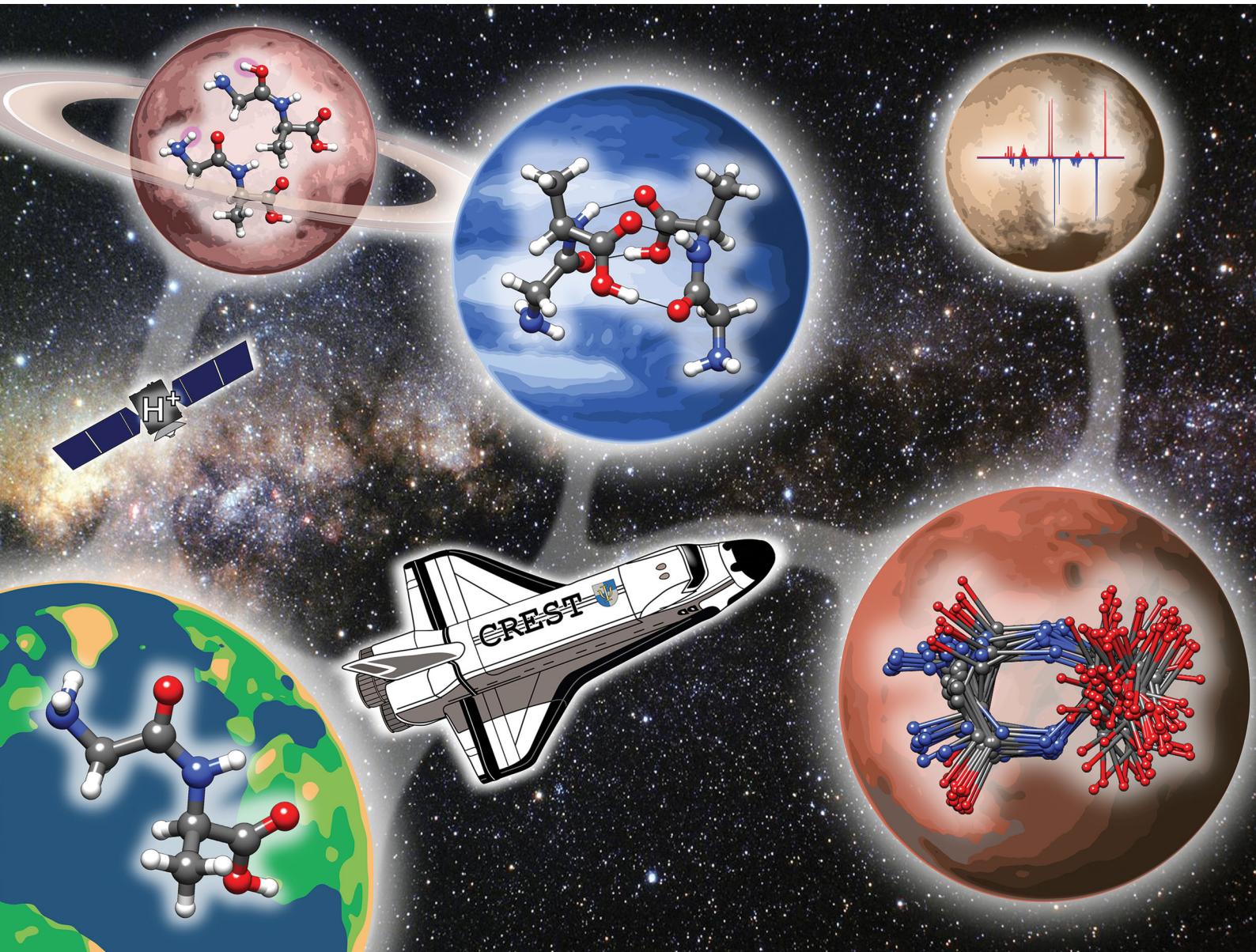


PCCP

Physical Chemistry Chemical Physics

rsc.li/pccp



ISSN 1463-9076



Cite this: *Phys. Chem. Chem. Phys.*,
2020, 22, 7169

Automated exploration of the low-energy chemical space with fast quantum chemical methods

Philipp Pracht, Fabian Bohle and Stefan Grimme *

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

Received 20th December 2019,
Accepted 12th February 2020

DOI: 10.1039/c9cp06869d

rsc.li/pccp

1 Introduction

Over the past decades computational methods became a valuable tool in many modern fields of chemistry, and some kind of quantum chemical (QC) calculation can be found in almost every new publication. The big popularity of computational chemistry is also founded on recent advances in density functional theory (DFT) methods, which nowadays can routinely provide gas- or condensed-phase structures and energies for roughly a few hundred atoms.^{1,2} For many interesting applications in biochemistry or supramolecular chemistry, however, those calculations are still too expensive. Classical force-fields (FFs) are often employed as alternatives for long molecular dynamics simulations³ but their limitations are manifold and they are not suited for general use, *e.g.*, for metallic systems. Additionally chemoinformatic procedures are used in the drug discovery with increasing popularity.^{4–7}

Among the most important application of low-cost atomistic methods is the large scale structural sampling of molecular geometries, *i.e.*, the generation of an ensemble of low-energy

structures, generally referred to as conformers. The knowledge about a molecules' conformations is highly important since all its properties are rooted in a thermodynamic ensemble average of the properties of its conformers that are accessible at finite temperature.^{8,9} This gave rise to a large framework of computational approaches to generate and screen three-dimensional molecular structures where many challenges have to be faced in the generation process, such as the correct distinction between different conformers or the handling of macrocyclic systems. Hence, a huge number of conformer generators based on different algorithmic approaches is available today.^{10–19} One of the most common types of conformer generators are knowledge-based algorithms, which chemoinformatically try to reproduce structures from reference data (often taken from experimental crystal structures), or generate structures based on heuristic rules.⁹ These approaches, sometimes also referred to as systematic methods, have the advantage of very short computation times, but are generally lacking a physically motivated methodology and thus are often not generally applicable. An example for this is the treatment of macrocyclic molecules, where special heuristics are required.^{20–23} By contrast a general workflow based on quantum chemical calculations has no need for specialized rules and should recover any structural information of the molecule by an analysis of

Mulliken Center for Theoretical Chemistry, Universität Bonn, Beringstr. 4,
53115 Bonn, Germany. E-mail: grimme@thch.uni-bonn.de; Tel: +49 0228 732351

the potential energy surface (PES). Furthermore such general algorithms can also be exploited in order to find different arrangements of noncovalent bound aggregates or to generate conformations under geometrical constraints.

A related problem to conformational sampling is the prediction of molecular protonation and deprotonation sites. While there are many approaches to obtain these sites based on knowledge, *i.e.*, databases, the computational QC approach is just as simple.^{24–26} Here the different structures in the chemical space have to be (automatically) generated and can then be ranked by their total energy (*i.e.*, proton affinities).^{27–29} These so-called protomers are linked to some other important properties such as the pK_a value.^{30–34} Furthermore a sequence of protonation and deprotonation at different positions can be employed to obtain all prototropic tautomers of a molecule. By employing QC models that can intrinsically form and break bonds, such a tautomer screening procedure has the advantage that even complicated rearrangements can be recovered, which would otherwise require complicated rules in chemoinformatic treatments.^{35,36}

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

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

2 Automatized quantum chemical procedures in the literature

Although there are many chemoinformatic schemes for the exploration of chemical space, there exist only a few automatized procedures driven by quantum chemical calculations. The reason

for this is the enormous amount of required calculations, leading to very high computational cost. However, the obvious inherent advantage of QC schemes is the possibility to generate and predict results in an *ab initio* fashion. They also yield much more chemical insight than purely heuristics guided results. There are several noteworthy efforts by our^{29,40–44} and other groups^{45–55} for efficient quantum chemistry driven algorithms to calculate various properties. One very large field of these automatized applications is the exploration of reaction mechanisms by so-called reaction networks, which recently gained popularity, *e.g.*, due to work by the groups of Maeda^{45,46,56–61} and Reiher.^{47,62–65} These reaction networks function by a throughout exploration of the normally reactive part of the PES under some pre-defined (*i.e.*, heuristics-guided) criteria, such as energy cut-offs or bias potentials.^{45,57} QC methods are primarily applied in order to perform structure optimizations and transition-state (TS) searches.^{46,59,62,63} Graph-based heuristic descriptors, from which vertices (*i.e.*, intermediate points on the different PES) and edges (*i.e.*, reactions) are created, are used to generate the eponymous networks, where a connected subgraph represents a single PES of the reaction. New intermediate structures are generated from a single starting point (referred to as the zeroth-generation structure) by identifying reactive sites and placing these in close proximity during a structure optimization. Here it is important to consider the relative orientation of the fragments. In the geometry optimization either the reactants are recovered or a new species is formed, which is detected and yields an approximate reaction path. The exploration of reaction paths can also be refined by the inclusion of conformational sampling.⁶³ Recently the concept of chemical reaction networks was extended with the application of a kinetic analysis by using semiempirical QC methods.^{66,67} This is done for the reduction of noise within the possible reaction pathways by removal of kinetically unfavorable structures. The motivation here is to reduce the computational cost as far as possible, while still maintaining a reasonable degree of accuracy, which is in fact also one of the main motivations of our work. All the different tasks and the building of the network itself can be automatized and parallelized in an efficient computer code, which makes the procedure feasible even at the underlying DFT or wave function theory (WFT) level. The automatized exploration of reaction mechanisms is a huge field in computational chemistry and more comprehensive reviews can be found in the literature (see ref. 46, 52, 66, 68 and 69).

