazine (HMP) is included, which is also the largest compound in the set, comprising 37 atoms. Lewis structures of all compounds and their protomers can be found in the SI†, Section S1, and xyz coordinate structures of all compounds and their protomers can be found in the electronic SI.

To evaluate the computed spectra, we employ the entropy similarity score (ESS),[75,76] which ranges from 0 (no agreement) to 1 (perfect agreement). An ESS of approximately 0.75 has been suggested as a meaningful threshold for reliable structure identification[75] and indicates very good agreement. For ESS calculations and spectral comparisons, theoretical spectra were generated using exact masses without isotope patterns, consistent with the mass selection of the monoisotopic precursor ion peak in the experiment. Additionally, computed masses below m/z 50 are omitted, as these cannot be detected on the instrument employed. In all figures, the experimental spectra intensities are plotted inverted (downward) to facilitate comparison with the calculated spectra. Exclusion of the molecular ion peak for ESS evaluation—previously shown to improve matching with library spectra[75,76]—does not affect our overall results (see SI†, S5). Unless otherwise stated, QCxMS2 calculations were performed at the "composite" level of theory, employing single-point energy calculations at the ωB97X-3c level on geometries optimized with GFN2-xTB.

Dependence on internal energy

First, we examine how the internal energy distribution settings in QCxMS2 influence the fragmentation patterns, in order to reproduce the experimental HCD and CID conditions. To this end, we computed spectra at five different internal energy settings, varying the *esi-atom* parameter from 0.2 to 0.6 eV per atom in steps of 0.05 eV, and determined the resulting spectral intensities.

The relationship between the internal energy settings in QCxMS2 and the experimental collision energies is illustrated using caffeine as an example. For caffeine, spectra were measured at different collision energies for CID in 3 eV increments up to 38 eV and for HCD in 5 eV increments up to 80 eV. Fragmentation was observed starting from 18 eV in CID and from 30 V in HCD.

Figure 1 shows the survival yield (SY) of the precursor ion in both the experiments and the QCxMS2 calculations, plotted against the experimental lab-frame collision energies and the computed internal energies, respectively. Notably, the characteristic logistic function behavior[39] observed in the CID and HCD experiments is well reproduced by QCxMS2. However, the absolute energies are challenging to compare between experiment and calculation due to uncertainties in the internal energy imparted by the ionization process, the number of collisions experienced by the precursor ion, and the computed absolute rate constants on which the SY depends exponentially (see Eq. 5). This demonstrates that the internal energy has to be adjusted to the experimental conditions to correctly reproduce the SY.

Figure 2 depicts the computed spectra of the lowest-energy protomer of caffeine using the energy settings that yield the highest ESS at different experimental collision energies. The spectra show that QCxMS2 is capable of computing both HCD and CID spectra at different experimental energies when appropriate internal energy settings are employed. For HCD, the best energy settings for experimental collision energies of 35, 70, and 80 eV are 0.3, 0.35, and 0.4 eV per atom, respectively, yielding good ESS values of 1.000, 0.698, and 0.624. Notably, at higher collision energies, the agreement decreases as the spectra become more complex and limitations of QCxMS2 become more pronounced. Due to incorrectly predicted peaks at m/z 95 and 94, likely due to errors at the employed QC level of theory, the best-matching computed spectra exhibit a higher SY than the experimental spectra. Internal energy settings that produce an SY comparable to the experimental value also lead to increased intensities of these artificial peaks, thereby reducing the ESS. For the same reason, ESS values for CID spectra also decrease at higher energies. However, because fewer experimental peaks are observed in CID, the evaluation of QCxMS2 for

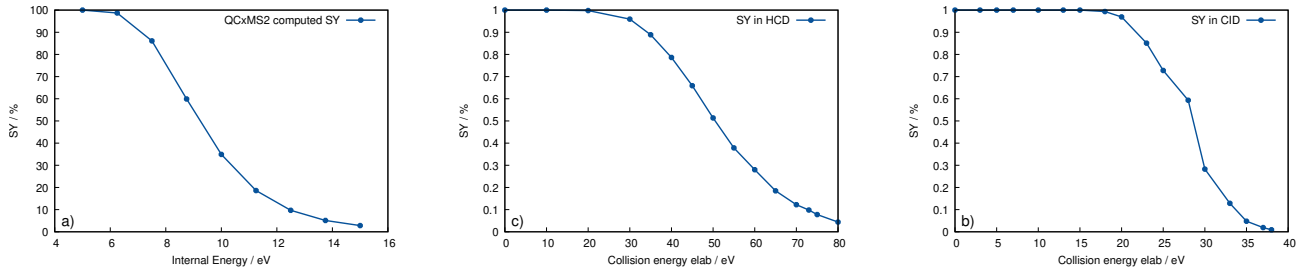


Figure 1: Survival Yield (SY) of the lowest-energy protomer of caffeine a) computed with QCxMS2, experimental b) HCD spectra and c) CID spectra.

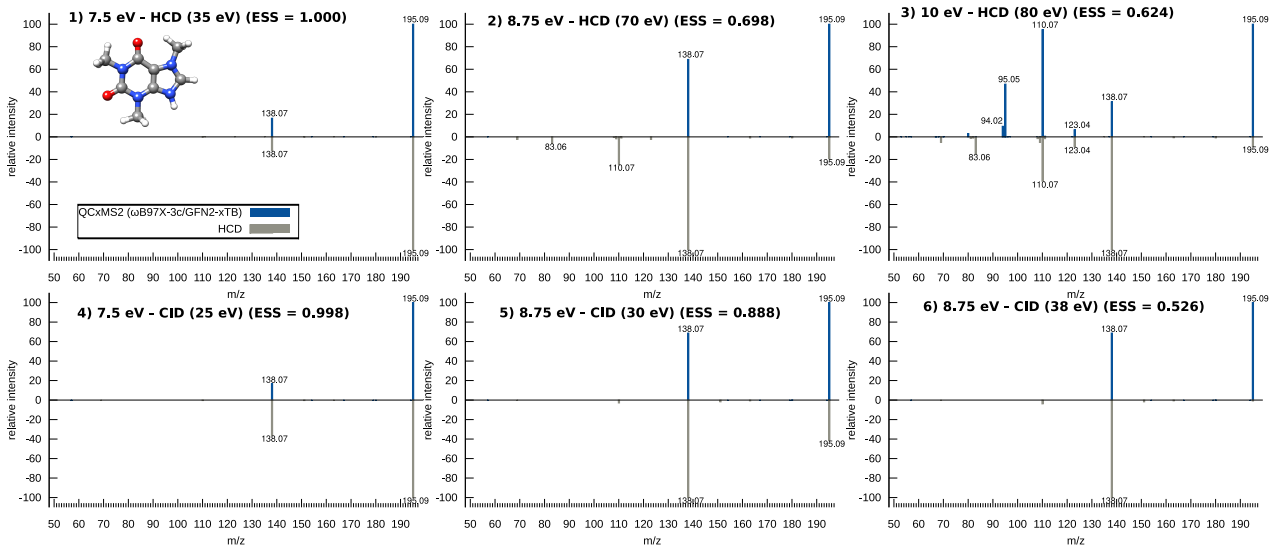


Figure 2: Calculated spectra with QCxMS2 at the "composite" (ωB97X-3c//GFN2-xTB) level of theory for the lowest-energy tautomer of caffeine at the best-matching internal energy, given in eV for each spectrum, compared to the inverted experimental spectra measured at different HCD (**1**-**3**) and CID (**4**-**6**) collision energies. The respective entropy similarity scores (ESS) are given in brackets.

these spectra is less conclusive.

To identify suitable general energy settings, we computed spectra at various average internal energies for all compounds in the test set and compared them with the experimental spectra. Figure 3 shows the average ESS for the test set, comparing the lowest-energy protomer and the best-matching protomer against the HCD and CID spectra using the same *esiatom* parameter across all compounds. The best-matching HCD spectra were obtained with the average internal energy set to $esiatom = 0.4\,\text{eV}$ per atom, when using the lowest-energy tautomer and 0.45 eV when selecting the best-matching tautomer.