Another approach to automated QC is to provide the infrastructure for computational workflows. Representative programs here are for example the PyADF⁵³ and QMflow⁵⁵ frameworks as pioneered by Visscher *et al.* or the atomistic simulation environment ASE.⁷⁰ In general these type of programs allow the setup of multiple quantum chemical calculations in a script-like manner, coupled to some semi-automated analysis of the results. However, many large computational chemistry program packages today also come with their own scripting environments such as the PLAMS driver distributed with ADF2019.^{71,72} Although the setup of these automation is certainly less exhausting than conventional

scripting, it can still be tedious and requires a large amount of user input. If, however, the setup was done once it can be re-used for multiple calculations.

The research in our own group concerning automated processes so far mainly focused on the generation of spectral data, such as UV-vis/circular dichroism (CD),⁷³ nuclear magnetic resonance (NMR),⁴⁴ and mass spectra.^{40–42} With the introduction of the tight-binding method GFN-xTB³⁷ (short for Geometries, Frequencies, and Noncovalent interactions – extended Tight-Binding) the efficient automation of screening processes at a semiempirical QC level of theory became feasible and was applied to conformers, protomers, and tautomers including all elements up to Radon. The first SQM based conformer generator was published in scope of the automated calculation of NMR spectra,⁴⁴ while the screening of protonation sites and tautomers/isomers had been published in ref. 29 and 43 accordingly. Since then, the procedures have been refined and implemented into a single computer code (*i.e.*, CREST), which is the subject of this work.

3 The automated conformational search algorithm

We introduce a new tool called CREST (short for Conformer-Rotamer Ensemble Sampling Tool), for the automatized

exploration of the low-energy chemical structure space normally not consisting of any covalent bond break/formation. As its name implies, the main application of CREST is the generation of conformer ensembles with an algorithm called iMTD-GC,⁷⁴ but other related applications, such as the screening of different noncovalently bound aggregates, or the screening for different protonation sites are also implemented. An overview of the different procedures and their general workflows is shown in Fig. 1.

3.1 Identification of conformer ensembles

Stereoisomers of a molecule that differ only in their conformation but have the same covalent topology are referred to as conformers and can be characterized by a distinct potential energy minimum. By rotation around covalent chemical bonds (or other complicated inversion-type processes) that interchange nuclei belonging to the same group of nuclides, as for example the interchange of H nuclei at a methyl group, so called rotamers arise. Rotamers have degenerate potential energy minima and thus are indistinguishable by any nuclear spin-independent quantum mechanical observable computed at the respective minima (see Fig. 2). In the following the Born–Oppenheimer-approximated equilibrium conformer including all its rotamers is referred to as “conformation”. A set of different conformers and their rotamers within a certain energy

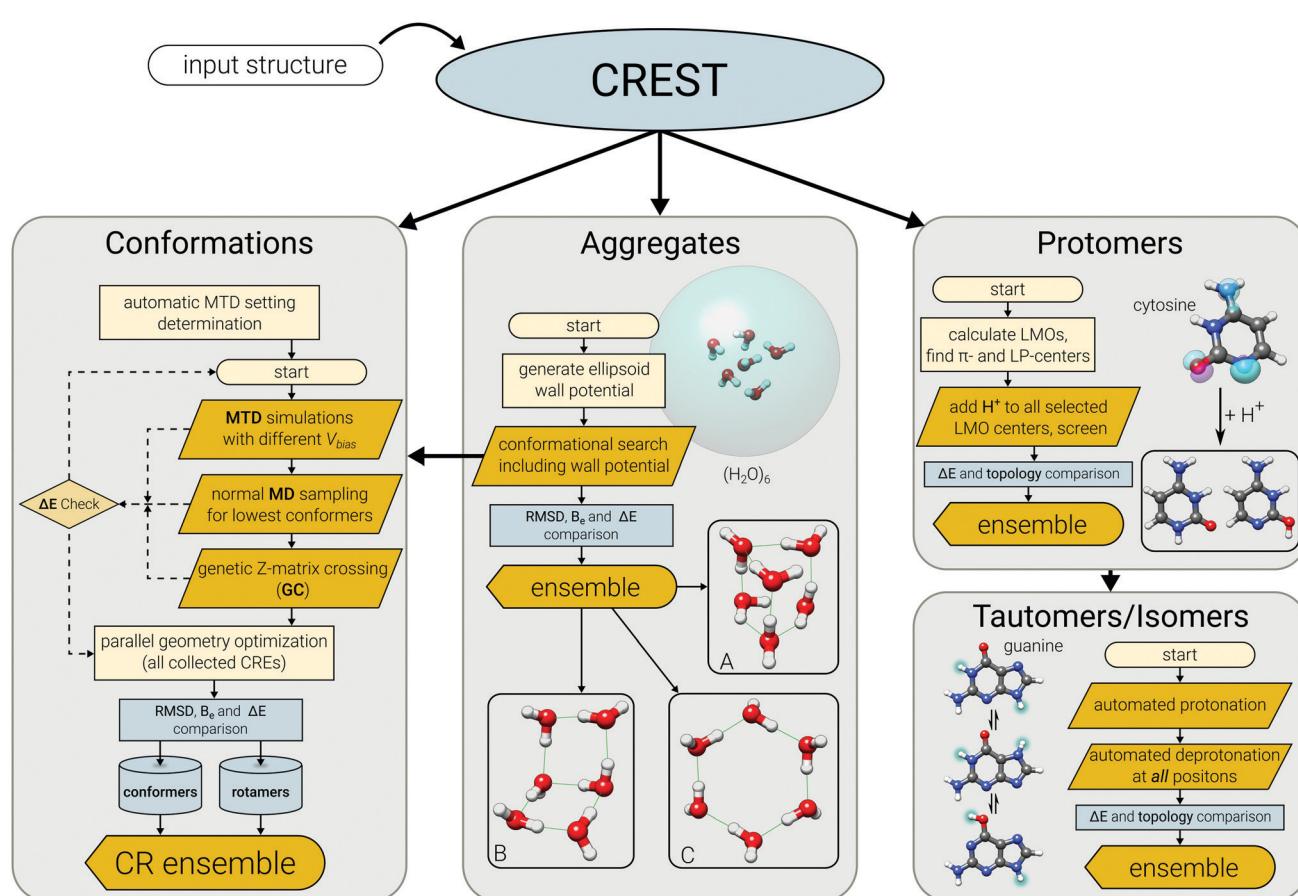


Fig. 1 Overview of the automatized quantum chemical sampling procedures that have been implemented in the CREST code.

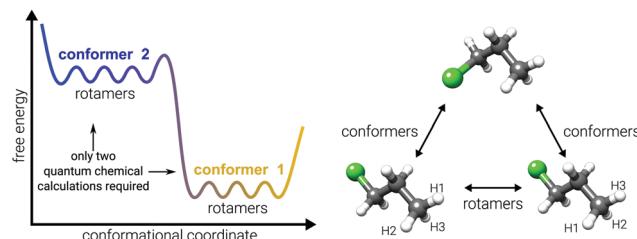


Fig. 2 Relation of conformers and rotamers using the example of 1-chloropropane and their schematic representation on the potential energy surface.

window around the same global covalent potential energy minimum is referred to as the conformer/rotamer ensemble (CRE).

For the calculation of properties it is often necessary to include different molecular conformations by averaging the individually obtained Boltzmann weighted property of each constituent in the ensemble. Some examples where this ensemble average is highly relevant are nuclear magnetic resonance (NMR) spectra,^{44,75–77} circular dichroism (CD) spectra,^{73,78–80} or pK_a values.^{43,81} In order to avoid double counting, which leads to incorrect Boltzmann averages and subsequently falsely averaged properties, the precise distinction between identical structures, conformers, and rotamers is mandatory. This distinction is possible on the basis of three dimensional structures and (free) energies of the isomers. The energy is employed as a criterion, since each conformer is characterized by its own minimum on the PES. For purely structure based comparisons the root-mean-square deviation (RMSD) of atomic Cartesian coordinates and the difference between rotational constants (B_e) of two molecules can be used. However, structural parameters do not include any information whether the structure is an equilibrium geometry or some higher-energetic intermediate.⁸² Therefore, structural information must always be combined with the energetics for correct identification of different conformations. In practice, however, many knowledge-based conformer generators still only employ structural criteria (two or three dimensional) for the distinction of conformers,^{10,19,83} which can be useful, *e.g.*, for large scale screening of databases.^{9,84}

The distinction between identical isomers, conformers, and rotamers on the basis of the energy, the atomic Cartesian RMSD and the molecular rotational constant is schematically outlined in Fig. 3. For practical reasons one has to work with predefined thresholds in order to eliminate the effect of numerical noise. Conformers are those structures that either have different PES minima ($\Delta E > E_{\text{thr}}$), or, if the energy difference is small, a high RMSD and unequal B_e . Two structures with similar energetics can be rotamers if their atomic coordinates differ, but at the same time the rotational constants are equal. Only if two structures have the same energy and matching structural criteria ($\text{RMSD} \approx 0$ and $\Delta B_e \approx 0$), they can be discarded as duplicates. The final CRE for further practical calculations typically consists of all unique conformers and rotamers within a certain energy window. The choice for this window depends on the accuracy of the used QC and the type of application (*i.e.*, sensitivity of the target property to details of the conformational ensemble).

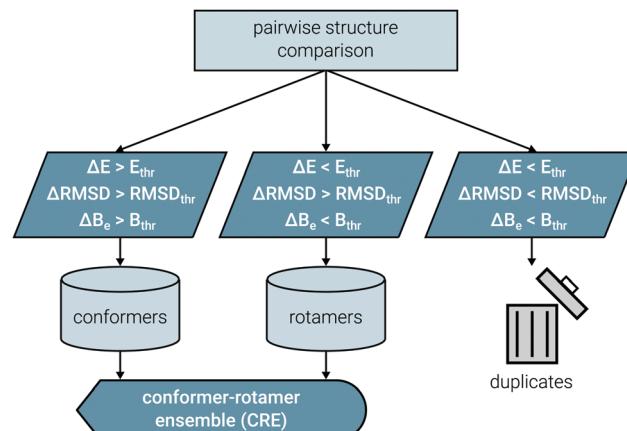


Fig. 3 Schematic representation of the sorting criteria to distinguish between identical structures, conformers and rotamers. E_{thr} , RMSD_{thr} , and B_{thr} are the respective predefined thresholds for the energy, atomic RMSD between the considered pair, and rotational constant.

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

$$S_{\text{CR}} = R \sum_{i=1}^{\text{CRE}} p_i \log p_i, \quad (1)$$

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

$$p_i = \frac{\exp(-\Delta E_i / RT)}{\sum_{j=1}^{\text{CRE}} \exp(-\Delta E_j / RT)}. \quad (2)$$

The ensemble entropy S_{CR} is also linked to the ensemble free energy (at $T = 298$ K) $G_{\text{CR}} = -TS_{\text{CR}}$, which is minimized for a complete CRE. This completeness criterion of a maximized S_{CR} only holds if the global minimum conformation is included in the ensemble and breaks down otherwise. Therefore, finding the lowest energy conformation of a molecule is the one of the defining tasks a conformer generator must be able to perform robustly. A maximized ensemble entropy was used for determining technical parameter sets that are employed in the CREST program and is discussed in ref. 74.

Practically, it is difficult to assign a quality and/or completeness to an ensemble, without knowing the “true” conformations. The practical approach for identifying the “true” low-energy conformations are experimental measurements. Crystal structures determined from X-ray are the most common source of experimental geometries. However, interpretation of CREs with respect to crystal structures can be highly problematic, since conformations in the solid can differ significantly from structures in the gas-phase or in solution, *e.g.*, due to packing effects.^{85,86} Furthermore, crystal structures intrinsically yield only one or a few conformers instead of entire ensembles. Other experimental techniques for structure elucidation, such as solution NMR (*e.g.*, using variable temperature NMR, NOESY, residual dipolar and 3J scalar couplings), microwave

spectroscopy, and gas electron diffraction, are less common. In general experimental conditions will always have an influence on the composition of the ensemble. Even under ideal experimental conditions identifying the global minimum might not be possible, *e.g.*, due to kinetic trapping. For these reasons theoretical calculations provide a valuable alternative to experiment for obtaining CREs under idealized environment. In the literature one can find a variety of benchmark studies^{84,87–90} where computer generated conformers are compared to experimentally observed ones in order to assess the performance of different conformer generators. Although a comparison like this gives some insight into the ensembles, it has to be evaluated with caution in respect to the ensemble completeness and the different “measurement” conditions. Therefore in the following sections, rather than benchmarking on crystal structures, we will compare selected conformer ensembles in the gas or liquid phase with spectroscopically evaluated structures.

3.2 Algorithmic details

Generating conformations, *e.g.*, by rotation around dihedral angles is impractical for larger and flexible molecules. In addition this approach requires the manual *a priori* definition of the conformational coordinates (*i.e.*, the angles). To remedy this, we recently proposed a meta-dynamics (MTD) simulation based screening procedure that can be routinely used for the generation of molecular conformations in the gas-phase or in implicit solvation.⁷⁴ A history-dependent biasing potential is applied, where the collective variables (CVs) for the meta-dynamics are previous structures on the PES, expressed as atomic RMSD between them, which is calculated according to a quaternion algorithm.⁹¹ Although being a well-known concept of MD simulations^{92,93} and being used before in the general context of conformational changes,^{94–98} it is, to our knowledge, the first combination of MTD simulations with atomic RMSDs in order to generate conformers. The biasing contribution is given in the form of a Gaussian potential by

$$V_{\text{bias}} = \sum_i^n k_i \exp(-\alpha_i \Delta_i^2), \quad (3)$$

where the RMSDs enter as collective variables Δ_i , n is the number of reference structures, k_i is the pushing strength and the parameter α_i determines the potentials’ shape. From this energy expression atomic forces are derived that enter as additional forces in the MTD simulations, which is also sometimes referred to as guiding forces.⁹⁵ Since the addition of each bias Gaussian potential drives the structure further away from previous geometries this allows otherwise unlikely high-barrier crossings where all atoms collectively explore huge regions of the PES. A schematic representation of a 1-dimensional PES that is filled by additive bias potentials over time is given in Fig. 4. For more realistic examples see Sections 4 and 6.1.

While calculations in this manuscript were conducted entirely at the semiempirical tight-binding level, it has to be noted that the application of a RMSD-based bias potential is a general approach that, in principle, works at all levels of theory,

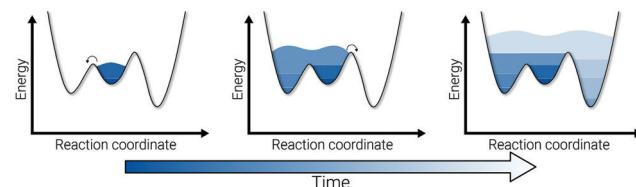


Fig. 4 Schematic one dimensional PES that is “filled” by several bias potentials over time, which allows larger barrier heights to be overcome.

i.e., at the FF, DFT or even WFT level. Furthermore, the employed CVs do not have to be atomic RMSDs, but could also be some other kind of alignment factor between structures, as long as forces can be obtained from the partial derivatives $\frac{\partial CV}{\partial r}$.

In the literature one can also find other MD based approaches with various bias potentials under the general keyword accelerated molecular dynamics (aMD).^{99–107} The basic idea of those aMD approaches is to smooth the PES and to fill its minima by reshaping the potential. This is done in order to decrease inter-conformational energy barriers and enable the simulation to explore larger regions of the PES. The aMD approach was already successfully applied in the generation of molecular conformations, *e.g.*, for macrocycles.¹⁰⁸ The fundamental difference of aMD to the RMSD based MTD approach is the missing directionality of the PES exploration. In aMD the shape of the energy surface is in general retained after addition of the bias potential, *i.e.*, during the simulation it is possible to arrive at the same minimum on the PES again. With the RMSD based approach however, previous minima on the PES are “occupied” by the V_{bias} , leading to history dependent forces and thus to an implied directionality of the simulation. Since the potential in aMD is modified by a single bias potential and retains the general shape of the energy surface, very long simulation times can still be required in order to sample the entire conformational space. This time can be expected to be much shorter with a history dependent guiding force. It must be stressed however, that the target quantities of the two approaches are slightly different: in aMD, the desired quantity usually is the canonical ensemble average of some observable on the unmodified PES, which can simply be obtained by back-correcting the observable average on the biased energy surface.^{99,107,109} In our MTD based approach the targets are the “true” quantum chemical energy minima, as defined in Section 3.1. Therefore, the latter requires a separate geometry optimization of the generated MTD structure snapshots.