We also determined the best-fitting *esiatom* value individually for each compound. The resulting ESS values relative to the experimental HCD spectra, using the best *esiatom* settings and best-matching protomer, are summarized in Table 1. Detailed ESS values for all protomers and for the CID spectra are provided in the SI†, Section S3. Using the best energy settings for each compound individually, an average ESS of 0.687 is achieved, which is significantly closer to the target value of 0.75 than the average ESS of 0.613 obtained when applying a uniform $esiatom = 0.45\,\text{eV}$ per atom across all compounds. Notably, the optimal energy settings are highly system-specific, and no clear general trend is noted. Contrary to expectations, for the compounds measured at lower collision energies, namely HMP, leucine, aspar-

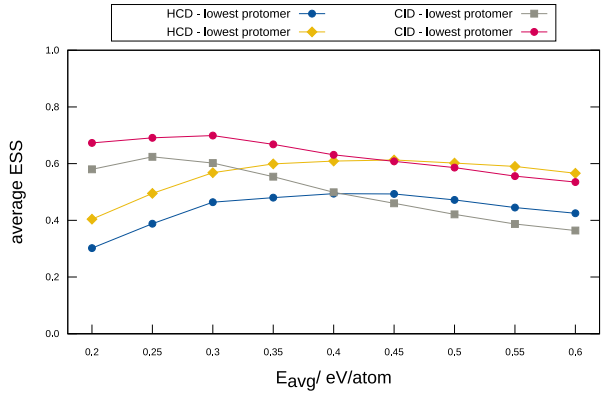


Figure 3: Average entropy similarity scores (ESS) of QCxMS2 spectra for the whole test set computed at different average internal energies E_{avg} per atom with HCD and CID spectra for the lowest-energy protomer and best-matching protomer.

Table 1: Experimental energies of HCD spectra E_{lab}, number of considered protomers n_{prot} and entropy similarity scores (ESS) of QCxMS2 spectra computed at best matching *esiatom* values in eV/atom for best matching protomer i_{best} of each compound compared to the experimental HCD spectra measured at E_{lab} in eV. For the respective ESS values with the CID spectra and their experimental energies, see SI†, S2.

compound	E_{lab}	*esiatom*	n_{prot}	i_{best}	ESS
caffeine	70	0.35	4	4	0.844
HMP	65	0.55	2	1	0.495
leucine	50	0.25	2	1	0.934
aspartic acid	50	0.6	3	3	0.732
arginin	50	0.5	4	4	0.648
lysine	70	0.35	2	1	0.549
nicotine	70	0.6	2	1	0.407
methacholine	70	0.35	1	1	0.679
histamine	70	0.35	3	1	0.756
serotonin	70	0.55	7	3	0.580
paracetamol	70	0.45	5	3	0.783
hydroxyproline	70	0.4	3	2	0.680
glutamic acid	50	0.25	3	3	0.843
average		0.43	-	-	0.687

tic acid, arginine, and glutamic acid, reduced internal energy settings did not result in improved simulation quality compared to the settings determined for the other compounds.

A similar trend is observed for the CID spectra: employing the best individual energy settings yields an average ESS of 0.773, whereas using the average optimal internal energy of *esiatom* = 0.3 eV per atom gives a lower value of 0.699. We attribute this pronounced system-specific energy dependence to the absence of explicit modeling of collision and collisional cooling processes, combined with uncertainties in the internal energy imparted during ionization, all of which influence the actual internal energy distribution of the ions.

In the following, we focus on discussing the HCD spectra, as they typically exhibit more fragment peaks, corresponding to greater spectral entropy,[75] and are therefore more meaningful for comparison with the computed spectra than the CID spectra.

Impact of protonation site

Most compounds in this study exhibit multiple protonation sites, resulting in different protomers that can lead to substantially distinct fragmentation patterns. Therefore, spectra of all protomers identified within a 40 kcal mol^{-1} free energy window at the DFT level were computed for each compound. Of the 13 compounds investigated, methacholine is inherently cationic and does not require additional protonation. HMP, leucine, lysine, and nicotine each exhibit two protomers. Aspartic acid, histamine, hydroxyproline, and glutamic acid each have three protomers, caffeine and arginine have four, paracetamol five, and serotonin the most with seven protomers. Relative free energies in the gas phase and in solution for all protomers are tabulated in the SI, Section S10. The energetic ordering of the protomers is very similar between the gas-phase and solvated calculations, consistent with previous studies on relative protomer stabilities in CID.[23,36] Only for nicotine and histamine does the lowest-energy protomer differ between phases.

Figure 4 shows the ESS values computed with QCxMS2 for all compounds in the test set at their respective best energy settings, comparing the best-matching protomer with the lowest-energy protomer. Notably, for seven compounds, a higher-energy protomer yields the

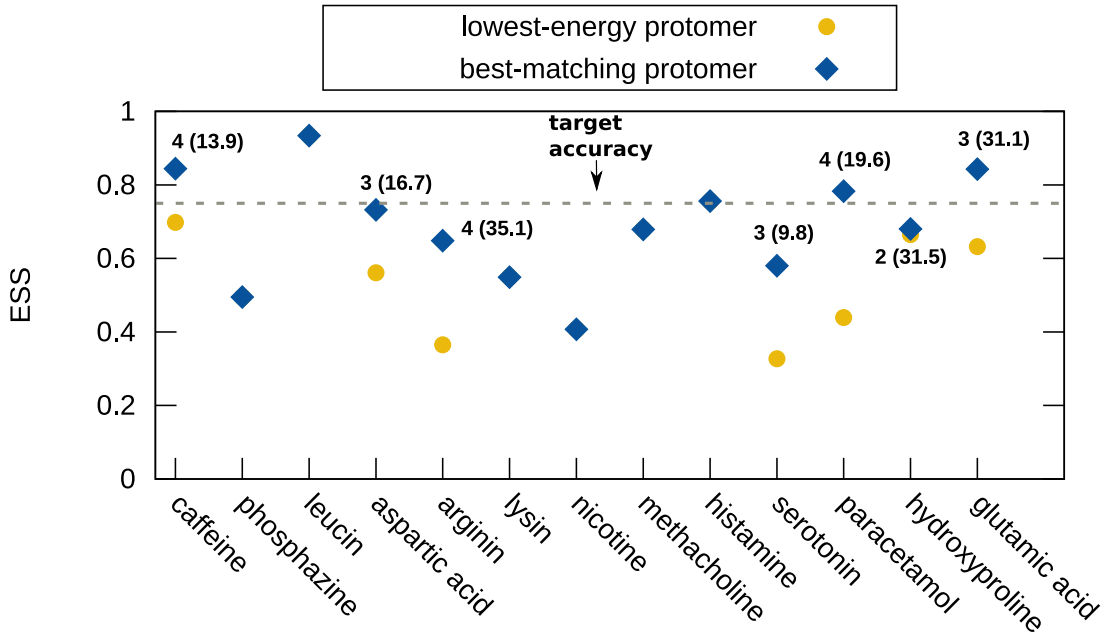


Figure 4: Entropy similarity scores (ESS) between experimental HCD spectra and theoretical spectra computed with QCxMS2 at the "composite" ωB97X-3c//GFN2-xTB levels at best energy settings for the lowest-energy protomer in the gas phase and the best-matching protomer for all compounds of the test set. The order number of the best matching protomer and its relative free energy (computed in gas phase) in kcal mol^{-1} is also given in parentheses. The target accuracy of an ESS of 0.75 with the experiment is indicated by the grey dashed line.

best agreement with experiment. In the case of arginine, the most relevant protomer is as much as 35.1 kcal mol^{-1} higher in energy. Overall, the average ESS is 0.577 when using the lowest-energy protomer in the gas phase (0.555 in solution) and improves to 0.687 when selecting the best-matching protomer for each compound. For CID (see SI†, Section S7), the same trend is observed, with corresponding values of 0.747, 0.692, and 0.790, respectively.

This finding aligns with several previous studies employing ab initio MD simulations with CIDMD,[36] as well as SQM MD simulations with QCxMS for larger molecules.[22,23] These works have demonstrated that multiple tautomers can contribute to the observed mass spectrum and that it is often not the thermodynamically most stable form, but rather the kinetically most labile one,i.e., the tautomer associated with the lowest-energy fragmentation pathway, that predominates. It is also possible that very labile tautomers fragment before mass selection due to in-source CID (IS-CID), as seen, for example, with histamine[29] and serotonin.[28] In such cases, the recorded spectra primarily originate from the fragmentation of the remaining, more stable protomers. Moreover, it is known that high-energy protomers can be kinetically stable and appear in the mass spectrum.[78,79]

Given that the computed spectra suggest multiple protomers are formed during the ionization process, we tried to address the role of their distribution. To this end, we Boltzmann-weighted the spectra of the protomers of each compound at various effective temperatures, ranging from 548 K (the capillary temperature in the experiment) up to infinity, where all protomers are equally weighted. Weighting at 548 K yields an average ESS of 0.506 using

gas-phase free energies and 0.491 when employing solution-phase free energies. Interestingly, the best average ESS across the tested temperatures is 0.560 at infinite temperature—that is, when all protomer spectra are equally weighted—which remains lower than simply using the lowest-energy protomer in each case. A similar trend is observed for the CID spectra (see SI†, Section S8 for details).

These observations highlight the challenge of determining the most appropriate weighting of different protomer spectra and underscore the non-equilibrium nature of ESI.[80,81] The situation is further complicated by the possibility that, according to the mobile proton theory, protomers can interconvert on the timescale of the mass spectrometric experiment, thereby allowing multiple protonation sites to contribute to the observed fragmentations.[82]

An illustrative example of this phenomenon is paracetamol, for which the QCxMS2-computed spectra of the four lowest-energy protomers are shown in Figure 5. The spectra were computed at 0.45 eV per atom, the internal energy setting at which the highest ESS is achieved with protomer **4**. Protomer **5** is not depicted, as it has an ESS of only 0.345 and a significantly higher relative free energy (35.5 kcal mol^{-1}) compared to the other four protomers.

Notably, the main signal observed in the experimental spectrum at m/z 110 stems from ketene loss, which is identified as the most favorable fragmentation pathway only for protomers **3** and **4**. By contrast, for protomers **1** and **2**, water loss is the dominant fragmentation channel, leading to a fragment at m/z 134, although this peak is much less pronounced in the experimental spectrum. Consequently, protomers **3** and **4** exhibit the best agreement with experiment, with ESS values of 0.648 and 0.783, respectively, compared to only 0.439 and 0.472 for **1** and **2**.

Interestingly, protomer **4** undergoes rearrangement to protomer **2** and subsequently to protomer **1**, thereby producing the fragments at m/z 134 and m/z 93. One possible explanation for the dominance of the higher-energy protomers in the spectrum is the aforementioned IS-CID of protomers **1** to **3**, which may effectively filter out these species before they reach the collision cell. As a result, only protomer **4** is sufficiently stable to survive mass selection and contribute to the observed spectrum. This interpretation is consistent with a study by Bahrami et al., who attributed fragmentation leading to m/z 110 to higher-energy protonated isomers of paracetamol.[83]

The QCxMS2-computed SY at 0.45 eV per atom for protomers **1** to **4** is 8.4%, 0.2%, 0.1%, and 12.7%, respectively, indicating that protomer **4** is the kinetically most stable and thus contributes most significantly to the spectrum in this case. However, it remains unclear how the SY could be incorporated into an automated, general procedure to predict a combined spectrum arising from the various contributions of different protomers. Even with the kinetics-based QCxMS2 approach, it is still necessary to compute spectra for all relevant protomers, typically within an energy window of up to approximately 40kcal mol^{-1} .

Effect of level of theory

Next, we examine the effect on the spectra of using ωB97X-3c-optimized geometries instead of GFN2-xTB geometries, with the examples of histamine and aspartic acid, as depicted in Figure 6. For both compounds, the best-matching protomer and energy settings for the ωB97X-3c calculations are shown.

The spectrum of the lowest-energy protomer of histamine computed using GFN2-xTB geometries shows generally good agreement with the experimental data, including the dominant loss of NH_3, yielding an ESS of 0.709. However, some deviations in signal intensities are observed, and the experimentally measured peak at m/z 83.06 is missing, while instead a peak at m/z 82.05 is predicted. When using geometries optimized at the ωB97X-3c level, the agreement improves substantially, resulting in a higher ESS of 0.797. The overall fragmentation pattern is well reproduced, although the peak at m/z 83.06 is still underestimated by one mass unit. Such deviations, where peaks are shifted by ± 1 m/z, have also been observed in computed EI-MS spectra with QCxMS2[26] and

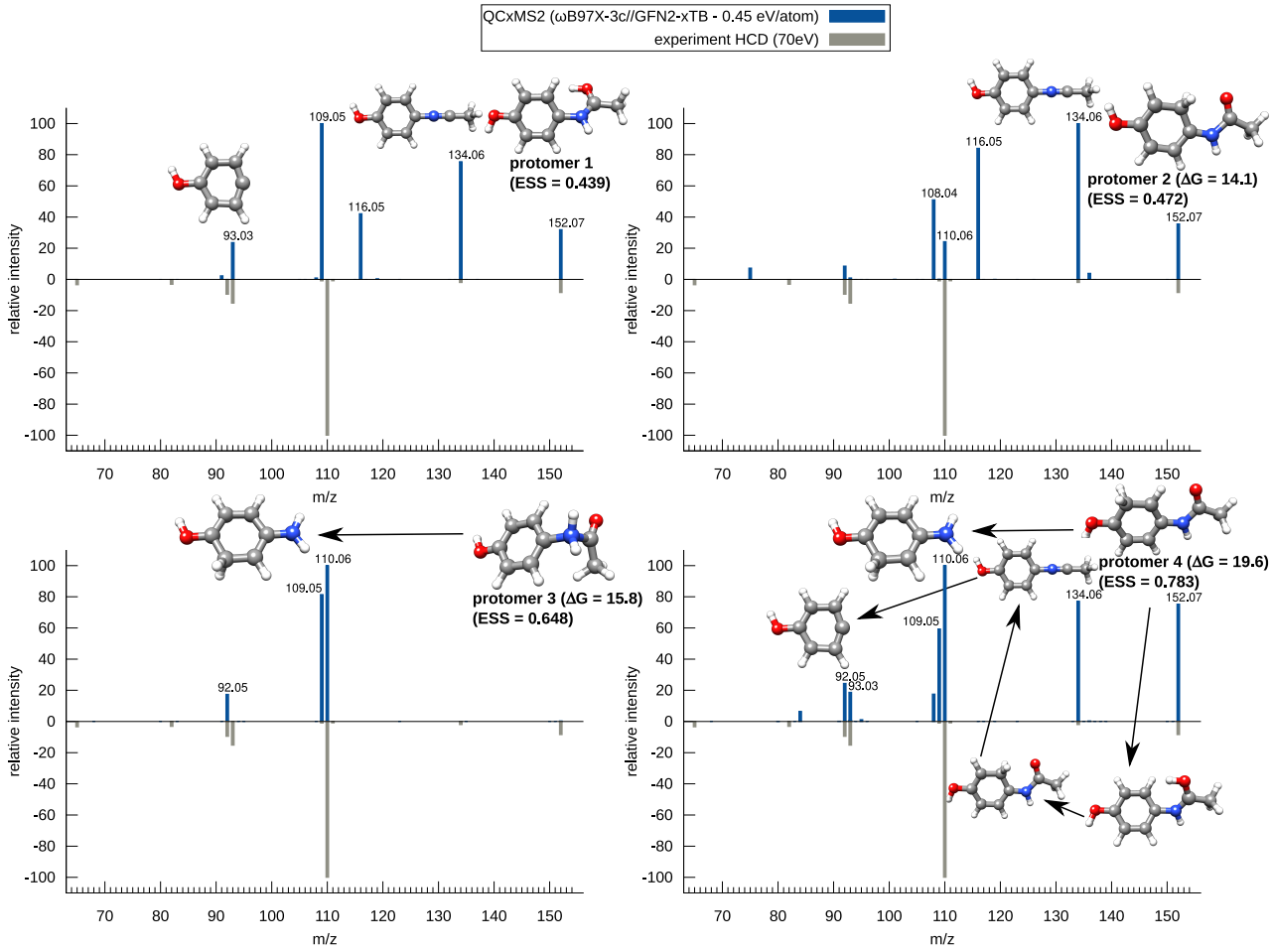


Figure 5: QCxMS2 computed spectra of the four lowest-energy protomers of paracetamol with entropy similarity scores (ESS) compared to the experimental HCD spectrum and relative free energies in the gas phase of the protomers ΔG in kcal mol^{-1} .

are attributed to the challenges of accurately modeling hydrogen transfer and rearrangement reactions.