For the automated generation of conformers we developed a composite approach consisting of MTD sampling, regular MD sampling and a procedure that is related to genetic structure crossing algorithms (GC).^{11,110,111} Hence, the procedure was termed iMTD-GC, where the lowercase *i* indicates an iterative strategy within the algorithm. As mentioned above, the approach heavily relies on the semiempirical GFN n -xTB methods ($n = 0–2$),^{37–39} which offer the possibility for fast and robust calculations. The general workflow is outlined in Fig. 5.

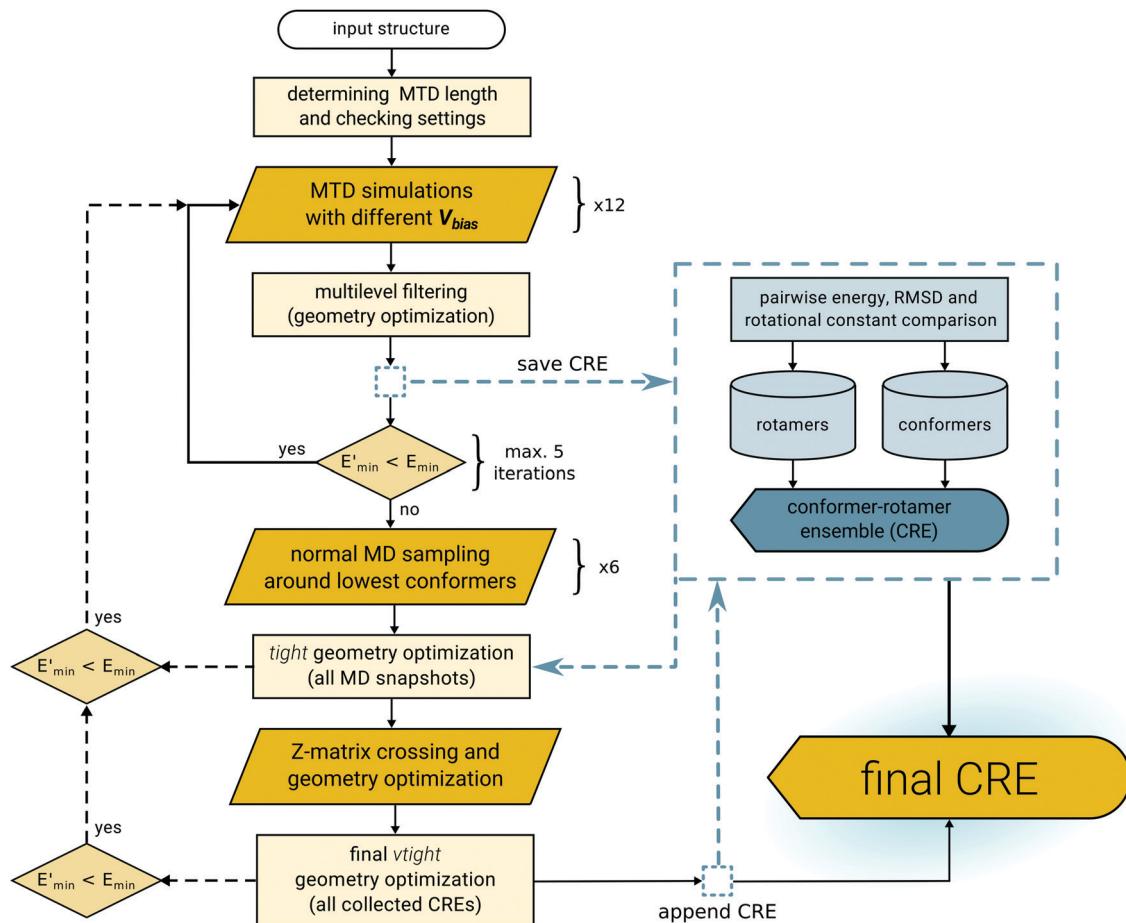


Fig. 5 Outline of the iMTD-GC workflow.

First, the maximum MTD length is determined, which mainly depends on the molecular size and flexibility of the system. Then, technical general settings are evaluated, to check if the MTDs will run stable. The main step of each iMTD-GC conformational search is the MTD sampling. Since the automatization is the key step and different molecules require adjusted pairs of k_i and α_i to produce the best results, a set of twelve MTDs is performed with different settings for the V_{bias} parameters. Here, α_i typically has values between 0.1 and 1.3 Bohr⁻², which can be seen as the “range” of the bias. The constant k_i is scaled by the number of atoms N , where k_i/N has magnitudes of 0.75 to 3.00 mE_h. Within the MTD simulation a new structure is added to the V_{bias} potential every 1.0 ps, which constantly drives the molecule into new conformations as time progresses. However, since the PES is constantly modified by the bias potential, the conformers form the MTD trajectory can not directly compared to each other and have to be re-optimized without the biasing potential. This is done in a two-step filtering procedure, first with very crude and then with tight convergence criteria. Afterwards the re-optimized structure snapshots from the trajectory are sorted according to the procedure outlined in Fig. 3, which yields an initial CRE. If a new conformer is found that is lower in energy than the input structure, the entire procedure is restarted on this conformer,

otherwise the workflow is continued. By default, the MTD iteration is restarted at least once, but not more than five times. All intermediate CREs are saved to be compared at a later stage. In the second step two unbiased MD simulations (*i.e.*, at two different temperatures 400 K and 500 K) are run on the three lowest conformers. This is done to get conformations with low-energy barrier crossings, opposed to high barrier conformational changes that can be obtained by the MTD simulations. The low-energy barrier crossings include simple torsional motions, such as group rotations, which are needed to complete the CRE regarding the rotamers. All structures from the MD simulations are sorted again and included in the intermediate CRE. In the final step the genetic structure crossing (GC) is performed with automatically generated Z-matrices, as was described in previous publications.^{44,74} Together with the regular MD simulations this approach helps to further complete the CRE and is particularly useful for flexible systems, *e.g.*, with many alkyl chains. If in the MD or GC step a new lowest energy conformer is found, the entire procedure is restarted. However, unlike the MTD iterations, these iterations do not have a maximum number of cycles and will only terminate if the lowest conformer does not change any more. The advantage of these restarts is mainly observed for larger molecules whose global minimum structure can be way off the

initial input geometry. All collected CREs are then optimized once more with very tight energy convergence criteria and the final CRE is created. Various energy thresholds and other MTD settings are employed within the workflow, which will be discussed in detail below.

3.3 Conformations at low-cost QM level

Conformations are generated at the GFN n -xTB level within the iMTD-GC workflow, as implemented in the CREST program. For the reliable generation of conformers at a semiempirical level there are two main questions that have to be answered: first, how trustworthy are SQM conformations (ΔE and geometries) compared to higher level theoretical methods such as density functional theory (DFT)? And secondly, can experimentally observed low-energy conformations be reproduced at a low-cost level of theory?

Concerning the first question a huge amount of literature exists in which various theoretical methods are benchmarked on conformational energies and geometries. It seems to be consensus in the literature that geometries are often quite well reproduced by SQM methods.^{112,113} This is particularly true for methods of the GFN n -xTB family, which, as their name conveys, are parameterized to yield reasonable structures.^{37,38,113} However, the calculation of accurate conformational energies at a semiempirical level remains difficult since small energy differences have to be described quite accurately. In Fig. 6 the mean average deviations (MADs) for some SQM methods are shown as evaluated on subsets of the GMTKN55 database¹¹⁴ that investigate conformational energies and the MALT205 set containing the energies for 205 conformers of maltose.¹¹⁵ The deviation of conformational energies from highly accurate WFT reference values is very dependent on the type of molecular system. *E.g.*, for simple alkane isomers in the ACONF set the MADs of all depicted semiempirical methods are well below 1 kcal mol⁻¹, while for the conformers of maltose (MALT205 set) all methods show deviations >3 kcal mol⁻¹. Likewise, different semiempirical methods do not describe all systems

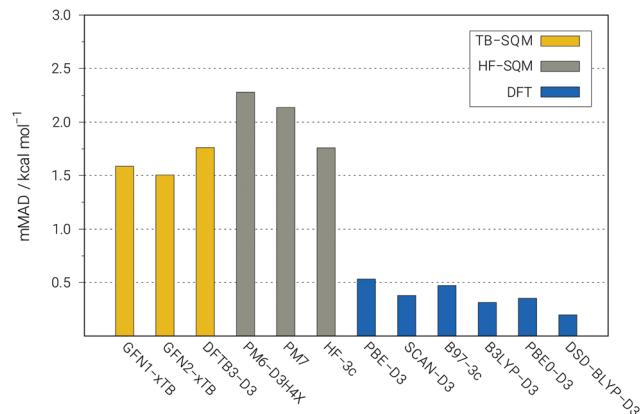


Fig. 7 Averaged MADs for the subsets shown in Fig. 6, evaluated for different levels of theory.

equally well. For example the Hartree–Fock derived method PM6-D3H4X^{116–118} has a MAD for the SCONF set that is more than twice the MAD of the tight-binding based methods, while for the melatonin conformers in MCONF it is the best performing semiempirical method. In Fig. 7 the mean average deviations (MADs) from Fig. 6 are averaged (mMAD) for different levels of theory, which provides a more general overview for the average performance of conformational energies. Their accurate description requires a balanced description of covalent and noncovalent energy contributions. As Fig. 7 depicts, this balance in general is much better at the DFT level than at any semiempirical level. The MAD of conformational energies calculated by semiempirical methods is on average more than three times higher than at even a “cheap” DFT level. At the DFT level, the PES appear to be much smoother and consistently shaped. The same observation was also made in other publications.¹¹⁹ Nevertheless semiempirical methods allow much shorter computation times while still maintaining a reasonable level of accuracy, which is sufficient for qualitative results. Several recent studies also show that the GFN n -xTB methods are among the best performing semiempirical methods for conformational energies and geometries.^{82,120} The good trade-off between accuracy and computational cost enables the use of GFN n -xTB for the generation of conformers with the iMTD-GC workflow. Although it is technically possible, any higher level theoretical method, even low cost DFT with a small basis set, would be much too expensive for the vast amount of required geometry optimizations and evaluations.