For the lowest-energy protomer of aspartic acid, the composite level yields very poor agreement with experiment, achieving an ESS of only 0.361. It is noteworthy that, probably due to error compensation, the best-matching internal energy and even the best-matching protomer can differ depending on the level of theory employed. At this level of theory, a good ESS of 0.732 is obtained when using protomer **3** at 0.6 eV per atom. In contrast, using ωB97X-3c for protomer **1** yields almost perfect agreement with experiment, with an ESS of 0.961.

Due to the size of the systems studied, we employed only the composite level of theory (ωB97X-3c//GFN2-xTB) and GFN2-xTB//GFN2-xTB and performed full DFT calculations only for a small subset of molecules, namely aspartic acid, hydroxyproline, and histamine. Additionally, we evaluated the new g-xTB semiempirical method on GFN2-xTB-optimized geometries. Detailed ESS values for these calculations are provided in the SI†, Section 11. The results obtained with g-xTB//GFN2-xTB are very similar to those from ωB97X-3c//GFN2-xTB, yielding an average ESS of 0.694, which demonstrates the robustness and high accuracy of g-xTB for these systems.

As expected, computing the barriers with GFN2-xTB alone gives a lower average ESS of 0.618, significantly worse than that achieved with the composite approach. Conversely, using ωB97X-3c for both geometries and energies im-

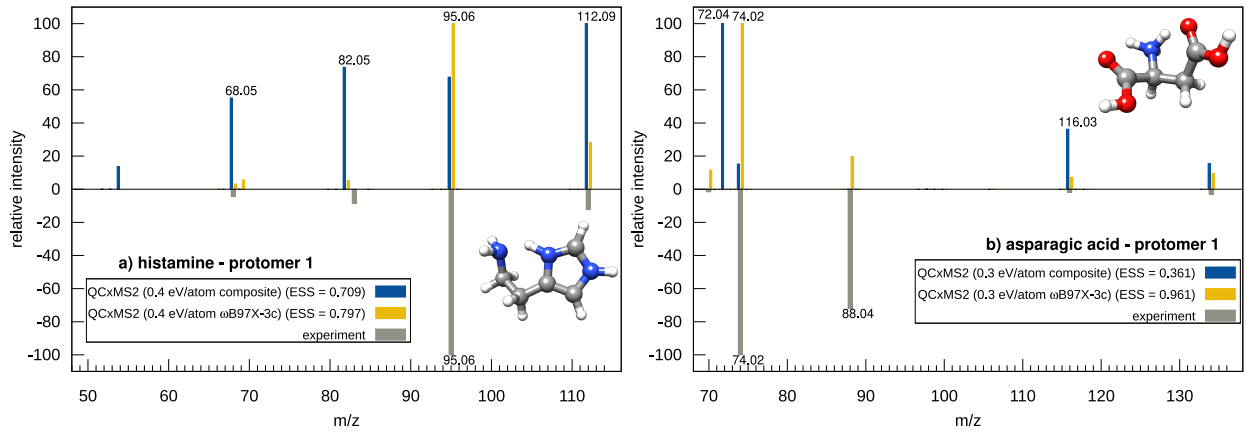


Figure 6: Calculated spectra with QCxMS2 at the "composite" (ωB97X-3c//GFN2-xTB), and ωB97X-3c levels of theory compared to the inverted experimental HCD spectrum of a) histamine and b) aspartic acid. Theoretical spectra were shifted by $\pm$ 0.25 m/z units for better visibility. The respective entropy similarity scores (ESS) to the experimental spectra are also given.

proves the average ESS for the three molecules to 0.827, compared to 0.723 with the default composite level. For hydroxyproline specifically, ωB97X-3c yields a better result of 0.724 at 0.35 eV per atom with protomer **1**, whereas the composite level gives a maximum ESS of 0.680 at 0.4 eV per atom with protomer **2**.

The expected trend that a higher level of theory gives more accurate spectra has also been observed for EI-MS with QCxMS2[26] and strongly supports the validity of the QCxMS2 approach for computing CID-MS spectra.

Comparison to QCxMS

Next, we compare the accuracy of QCxMS2 with its predecessor, QCxMS. QCxMS calculations were performed at the GFN2-xTB level in the thermal activation *temprun* mode, i.e., also without explicit collisions, for the sake of direct comparison. The internal energy was scaled to *esi 6* eV. For HCD, an average ESS of only 0.377 is achieved, which is clearly surpassed by QCxMS2, yielding an average ESS of 0.613 under the same internal energy settings of 0.45 eV per atom for all compounds.

For the HCD spectra, we also computed QCxMS spectra in the *collauto 6* mode at *elab 70* eV to investigate the effect of explicitly modeling collisions. However, no improvement is observed, with an average ESS of 0.382. For CID, using *esi 2* eV, an average ESS of 0.626 is obtained, which also remains significantly lower than the accuracy achieved by QCxMS2. Detailed ESS values obtained with QCxMS for each compound are provided in the SI†, Section S9.

It should be noted that the ESS values of QCxMS could likely be improved by individually tuning the energy and collision settings for each compound. However, this is beyond the scope of the present study. Figure 7 **a)** and **b)** show the computed spectra for glutamic acid and HMP using both QCxMS2 and QCxMS. For glutamic acid, QCxMS shows poor agreement with the experiment, exhibiting almost no molecular peak and minimal intensity for the main experimental fragment at m/z 84, resulting in an ESS of only 0.323. In contrast, QCxMS2 predicts all major peaks correctly and achieves a very good ESS of 0.843. For HMP, QCxMS performs even worse, failing to reproduce essentially any of the major experimental peaks and yielding an ESS of just 0.136. In contrast, QCxMS2 provides a reasonable ESS of 0.495 and recovers most of the key peaks, which is a good result given the challenging nature of this molecule. It should be noted that signals at m/z 308 and m/z 278 appear in the experimental spectra of HMP, with intensities varying by day and instrument. We attribute these to

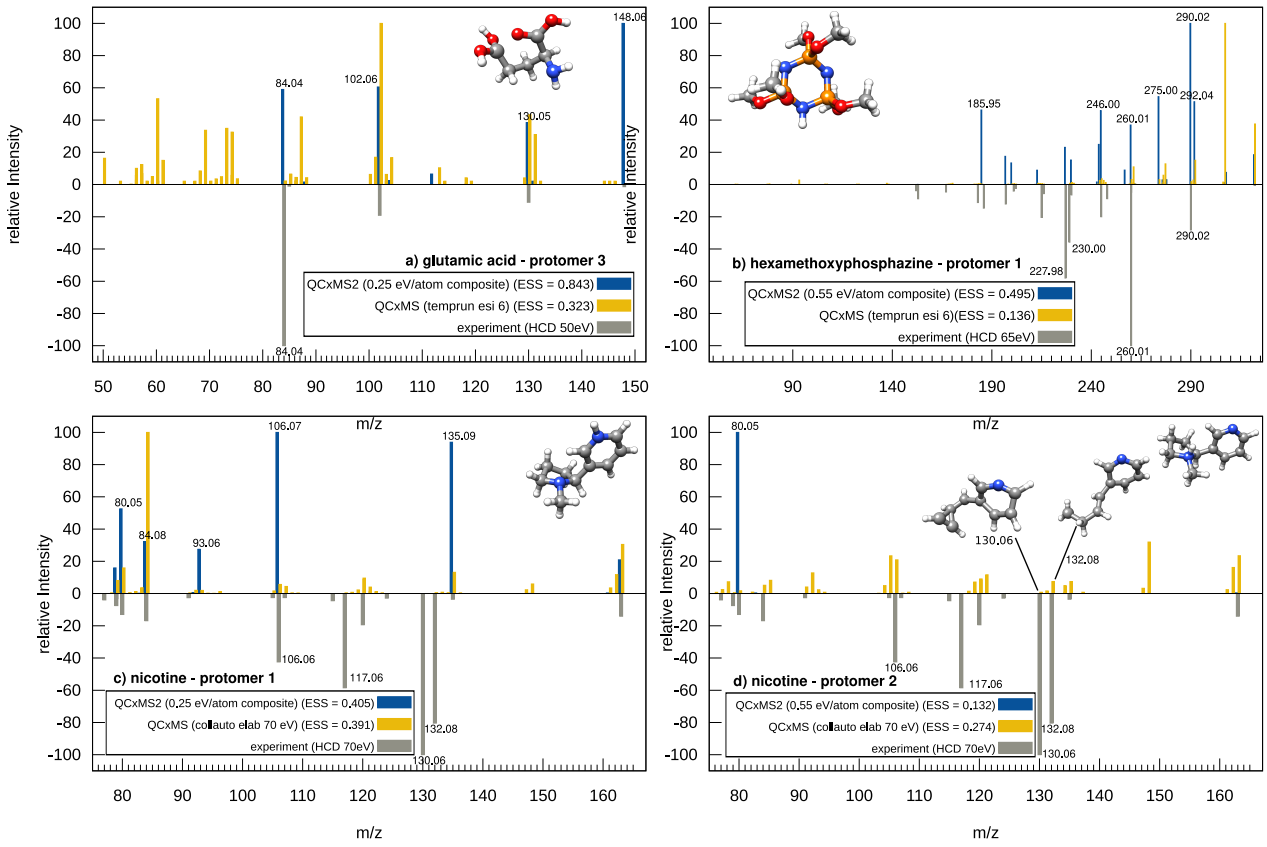


Figure 7: Calculated spectra of QCxMS2 at the ωB97X-3c//GFN2-xTB level and QCxMS at the GFN2-xTB level for a) glutamic acid, b) hexamethoxyphosphazine, and protomers **1** (c)) and **2** (d)) of nicotine. The respective program-specific energy settings are given in the spectra. Masses of theoretical spectra were shifted by $\pm$ 0.25 m/z units for better visibility. Entropy similarity scores (ESS) to the experimental spectra are also given.

ion/molecule reactions with background water and therefore exclude them from comparisons with computed spectra.

While QCxMS2 generally shows good accuracy across the spectra, there is one notable exception, namely nicotine, where a relatively low ESS of 0.407 is obtained. To investigate whether the lack of explicit collision modeling could explain this discrepancy, we computed spectra for both protomers using QCxMS with the "*collauto 6*" mode and an "*elab* of 70" eV. The resulting spectra are shown in Figure 7 **c)** and **d)**. In the QCxMS2 calculation, methylamine loss leading to the peak at m/z 132 is not predicted by the MSREACT fragment generator, and consequently, this peak is absent. In QCxMS, the ion at m/z 132 does appear for protomer **2**, albeit with low intensity, suggesting a fragmentation pathway involving multiple collisions and likely non-statistical, direct bond-breaking processes. However, even with QCxMS, the intensity of this peak is low and the ESS remains low at 0.274.

To investigate this further, we manually computed the barrier for the QCxMS-predicted fragment at the composite level, finding a high barrier of 4.98 eV, compared to just 0.74 eV for the peak at m/z 80. This substantial barrier likely explains why MSREACT does not generate this fragment and supports the idea that it arises from a non-statistical direct cleavage process. Nonetheless, it remains unclear why this particular fragmentation would be unusually prone to such a mechanism, especially given that even QCxMS predicts only a low intensity for this peak. In an experimental study, Williams and coworkers proposed a multi-step pathway involving several hydrogen rearrange-

ments to form this fragment,[32] consistent with the type of mechanism suggested by QCxMS.

Timing comparison

Finally, we discuss the computational timings using the example of the lowest-energy protomer of aspartic acid, a molecule comprising 17 atoms. QCxMS2 calculations were performed on 16 CPU cores, except for the more demanding ωB97X-3c spectra calculations, which were executed on 96 cores. For the QCxMS runs, the number of cores matched the number of trajectories. However, to enable a meaningful comparison of practical performance, all timings were scaled to a standard resource setting of 16 CPU cores.

Figure 8 shows the computational timings, normalized to 16 Intel® Xeon® "Sapphire Rapids" v4 @ 2.10 GHz CPU cores, for the spectra calculations with QCxMS2 at different levels of theory: GFN2-xTB//GFN2-xTB, the "composite" level ωB97X-3c//GFN2-xTB, and full ωB97X-3c//ωB97X-3c. Results are also shown for the recently published g-xTB method, applied for single-point and IP calculations denoted as g-xTB//GFN2-xTB, as well as for QCxMS using GFN2-xTB. The QCxMS2 calculation at the GFN2-xTB level takes approximately 66 minutes and yields a reasonable average ESS of 0.618. Refining the barriers at the ωB97X-3c level requires only about 23 minutes more computation time and significantly improves the ESS to 0.687. By contrast, computing the geometries and performing the reaction path searches entirely at the ωB97X-3c level is extremely costly, increasing the computational time to roughly 2500 hours.

With g-xTB, no refinement of close IPs at the DFT level is necessary (as supported by its very good performance on the G21IP benchmark set[68,84]), resulting in a very fast computation time of about 39 minutes while still achieving an ESS of 0.694, comparable to the DFT single-point approach. In the future, with the release of an analytical nuclear gradient, geometry optimizations at the g-xTB level will also become feasible at computational costs similar to GFN2-xTB, and are expected to yield ESS values close to those obtained from ωB97X-3c spectra.

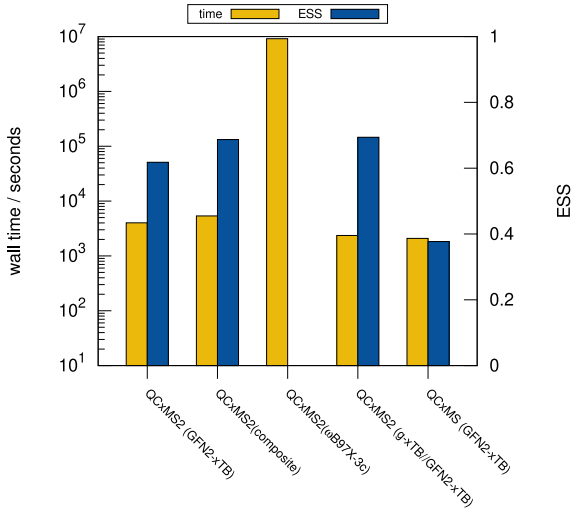


Figure 8: Computational wall times on 16 Intel® Xeon® "Sapphire Rapids" v4 @ 2.10 GHz CPU cores for the calculation of the lowest-energy protomer of aspartic acid with QCxMS2 using GFN2-xTB//GFN2-xTB, the "composite" level ωB97X-3c//GFN2-xTB, ωB97X-3c//ωB97X-3c, g-xTB//GFN2-xTB (without refinement of close IPs with ωB97X-3c) and QCxMS with GFN2-xTB and average entropy similarity scores (ESS) for best protomer and internal energy compared to experimental HCD spectra of the test set.

For comparison, the QCxMS calculation at the GFN2-xTB level takes only 35 minutes using the default 425 trajectories for a molecule of this size, but yields a notably lower average ESS of 0.377.

It should be noted that the ESS values reported for QCxMS2 were obtained using energy settings optimized against the experimental spectra. However, this process is very efficient in QCxMS2, as only the computationally inexpensive kinetic modeling of the precomputed reaction network needs to be repeated to screen different energies and fit the spectrum to the experimental SY. In contrast, QCxMS requires full MD simulations to be repeated for each new energy setting, representing a major

practical advantage of QCxMS2.

Conclusion and Outlook

The computation of CID spectra is highly valuable for supporting structure elucidation workflows, but remains computationally challenging with existing approaches. To this end, we employed the recently developed QCxMS2 program, which is based on automated reaction discovery, for the calculation of CID and HCD spectra. Herein, the energy distribution within QCxMS2 was adapted to account for the different internal energy conditions characteristic of CID and HCD experiments.

Spectra were computed for a test set of 13 organic compounds, for which both CID and HCD spectra at various collision energies were measured under consistent experimental conditions. As expected from results obtained with its predecessor QCxMS, we could reliably compute both HCD and CID spectra with good accuracy, without explicit modeling of collisions, by employing an appropriate average internal energy to approximate the experimental internal energy distribution.

Although the average internal energy matching best to the experiment is system-dependent, we achieved, on average, good entropy similarity scores (ESS) of 0.603 and 0.699 using general values of 0.45 eV per atom for HCD and 0.3 eV per atom for CID, respectively. Furthermore, QCxMS2 allows for computationally inexpensive kinetic modeling of the reaction network at different internal energies, enabling efficient adjustment to match experimental conditions. This represents a major advantage over MD-based approaches, which require costly QC calculations to be repeated for each new energy setting.