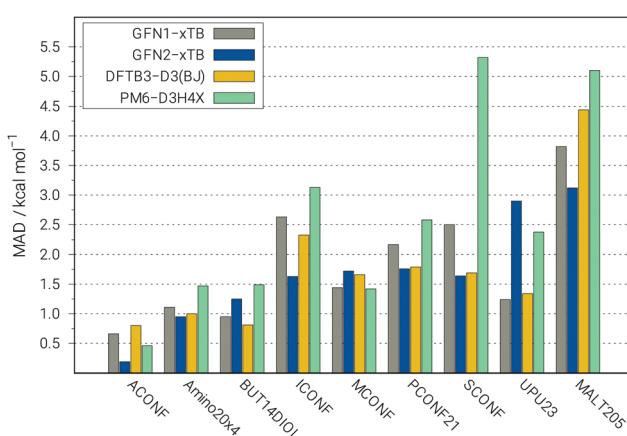


Fig. 6 MADs for conformational energies evaluated for GFN1-xTB, GFN2-xTB, DFTB3-D3, and PM6-D3H4X on the ACONF, Amino20x4, BUT14DIOL, ICONF, MCONF, PCONF21, SCONF, and UPU23 subsets of the GMTKN55 database and the MALT205 benchmark.

4 Selection of default thresholds and settings in CREST

Any (semi-)automated screening procedure requires the application of pre-defined thresholds of various kinds. The by far most important threshold is the energy window, *i.e.*, the maximum relative energy up to which structures are considered further. This is naturally related to the population of the

structures at finite temperature, which is calculated according to eqn (2). For conformational energies at a semiempirical level the default size of this window is 6 kcal mol⁻¹ in CREST, which is a reasonable but still conservative choice for many systems.^{74,120} This window is applied even though significant Boltzmann populations at 298 K are obtained only up to approximately 2 kcal mol⁻¹. The larger “save choice” value should account for the non-parallel PES of the semiempirical and higher level QM methods. In fact, the results from benchmark studies suggest (*cf.* Fig. 6), that in some cases it may be necessary to increase the energy window to 10 kcal mol⁻¹ or more. For applications that involve chemical changes such as the protonation, even larger energy windows have to be applied in order to recover all thermodynamically accessible structures. By default, the corresponding energy threshold for the protonation, deprotonation and tautomerization applications is 30 kcal mol⁻¹. Other thresholds are applied for the identification of conformers and rotamers as discussed in Section 3.1. These thresholds are used to quantify the difference between two structures according their relative energies (E_{thr}), atomic RMSD (RMSD_{thr}), and rotational constants (B_{thr}). The default values are given in Table 1. The selection of default threshold values has a significant influence on the performance and results of an automated procedure, particularly on the computational wall-times. Hence, it is important to carefully choose and adjust these settings. If for example a smaller energy window is chosen, more structures will be discarded, which leads to shorter computation times but also to less complete ensembles. Or, as another example, if the RMSD threshold is increased, more structures will be (falsely) identified as the same conformer, also leading to smaller ensembles and consequentially lower computational cost. Other settings, such as the simulation time of the MTD simulations, have a more direct influence on the performance of the workflow and are system dependent. Due to this dependence such settings have to be determined dynamically for each simulation. In case of the MTD time t_{mtd} (in picoseconds) the scaling is chosen to be dependent on an effective number of atoms N_{eff} ,

$$t_{\text{mtd}} = 0.1(N_{\text{eff}} + 0.1N_{\text{eff}}^2), \quad \text{where } 5 \text{ ps} \leq t_{\text{mtd}} \leq 200 \text{ ps.} \quad (4)$$

This is justified because larger molecules typically require longer simulations in order to undergo conformational changes. The effective atom number is obtained from the total atom number N and a flexibility measure according to

Table 1 Overview of various default thresholds applied in the CREST program for structural comparisons

Threshold	Value
E_{win} energy (conformers)	6.0 kcal mol ⁻¹
E_{win} energy (prot./deprot./taut.)	30.0 kcal mol ⁻¹
E_{thr} energy (between conformers)	0.1 kcal mol ⁻¹
RMSD _{thr}	0.125 Å
B_{thr} (rot. constant)	15.0 MHz

$N_{\text{eff}} = N\mu_f$, where μ_f is given by

$$\mu_f = \sqrt{\frac{1}{N_{\text{bonds}}} \left(\sum_i^{N_{\text{bonds}}} \left(1 - e^{-5(B_{AB}-2)^{10}} \right)^2 \frac{4}{N_A^{\text{neigh}} N_B^{\text{neigh}}} \left(R_i^{(\text{f})} \right)^2 \right)^{\frac{1}{2}}}. \quad (5)$$

Here, the summation runs over all non-terminal bonds i with the involved atoms A and B (*i.e.*, A, B $\in i$), B_{AB} is the Wiberg–Mayer bond order^{121,122} between the two atoms as obtained from a GFNn-xTB calculation, and $N_{A,B}^{\text{neigh}}$ are the numbers of neighboring atoms of A and B. The predefined factor $R_i^{(\text{f})}$ is 1 if the bond i is not part of a ring and <1 (depending on the ring size) if it is. In total, the flexibility measure μ_f can be $0 < \mu_f < 1$, where values close to 1 indicate a highly flexible system, *e.g.* *n*-alkanes, and values $\ll 1$ indicate rigid systems. For other proposed flexibility measures, see *e.g.* ref. 123 and 124.

Another important system dependent variable is the bias potential V_{bias} (see eqn (3)). Since the RMSD is a normalized variable, potentials of the same order of magnitude would be obtained for any system. However, larger molecules do require higher bias to undergo conformational changes. Therefore, the potentials V_{bias} have to be scaled by the system size, which in practice is done by scaling the pre-factors in eqn (3) by the number of atoms, *i.e.*, $k_i = k'_i N_{\text{at}}$. Finding and optimizing a reasonable set of pairs for the variables k_i and α_i is a non-trivial task and was done by hand.⁷⁴ Different combinations of the two factors will act differently upon any system and hence a set of all combinations of three different k'_i and four different α_i (yielding in total 12 MTD simulations, see Fig. 5) is applied. Additionally, two further simulations are performed with extreme value combinations. For specialized applications such as the NCI-iMTD (for noncovalently bound complexes, see Section 7.2) different sets of parameters are employed. The default combinations of k'_i and α_i are listed in Table 2.

The influence of a single bias potential on a two dimensional PES of 1-bromo-3-chloropropane, obtained by rotation around two dihedral angles, is shown in Fig. 8. For this small molecule the entire conformational PES is described by the two dihedral angles φ and θ . As can be seen from Fig. 8b the surface is symmetric with two global minima (θ, φ) at (68°, 66°) and (292°, 294°), which correspond to the different enantiomers of 1-bromo-3-chloropropane. Since two enantiomers are distinguishable by their atomic RMSD it is possible to “fill” one of the global minima with a bias potential while retaining the other. As a consistency check, by choosing the enantiomer (292°, 294°) (Fig. 8a) as the center for V_{bias} , the biased PES in Fig. 8c is obtained, which has only the global minimum at (68°, 66°). Furthermore, in this example a combination of k_i and α_i similar to setting 10 of Table 2 was employed, which provides valuable insight about their effect on the PES.