Consistent with previous studies, we found that higher-energy protomers often need to be considered to achieve good agreement with experiment. For several compounds in the test set, the mobile proton model is validated, and QCxMS2 predicts interconversion between different protomers.

At the composite theory level ωB97X-3c//GFN2-xTB, using the best internal energy settings and the best-matching protomer, QCxMS2 achieves a good ESS of 0.687 compared to the experimental HCD spectra and an excellent ESS of 0.790 for the CID spectra. The newly developed g-xTB method was also tested for single-point energy calculations and yielded results comparable to the ωB97X-3c level, but at substantially lower computational cost. Extending this method to geometry optimizations is expected to enable routine calculations for larger molecules of up to 50 atoms.

We attribute the remaining deviations from the experiment primarily to the uncertainties associated with the level of theory feasible for these studies. Only for nicotine, we attribute larger discrepancies to direct bond-breaking processes induced by collisions. For modeling such non-statistical fragmentation processes, we refer to approaches like QCxMS[22] or the CID-MD method.[20] Nevertheless, despite not explicitly simulating collisions, QCxMS2 achieves significantly higher average accuracy on the test set than QCxMS.

For future studies, we plan to implement options for computing multiply charged and negatively charged molecules. Because QCxMS2 is systematically improvable and generally more efficient than MD based simulations, it is particularly well suited for the integration of advanced electronic structure methods. The implementation of the new g-xTB method is also expected to enhance accuracy due to an improved description of transition state geometries and energies. Since QCxMS2 enables the reliable and computationally efficient calculation of CID mass spectra, it is a promising tool for generating accurate in silico spectra to support automated structure elucidation workflows.

Funding

This work was supported by the DFG grant no. 533949111, "Quantum Chemical Calculation of Mass Spectrometry via Automated Transition State Search".

Conflict of interest

The authors declare no competing financial interest.

Author contributions

Johannes Gorges: conceptualization (equal); theoretical data curation (lead); methodology (equal); software (lead); writing – original draft (lead); writing – review & editing (equal). Marianne Engeser: conceptualization (supporting); experimental data curation (lead); writing – original draft (supporting) writing – review & editing (equal). Stefan Grimme: conceptualization (equal); methodology (equal); writing – original draft (supporting); writing – review & editing (equal).

Acknowledgement This work was supported by the DFG grant no. 533949111, "Quantum Chemical Calculation of Mass Spectrometry via Automated Transition State Search". We gratefully acknowledge the access to the Marvin cluster of the University of Bonn. We thank Dr. Jeroen Koopman from FACCTs, Jonathan Schöps, and Thomas Froitzheim for fruitful discussions. Jasmin Klotz is thanked for testing the QCxMS2 program for the calculation of CID mass spectra and collection of parts of the data set. Furthermore, we thank Nguyen Thuy Duong Vu for helping with the experimental measurements.

Supporting Information Available

†: The supporting Information contains in the pdf file:
-Figures of the computed spectra, which are not contained in the main publication, as well as Tables with additional data analysis of the computed spectra
-peak list and assignment of experimental data
In the electronic supporting information:
-xyz coordinate files of all input structures of the test set taken for the calculations
-All computed and experimental spectra as csv files

Code availability

The QCxMS2 program is open-source and freely available at https://github.com/grimme-lab/QCxMS2

Data availabilty

QCxMS2 and QCxMS-computed spectra, input structures, relative free energies of protomers and experimental spectra are available at https://github.com/grimme-lab/QCxMS2-CID-data

References

(1) Urban, P. L. Quantitative mass spectrometry: an overview. *Philos. Trans. R. Soc.* **2016**, *374*, 20150382.

(2) Fenn, J. B. Electrospray wings for molecular elephants. *Angew. Chem. Int. Ed.* **2003**, *42*, 3871–3894.

(3) Kebarle, P.; Verkerk, U. H. Electrospray: from ions in solution to ions in the gas phase, what we know now. *Mass Spectrom. Rev.* **2009**, *28*, 898–917.

(4) Schäfer, M.; Drayß, M.; Springer, A.; Zacharias, P.; Meerholz, K. Radical cations in electrospray mass spectrometry: formation of open-shell species, examination of the fragmentation behaviour in ESI-MSn and reaction mechanism studies by detection of transient radical cations. *Eur. J. Org. Chem.* **2007**, *2007*, 5162–5174.

(5) Sleno, L.; Volmer, D. A. Ion activation methods for tandem mass spectrometry. *J. Mass Spectrom.* **2004**, *39*, 1091–1112.

(6) Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M.

Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **2007**, *4*, 709–712.

(7) Kind, T.; Tsugawa, H.; Cajka, T.; Ma, Y.; Lai, Z.; Mehta, S. S.; Wohlgemuth, G.; Barupal, D. K.; Showalter, M. R.; Arita, M.; others Identification of small molecules using accurate mass MS/MS search. *Mass Spectrom. Rev.* **2018**, *37*, 513–532.

(8) da Silva, R. R.; Dorrestein, P. C.; Quinn, R. A. Illuminating the dark matter in metabolomics. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 12549–12550.

(9) Lai, Y.; Koelmel, J. P.; Walker, D. I.; Price, E. J.; Papazian, S.; Manz, K. E.; Castilla-Fernández, D.; Bowden, J. A.; Nikiforov, V.; David, A.; others High-resolution mass spectrometry for human exposomics: Expanding chemical space coverage. *Environ. Sci. Technol.* **2024**, *58*, 12784–12822.

(10) Murphy, M.; Jegelka, S.; Fraenkel, E.; Kind, T.; Healey, D.; Butler, T. Efficiently predicting high resolution mass spectra with graph neural networks. *arXiv preprint arXiv:2301.11419* **2023**,

(11) Allen, F.; Greiner, R.; Wishart, D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* **2015**, *11*, 98–110.

(12) Goldman, S.; Li, J.; Coley, C. W. Generating molecular fragmentation graphs with autoregressive neural networks. *Anal. Chem.* **2024**, *96*, 3419–3428.

(13) Young, A.; Wang, F.; Wishart, D.; Wang, B.; Röst, H.; Greiner, R. FraGNNet: A deep probabilistic model for mass spectrum prediction. *arXiv preprint arXiv:2404.02360* **2024**,

(14) Nowatzky, Y.; Russo, F. F.; Lisec, J.; Kister, A.; Reinert, K.; Muth, T.; Benner, P. FIORA: Local neighborhood-based prediction of compound mass spectra from single fragmentation events. *Nat. Commun.* **2025**, *16*, 2298.

(15) Bohde, M.; Manjrekar, M.; Wang, R.; Ji, S.; Coley, C. W. DiffMS: Diffusion Generation of Molecules Conditioned on Mass Spectra. *arXiv preprint arXiv:2502.09571* **2025**,

(16) Bushuiev, R.; Bushuiev, A.; de Jonge, N.; Young, A.; Kretschmer, F.; Samusevich, R.; Heirman, J.; Wang, F.; Zhang, L.; Dührkop, K.; others MassSpecGym: A benchmark for the discovery and identification of molecules. *Adv. Neural. Inf. Process Sys.* **2024**, *37*, 110010–110027.

(17) Borges, R. M.; Colby, S. M.; Das, S.; Edison, A. S.; Fiehn, O.; Kind, T.; Lee, J.; Merrill, A. T.; Merz Jr, K. M.; Metz, T. O.; others Quantum chemistry calculations for metabolomics: Focus review. *Chem. Rev.* **2021**, *121*, 5633–5670.

(18) Hu, X.; Hase, W. L.; Pirraglia, T. Vectorization of the general Monte Carlo classical trajectory program VENUS. *J. Comput. Chem.* **1991**, *12*, 1014–1024.

(19) Lourderaj, U.; Sun, R.; Kohale, S. C.; Barnes, G. L.; de Jong, W. A.; Windus, T. L.; Hase, W. L. The VENUS/NWChem software package. Tight coupling between chemical dynamics simulations and electronic structure theory. *Comput. Phys. Commun.* **2014**, *185*, 1074–1080.

(20) Lee, J.; Tantillo, D. J.; Wang, L.-P.; Fiehn, O. Predicting Collision-Induced-Dissociation Tandem Mass Spectra (CID-MS/MS) Using Ab Initio Molecular Dynamics. *J. Chem. Inf. Model* **2024**, *64*, 7470–7487.