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

Table 2 Combinations of the parameters k'_i and α_i as applied by default in the CREST program. k'_i is multiplied by the number of atoms N_{at} in order to obtain k_i of eqn (3)

MTD	k'_i [mE _h]	α_i [Bohr ⁻²]
1	3.00	1.300
2	1.50	1.300
3	0.75	1.300
4	3.00	0.780
5	1.50	0.780
6	0.75	0.780
7	3.00	0.468
8	1.50	0.468
9	0.75	0.468
10	3.00	0.281
11	1.50	0.281
12	0.75	0.281
13	1.00	0.100
14	5.00	0.800

5 Computational details

All shown screening procedures are implemented in a computer program called CREST which makes use of the xtb program. CREST makes full use of single node (OMP) parallelization in order to execute several independent xtb calculations at once. All calculations executed with xtb were performed using the 6.2 release version of the program. DFT calculations were performed with the TURBOMOLE 7.3.1 program. The resolution-of-identity (RI) approximation for the Coulomb integrals¹²⁵ was generally applied using the matching default auxiliary basis sets.¹²⁶ The integration of the exchange–correlation contribution was evaluated on the numerical quadrature grids m4. The default convergence criteria for single-point energies were 10^{-7} E_h.

6 Conformational search examples

6.1 Conformations of (*S*)-citronellal

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

The (*S*)-stereoisomer and its conformers were investigated. Because of the acyclic geometry and with five freely rotatable C–C bonds it is to be expected that the conformational space of citronellal is relatively large. Fig. 10b shows a part of the multidimensional unbiased PES of citronellal along two dihedral angles. The two dimensional PES shows many deep pocketed minima connected by high energetic barriers of approximately 5–6 kcal mol⁻¹. In Fig. 10c, the 2D-PES in a MTD run (with bias potential) is shown, which reduces the complexity of the PES in terms of accessible conformations considerably.

The gas-phase iMTD-GC calculation generated 262 conformers, within the conservatively chosen energy window of 6 kcal mol⁻¹.

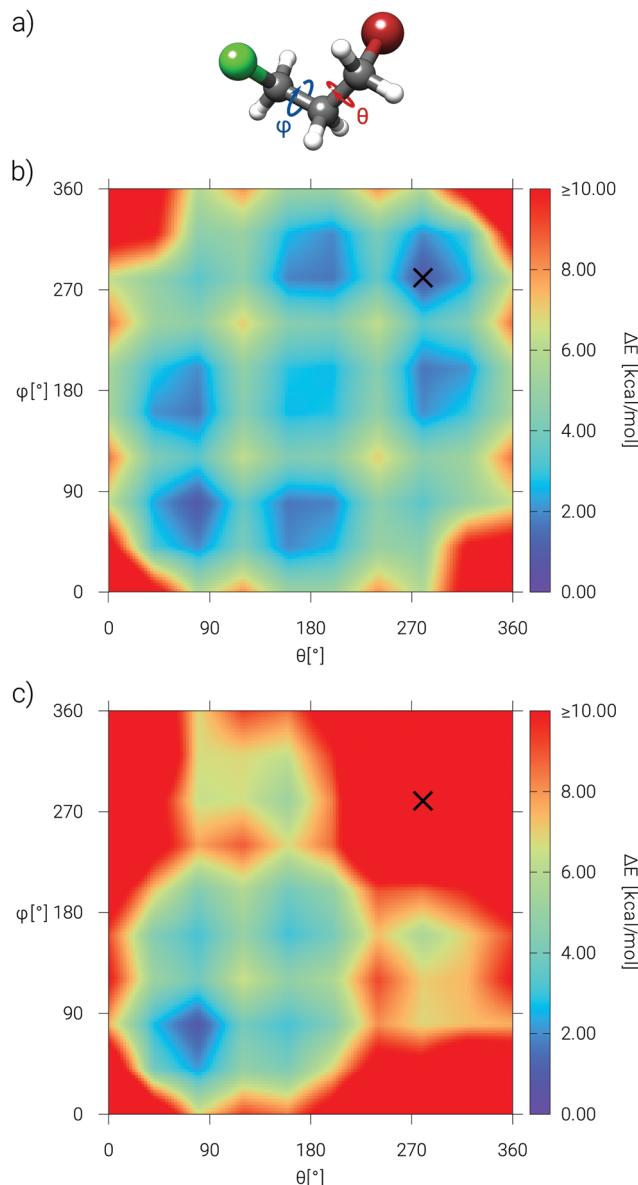


Fig. 8 Comparison of the unbiased and biased two dimensional PES of 1-bromo-3-chloropropane at the GFN2-xTB level. (a) Molecular structure and the dihedral angles φ and θ used for the construction of the two dimensional PES of 1-bromo-3-chloropropane. (b) Unbiased PES at the GFN2-xTB level. (c) PES under influence of a single bias potential V_{bias} with $k_i = 0.03$ E_h and $\alpha_i = 0.3$ Bohr⁻² at the GFN2-xTB level. The center of V_{bias} is marked by "X", which corresponds to the global minimum structure shown in (a).

As already mentioned, large energy windows allow the compensation of non-parallel energy surfaces of SQM and DFT (see Fig. 7) thereby preventing the loss of potentially low lying conformers at DFT level after reoptimization. The conformers can be classified into two main types: chain-like and globular-folded conformations. By visual inspection of the CRE it becomes apparent that the relative orientation of the aldehyde functional group has a large influence on the conformation of citronellal and the globular-folded conformation is stabilized by intramolecular noncovalent interactions of the aldehyde and C–H groups. The conformational

flexibility of citronellal has been experimentally investigated by the group of M. Schnell using chirped-pulse Fourier-transform microwave (CP-FTMW) spectroscopy where 15 gas-phase conformers were identified.¹²⁸ Rotational constants contain information about the entire geometry of the molecule and therefore conformational aspects as well. Since rotational constants are rather method dependant, the CRE is reoptimized at the PBEh-3c (DFT) level to get a good comparison to the experiment. Conformers with small mean deviations of the rotational constants between theory and experiment were visually inspected for matching geometries in the literature.¹²⁸ All 15 experimentally identified conformers are found in the iMTD-GC-CRE. The first five of the 15 conformers are shown in Fig. 9a.

Post-optimization of the ensemble at a higher level is important as also shown by the fact that the 30th conformer of the initial conformer search at GFN2-xTB level corresponds to the lowest lying conformer of the refinement at PBEh-3c level, highlighting a noticeable reranking. The conformers shown in Fig. 9c are identified by iMTD-GC//GFN2-xTB to be the highest populated and were also predicted by the computational investigation of Schnell *et al.* Interestingly, the conformers CONF1 and CONF2 could not be identified within the experimental spectrum. Experiments with different carrier gases suggested that conformational relaxation towards more stable conformers is facilitated due to collision in the supersonic expansion. The absence of CONF1 and CONF2 can be explained by conformational relaxation and their too high enthalpy after refinement at DFT level.

After demonstrating that relevant conformers were found within the gas-phase CRE, the completeness of the ensemble concerning the populated conformers is investigated by comparing

calculated and experimental ¹H-NMR spectra in solution. To this end, a new CRE was generated in chloroform and the ¹H-NMR spectrum was calculated (see Fig. 11) with the procedure detailed in ref. 44. The spectrum calculated with only one conformer clearly shows that the multiplet splittings of the proton signals are not correct. Only if the whole populated ensemble of 35 conformers is taken into account, the qualitatively correct splittings are obtained. The differences between the experimental and calculated chemical shifts partly stem from the neglect of zero-point effects, vibrational averaging, and errors in the respective DFT calculations. It has to be stressed that in addition to the conformers the identification of rotamers is crucial for calculating NMR spectra. Rotamers are very important since they are necessary to describe the correct averaging of NMR parameters due to the fast interchange of nuclei at the slow time scale of the NMR experiment. The overall good agreement between the calculated and experimental multiplicities indicates that all major conformers were found. This in turn highlights the sophistication and robustness of the iMTD-GC algorithm for generating conformer-rotamer-ensembles (CRE).