(21) Grimme, S. Towards first principles calculation of electron impact mass spectra of molecules. *Angew. Chem. Int. Ed.* **2013**, *52*, 6306–6312.

(22) Koopman, J.; Grimme, S. From QCEIMS to QCxMS: A tool to routinely calculate CID mass spectra using molecular dynamics. *J. Am. Soc. Mass Spectrom.* **2021**, *32*, 1735–1751.

(23) Koopman, J.; Grimme, S. Calculation of mass spectra with the QCxMS method for negatively and multiply charged molecules. *J. Am. Soc. Mass Spectrom.* **2022**, *33*, 2226–2242.

(24) Schreckenbach, S. A.; Anderson, J. S.; Koopman, J.; Grimme, S.; Simpson, M. J.; Jobst, K. J. Predicting the Mass Spectra of Environmental Pollutants Using Computational Chemistry: A Case Study and Critical Evaluation. *J. Am. Soc. Mass Spectrom.* **2021**, *32*, 1508–1518.

(25) Wang, S.; Lin, C.; Zhao, L.; Gong, X.; Zhang, M.; Zhang, H.; Hu, P. Identifying isomers in Chinese traditional medicine via density functional theory and ion fragmentation simulation software QCxMS. *J. Chromatogr. A* **2024**, *1730*, 465122.

(26) Gorges, J.; Grimme, S. QCxMS2 – a program for the calculation of electron ionization mass spectra via automated reaction network discovery. *Phys. Chem. Chem. Phys.* **2025**, *27*, 6899–6911.

(27) Rosenstock, H. M.; Wallenstein, M. B.; Wahrhaftig, A. L.; Eyring, H. Absolute Rate Theory for Isolated Systems and the Mass Spectra of Polyatomic Molecules. *Proc. Natl. Acad. Sci.* **1952**, *38*, 667–678.

(28) Asakawa, D.; Mizuno, H.; Sugiyama, E.; Todoroki, K. Fragmentation study of tryptophan-derived metabolites induced by electrospray ionization mass spectrometry for highly sensitive analysis. *Analyst* **2021**, *146*, 2292–2300.

(29) Asakawa, D.; Todoroki, K.; Mizuno, H. Fragmentation of protonated histamine and histidine by electrospray ionization in-source collision-induced dissociation. *J. Am. Soc. Mass Spectrom.* **2022**, *33*, 1716–1722.

(30) Zhang, P.; Chan, W.; Ang, I. L.; Wei, R.; Lam, M. M.; Lei, K. M.; Poon, T. C. Revisiting fragmentation reactions of protonated α-amino acids by high-resolution electrospray ionization tandem mass spectrometry with collision-induced dissociation. *Scientific reports* **2019**, *9*, 6453.

(31) Smyth, T. J.; Ramachandran, V.; McGuigan, A.; Hopps, J.; Smyth, W. F. Characterisation of nicotine and related compounds using electrospray ionisation with ion trap mass spectrometry and with quadrupole time-of-flight mass spectrometry and their detection by liquid chromatography/electrospray ionisation mass spectrometry. *Rapid Commun. Mass. Spectrom.* **2007**, *21*, 557–566.

(32) Williams, J. P.; Nibbering, N. M.; Green, B. N.; Patel, V. J.; Scrivens, J. H. Collision-induced fragmentation pathways including odd-electron ion formation from desorption electrospray ionisation generated protonated and deprotonated drugs derived from tandem accurate mass spectrometry. *J. Mass Spectrom.* **2006**, *41*, 1277–1286.

(33) Celma, C.; Allue, J.; Prunonosa, J.; Peraire, C.; Obach, R. Simultaneous determination of paracetamol and chlorpheniramine in human plasma by liquid chromatography–tandem mass spectrometry. *J. Chromatogr. A* **2000**, *870*, 77–86.

(34) Li, H.; Zhang, C.; Wang, J.; Jiang, Y.; Fawcett, J. P.; Gu, J. Simultaneous quantitation of paracetamol, caffeine, pseudoephedrine, chlorpheniramine and cloperastine in human plasma by liquid chromatography–tandem mass spectrometry. *J. Pharm. Biomed. Anal.* **2010**, *51*, 716–722.

(35) Ásgeirsson, V.; Bauer, C. A.; Grimme, S. Unimolecular decomposition pathways of negatively charged nitriles by ab initio molecular dynamics. *Phys. Chem. Chem. Phys.* **2016**, *18*, 31017–31026.

(36) Lee, J.; Tantillo, D. J.; Wang, L.-P.; Fiehn, O. Impact of Protonation Sites on Collision-Induced Dissociation-MS/MS Using CIDMD Quantum Chemistry Modeling. *J. Chem. Inf. Model.* **2024**, *64*, 7457–7469.

(37) Seritan, S.; Bannwarth, C.; Fales, B. S.; Hohenstein, E. G.; Isborn, C. M.; Kokkila-Schumacher, S. I.; Li, X.; Liu, F.; Luehr, N.; Snyder Jr, J. W.; others TeraChem: A graphical processing unit-accelerated electronic structure package for large-scale ab initio molecular dynamics. *WIREs Comp. Mol. Sci.* **2021**, *11*, e1494.

(38) Drahos, L.; Vékey, K. MassKinetics: a theoretical model of mass spectra incorporating physical processes, reaction kinetics and mathematical descriptions. *J. Mass Spectrom.* **2001**, *36*, 237–263.

(39) Kuki, Ákos and Shemirani, Ghazaleh and Nagy, Lajos and Antal, Borbála and Zsuga, Miklós and Kéki, Sándor Estimation of Activation Energy from the Survival Yields: Fragmentation Study of Leucine Enkephalin and Polyethers by Tandem Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2013**, *24*, 1064–1071, PMID: 23661424.

(40) Box, G. E.; Muller, M. E. A note on the generation of random normal deviates. *Ann. Math. Stat.* **1958**, *29*, 610–611.

(41) Meroueh, S. O.; Wang, Y.; Hase, W. L. Direct dynamics simulations of collision- and surface-induced dissociation of N-protonated glycine. Shattering fragmentation. *J. Phys. Chem. A* **2002**, *106*, 9983–9992.

(42) Martin-Somer, A.; Martens, J.; Grzetic, J.; Hase, W. L.; Oomens, J.; Spezia, R. Unimolecular fragmentation of deprotonated diproline [Pro2-H]- studied by chemical dynamics simulations and IRMPD spectroscopy. *J. Phys. Chem. A* **2018**, *122*, 2612–2625.

(43) Vékey, K. Internal energy effects in mass spectrometry. *J. Mass Spectrom.* **1996**, *31*, 445–463.

(44) Armentrout, P. Threshold collision-induced dissociations for the determination of accurate gas-phase binding energies and reaction barriers. *Modern mass spectrometry* **2003**, 233–262.

(45) Gross, J. H. *Mass spectrometry: a textbook*; Springer Science & Business Media, 2006.

(46) Eyring, H. The activated complex in chemical reactions. *J. Chem. Phys.* **1935**, *3*, 107–115.

(47) Bure, C.; Lange, C. Comparison of dissociation of ions in an electrospray source, or a collision cell in tandem mass spectrometry. *Current Organic Chemistry* **2003**, *7*, 1613–1624.

(48) Parcher, J. F.; Wang, M.; Chittiboyina, A. G.; Khan, I. A. In-source collision-induced dissociation (IS-CID): Applications, issues and structure elucidation with single-stage mass analyzers. *Drug testing and analysis* **2018**, *10*, 28–36.

(49) Program package for the quantum mechanical calculation of EI mass spectra using automated reaction network exploration `qcxms2`. `https://github.com/grimme-lab/QCxMS2`, Accessed: 2025-7-1.

(50) Bannwarth, C.; Ehlert, S.; Grimme, S. 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.* **2019**, *15*, 1652–1671.

(51) Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. Extended tight-binding quantum chemistry methods. *WIREs Comput. Mol. Sci.* **2020**, *11*, e01493.

(52) Müller, M.; Hansen, A.; Grimme, S. ωB97X-3c: A composite range-separated hybrid DFT method with a molecule-optimized polarized valence double-ζ basis set. *J. Chem. Phys.* **2023**, *158*, 014103.

(53) Automated and efficient quantum chemical determination and energetic ranking of molecular protonation sites. *J. Comput. Chem.* **2017**, *38*, 2618–2631.