6.2 Conformations of macrocyclic molecules

The treatment of macrocyclic rings with chemoinformatic conformational sampling algorithms is challenging since it requires special treatments or heuristic rules.⁹ Molecular dynamic-based approaches prove to be useful as their application is straightforward and no additional adjustments are required for these compounds.^{21,90,134} In particular, aMD based approaches seem to be promising as the PES modification helps to overcome even high energetic ring-interconversion barriers.¹⁰⁸ Therefore, it can also be expected for the MTD approach to yield similarly good results. To assess the performance

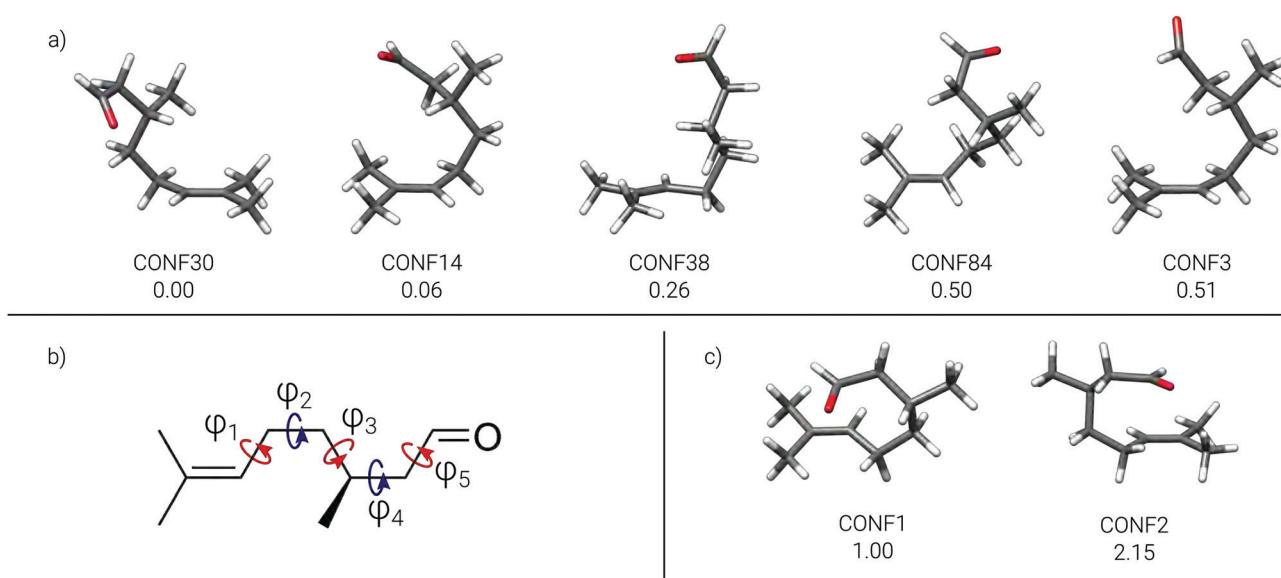


Fig. 9 Citronellal gas-phase conformers depicted with calculated enthalpies at 0 K in kcal mol⁻¹. (a) 15 conformers were experimentally identified by microwave spectroscopy and all 15 conformers were found in the calculated CRE. Only the five highest populated conformers are shown. Enthalpies were calculated at 0 K as the sum of PBEh-3c¹²⁷ energies and zero point vibrational energies from GFN2-xTB, (b) Lewis structure of (S)-citronellal with unhindered non-methyl-dihedral angle rotations highlighted, (c) these conformers obtained initially at xTB level are too high in energy at DFT level and not found under experimental conditions.

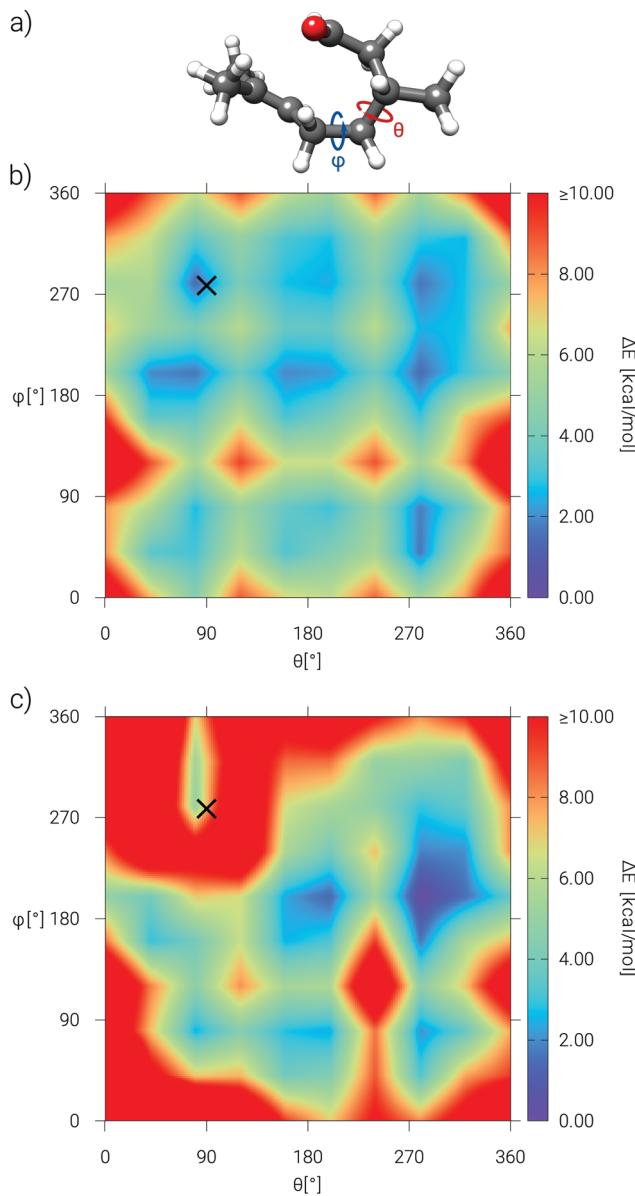


Fig. 10 Comparison of the unbiased and biased two dimensional PES of citronellal at the GFN2-xTB level. (a) Citronellal with indicated dihedral angles φ and θ for the construction of the two dimensional PES. (b) Unbiased PES of citronellal. (c) PES with a single bias potential V_{bias} ($k_i = 0.06 E_h$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

of iMTD-GC for macrocycles, a conformational search was conducted for three different macrocyclic systems taken from ref. 90. Crystal structure geometries of the three compounds were obtained from the Cambridge Crystal Structure Database (CCSD)¹³⁵ with the IDs POXTRD, CAMVES, and CHPSAR. The geometries were optimized in the gas-phase with GFN2-xTB. Their composition and size represent typical organic systems in the target range of the iMTD-GC. The (gas-phase) conformer ensembles can be expected to be sufficiently diverse if: (A) a high

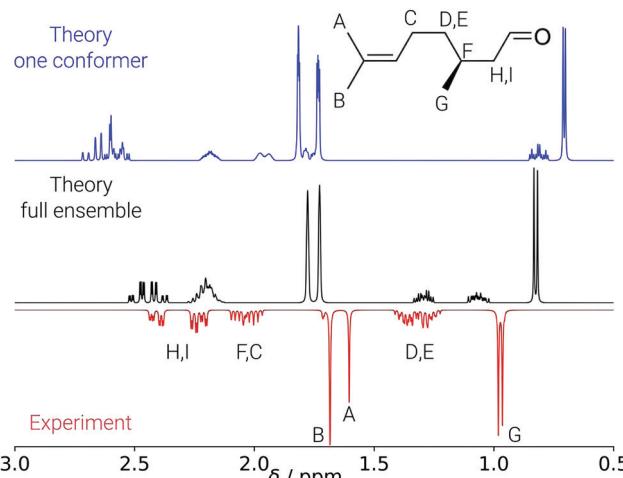


Fig. 11 Comparison of calculated and experimental ^1H -NMR spectra of citronellal with the focus on the aliphatic region. The upper theoretical spectrum (in blue) is calculated with only one conformer. The theoretical spectrum (in black) better reproduces the experimental spectrum and is calculated with the full CRE consisting of 35 conformers. The lower experimental spectrum¹²⁹ (red, inverted) was measured in CHCl_3 at 400 MHz, 298 K and the theoretical spectra were calculated with PBE0/def2-TZVP[COSMO]^{130–132}/PBEh-3c[DCOSMO-RS]¹³³ for the coupling and shielding constants. The assignment of the multiplets is indicated by capital letters.

RMSD to the input (crystal structure) conformation is observed, (B) a large number of conformers within the default energy window is obtained, and (C) there is a large energetic difference between the input structure and the lowest conformer. Results according to these criteria are depicted in Fig. 12.

For all three systems the defined criteria for a diverse ensemble are fulfilled. The smallest macrocycle POXTRD has a huge number of distinct conformers, while the other two examples (CHPSAR, CAMVES) have smaller CREs. By visual inspection of the ensembles of CHPSAR and CAMVES, pairs of hydrogen bonds are identified within the ring-systems. The hydrogen bonding strongly stabilizes a few selected conformational motifs, leading to more compact ensembles compared to POXTRD within the energy window of 6 kcal mol^{-1} . The results show that ring-interconversion poses no problem for the iMTD-GC procedure and highlights that different molecular classes can be treated with the same set of search parameters.