(54) Pracht, P.; Grimme, S.; Bannwarth, C.; Bohle, F.; Ehlert, S.; Feldmann, G.; Gorges, J.; Müller, M.; Neudecker, T.; Plett, C.; others CREST—A program for the exploration of low-energy molecular chemical space. *J. Chem. Phys.* **2024**, *160*, 114110.

(55) van Staalduinen, N.; Bannwarth, C. MolBar: a molecular identifier for inorganic and organic molecules with full support of stereoisomerism. *Digit. Discov.* **2024**, *3*, 2298–2319.

(56) A Molecular Identifier for Inorganic and Organic Molecules with Full Support of Stereoisomerism. `https://git.rwth-aachen.de/bannwarthlab/molbar`, Accessed: 2024-10-29.

(57) Mardirossian, N.; Head-Gordon, M. 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.* **2014**, *16*, 9904–9924.

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

(59) Weigend, F.; Ahlrichs, R. 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.* **2005**, *7*, 3297.

(60) Grimme, S.; Hansen, A.; Ehlert, S.; Mewes, J.-M. r2SCAN-3c: A "Swiss army knife" composite electronic-structure method. *J. Chem. Phys.* **2021**, *154*, 064103.

(61) Grimme, S. Supramolecular binding thermodynamics by dispersion corrected density functional theory. *Chem. Eur. J.* **2012**, *18*, 9955–9964.

(62) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. 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. *Phys. Chem. B* **2009**, *113*, 6378–6396.

(63) CREST - A program for the automated exploration of low-energy molecular chemical space. `https://github.com/crest-lab/crest`, Accessed: 2025-1-16.

(64) Neese, F. Software update: The ORCA program system—Version 5.0. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2022**, *12*, e1606.

(65) Vahtras, O.; Almlöf, J.; Feyereisen, M. W. Integral approximations for LCAO-SCF calculations. *Chem. Phys. Lett.* **1993**, *213*, 514–518.

(66) Weigend, F. Accurate Coulomb-fitting basis sets for H to Rn. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057–1065.

(67) Neese, F.; Wennmohs, F.; Hansen, A.; Becker, U. Efficient, approximate and parallel Hartree–Fock and hybrid DFT calculations. A 'chain-of-spheres' algorithm for the Hartree–Fock exchange. *Chem. Phys.* **2009**, *356*, 98–109.

(68) Froitzheim, T.; Müller, M.; Hansen, A.; Grimme, S. g-xTB: A General-Purpose Extended Tight-Binding Electronic Structure Method For the Elements H to Lr (Z=1–103). ChemRxiv, 2025; `https://doi.org/10.26434/chemrxiv-2025-bjxvt`.

(69) A general-purpose semiempirical quantum mechanical method gxtb. https://github.com/grimme-lab/g-xtb, Accessed: 2025-6-25.

(70) Interpolation of molecular geometries through geodesics in redundant internal coordinate hyperspace for complex transformations. https://github.com/virtualzx-nad/geodesic-interpolate, Accessed: 2024-10-29.

(71) Semiempirical Extended Tight-Binding Program Package xtb. https://github.com/grimme-lab/xtb, Accessed: 2025-6-1.

(72) Light-weight tight-binding framework tblite. https://github.com/tblite, Accessed: 2025-06-12.

(73) Quantum mechanic mass spectrometry calculation program. https://github.com/qcxms, Accessed: 2024-10-29.

(74) Plot Mass Spectra (PlotMS) plotting program for the QCxMS program. https://github.com/qcxms/PlotMS, Accessed: 2025-06-27.

(75) Li, Y.; Kind, T.; Folz, J.; Vaniya, A.; Mehta, S. S.; Fiehn, O. Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. *Nat. Methods* **2021**, *18*, 1524–1531.

(76) Li, Y.; Fiehn, O. Flash entropy search to query all mass spectral libraries in real time. *Nat. Methods* **2023**, *20*, 1475–1478.

(77) Spectral entropy for mass spectrometry data. https://github.com/YuanyueLi/MSEntropy, Accessed: 2024-10-29.

(78) Wyttenbach, T.; Bowers, M. T. Structural stability from solution to the gas phase: native solution structure of ubiquitin survives analysis in a solvent-free ion mobility–mass spectrometry environment. *Phys. Chem. B* **2011**, *115*, 12266–12275.

(79) Warnke, S.; Seo, J.; Boschmans, J.; Sobott, F.; Scrivens, J. H.; Bleiholder, C.; Bowers, M. T.; Gewinner, S.; Schoöllkopf, W.; Pagel, K.; others Protomers of benzocaine: solvent and permittivity dependence. *J. Am. Chem. Soc.* **2015**, *137*, 4236–4242.

(80) Shelimov, K. B.; Clemmer, D. E.; Hudgins, R. R.; Jarrold, M. F. Protein structure in vacuo: gas-phase conformations of BPTI and cytochrome c. *J. Am. Chem. Soc.* **1997**, *119*, 2240–2248.

(81) Jarrold, M. F. Unfolding, refolding, and hydration of proteins in the gas phase. *Acc. Chem. Res.* **1999**, *32*, 360–367.

(82) Dongré, A. R.; Jones, J. L.; Somogyi, Á.; Wysocki, V. H. Influence of peptide composition, gas-phase basicity, and chemical modification on fragmentation efficiency: evidence for the mobile proton model. *J. Am. Chem. Soc.* **1996**, *118*, 8365–8374.

(83) Bahrami, H.; Farrokhpour, H. Corona discharge ionization of paracetamol molecule: Peak assignment. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2015**, *135*, 646–651.

(84) Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. 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.* **2017**, *19*, 32184–32215.

Acknowledgements

First and foremost, I would like to thank my supervisor and Doktorvater Prof. Dr. Stefan Grimme for his invaluable guidance throughout my doctoral studies. From my bachelor thesis onward, he inspired me with his unique perspective on how to approach and solve scientific problems. An influence that proved especially important for the challenging topic of simulating mass spectrometry.

I am also grateful to Prof. Dr. Thomas Bredow for kindly agreeing to review this thesis as the second examiner.

Special thanks go to PD Dr. Marianne Engeser, with whom I collaborated on the evaluation of QCxMS2 for CID spectra, and to Dr. Andreas Hansen for our fruitful work together on the LNCI16 and HS13L benchmarks.

During my time in the group, I had the pleasure of working with several talented students and contributing to the supervision of their thesis projects. I would particularly like to acknowledge Benedikt Bädorf for his outstanding work on the LNCI16 benchmark and for becoming a valued colleague thereafter, as well as Jasmin Klotz and Jonathan Schöps, who bravely took on the topic of mass spectrometry computation with me.

I also want to thank Dr. Jeroen Koopman, from whom I took over the mass spectra project, for both scientific and moral support. This was indeed a frustrating but also rewarding topic.

Furthermore, I would like to thank the members of our group for their support throughout the years. The time we shared was both enjoyable and enriching, and it helped me through the struggles of doing a PhD. In particular, the long evenings on our Dachterrasse filled with scientific (and sometimes not so scientific) discussions were one of the highlights of my time at the Mulliken Center. Dr. Sebastian Ehlert, Dr. Joachim Laun, Dr. Sebastian Spicher, Dr. Hagen Neugebauer, Dr. Thomas Rose, Dr. Julius Kleine Büning, Dr. Philipp Pracht, Dr. Fabian Bohle, Dr. Jan Mewes, my fellow Cala-Dor-enjoyer Dr. Christian Hölzer, my long-time office mate Dr. Julia Kohn, my companion on this journey since the very beginning Thomas Gasevic, my fellow metalhead Tim Schramm, my coffee companion Robin Dahl, Thomas Froitzheim, Christoph Plett, Marcel Stahn, Marcel Müller, Marvin Friede, Lukas Wittmann, Christian Selzer, and Abylay aka Albert Katbashev – to all of you, thank you for your dedication, stimulating discussions, and the many shared moments that made this time so rewarding. For administrative and technical support, I thank Claudia Kronz and Jens Mekelburger.

For proof-reading this thesis, I would like to thank Thomas Gasevic, Dr. Thomas Rose, Tim Schramm, Thomas Froitzheim, Jonathan Schöps, and my sister Katharina.

Finally, beyond the scientific community, I thank my family for their unconditional love and support, without which this work would not have been possible. I am also deeply grateful to my friends for their companionship and encouragement throughout this journey.