6.3 Conformations of Ac-Ala₁₉-LysH⁺

As a larger example we chose the protonated peptide Ac-Ala₁₉-LysH⁺, consisting of 20 amino acids (220 atoms). With the size and flexibility of this system we approach the current practical limit of the iMTD-GC workflow conducted at a SQM level, although the application to larger systems would easily be possible at a FF level. From combined theoretical and experimental studies it is known that the conformation of this molecule depends on the protonation site.^{136–138} It was found that the protonation at the C-terminus stabilizes an α -helical form of the peptide, which is shown in Fig. 13. The alternative N-terminal protonation destabilizes the helical conformation and

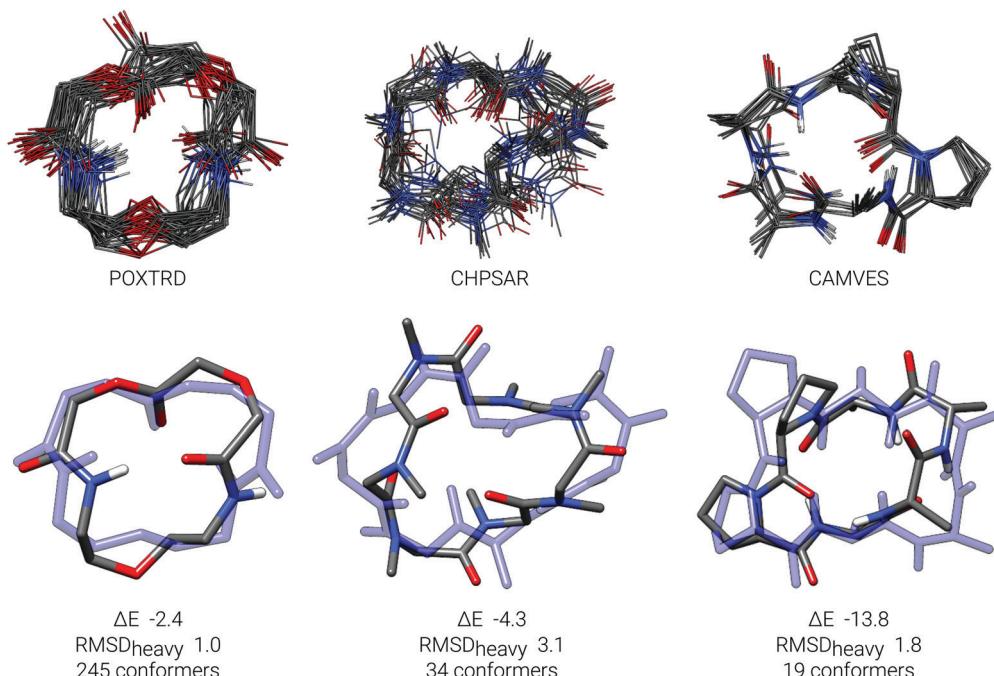


Fig. 12 iMTD-GC gas-phase ensembles of three macrocycles are shown. Below overlays between GFN2-xTB optimized geometries from the crystal structure and the lowest lying conformer are presented. For POXTRD only the first 45 conformers of the ensemble are shown. Energy differences (ΔE) of the lowest lying conformer and optimized input geometry are given in kcal mol^{-1} . The RMSD_{heavy} between the lowest lying conformer and input geometry is given in Å.

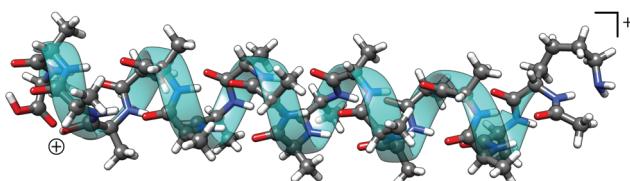


Fig. 13 The most stable conformer of Ac-Ala₁₉-LysH⁺ with protonation at the C-terminus. The protonation site is marked by the “+”-sign.

leads to more compact structures. Furthermore, the lysine protonated conformations are preferred in the gas-phase, where a single unique conformation was suggested.¹³⁸ Hence, the Ac-Ala₁₉-LysH⁺ system is an ideal molecule to evaluate the performance of iMTD-GC for larger systems. Three conformational searches were conducted: starting (A) from the α -helical conformer protonated at the C-terminus, (B) from a N-protonated helical structure, and (C) from the lowest-energy conformation proposed in ref. 138.

Starting from the C-terminal protonated conformations (case A), the findings in the literature can be confirmed already at the GFN2-xTB level. The entire ensemble consists of 54 α -helical structures that differ only by the orientation of Lys. The most stable conformation is shown in Fig. 13. No energetically close lying folded structures were found on the PES where the protonation site is at the C-terminus. During the iMTD-GC search several of these structures are created and then sorted out according to the energy threshold.

In the second conformational search (case B) the most stable helical conformation is taken as an input structure and the protonation

site is artificially set to the N-terminal lysine group. With this setup only folded conformations are generated in the ensemble (126 conformers). Here, the most stable conformers are very similar to the “unique” conformation from the literature. For an energy based comparison, the “unique” structure identified as lowest conformer in ref. 138 is taken as the reference point.

A comparison between the most stable conformation of the iMTD-GC ensemble and the reference structure is shown in Fig. 14. The folded conformers contain α - and 3_{10} -helical segments. However, at the GFN2-xTB level the most stable conformers generated by iMTD-GC are up to $1.64 \text{ kcal mol}^{-1}$ lower in energy than the reference structure.

When the conformational search is started from the reference structure (case C), a smaller ensemble compared to the previous searches is obtained by iMTD-GC (56 molecules). The energetically lowest conformers of ensemble B and C have similar structures and are up to $1.7 \text{ kcal mol}^{-1}$ more stable than the reference. For better comparison between the iMTD-GC structures and the reference all conformations were re-optimized at the PBEh-3c level and single-point energies were calculated at the higher PBE0-D4/def2-TZVPD level.¹³⁹ The relative energies for selected conformers are plotted in Fig. 15. Although the obtained conformers from the second conformational search are very similar to the reference structure and are favored by a few kcal mol^{-1} at the GFN2-xTB level, at the hybrid DFT level the reference structure is still preferred. However, by conducting the third conformational sampling (case C), the literature ensemble could be extended. At the PBE0-D4/def2-TZVPD level, there are three conformers in the new ensemble which are energetically lower

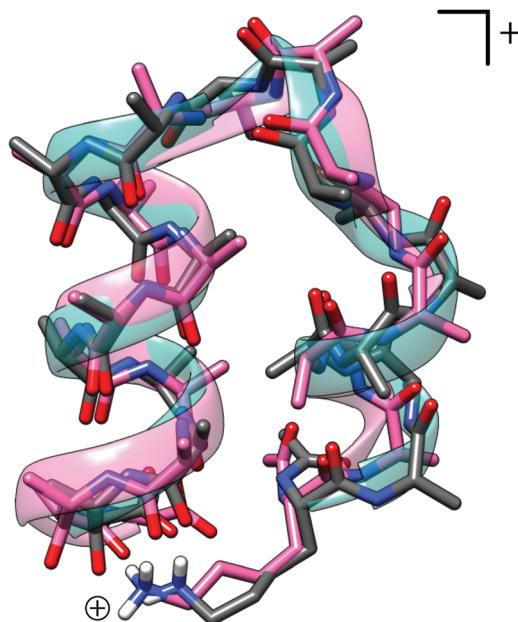


Fig. 14 Overlay between the most stable conformer of $\text{Ac-Ala}_{19}\text{-LysH}^+$ taken from literature (magenta) and the highest populated conformer of ensemble B generated by iMTD-GC (gray/green). The N-terminal protonation site is marked by the “ $+$ ”-sign.

(by 0.33 to 0.37 kcal mol^{-1}) than the reference. Furthermore, there are four other conformers which are less than 0.1 kcal mol^{-1} above the reference and thus would be significantly populated. Compared to the PBE0-D4 results, most conformations at the GFN2-xTB level are energetically close lying. This indicates a much flatter PES at the SQM level compared with the DFT level. Since energies of up to 0.4 kcal mol^{-1} are well within the error margin of conformational energies at a hybrid DFT level, it is not possible to

determine if these new conformers are truly the most stable structures of $\text{Ac-Ala}_{19}\text{-LysH}^+$. However, the ensemble indicates that there is not just one unique populated conformer of the peptide in the gas-phase but rather an ensemble of several energetically and structurally close conformations that are protonated at the lysine terminus. Furthermore, the example shows that qualitative results can be obtained with the iMTD-GC//GFN2-xTB method at comparatively low computational cost. In the original study several first-principles simulations based on replica-exchange molecular dynamics (REMD)^{140–143} at the DFT (PBE) level¹⁴⁴ were performed in order to find the low energy conformations.¹³⁸ This approach is extremely expensive, even without including PBE0 single-point calculations. In contrast, the iMTD-GC//GFN2-xTB sampling was conducted within a couple of days on a single workstation and yielded similar low-energy conformations.

6.4 Conformers of metal-organic systems

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

The first example is *trans*- $\text{Cu}^{II}(\text{l-valine})_2$. Each chelate ring can have an axial or equatorial conformation of the isopropyl-group. The valine residue can exhibit various conformations, identifiable by comparing the highlighted methine-proton position in Fig. 16, where six representative conformers are depicted. The generated structures were refined at the B97-3c

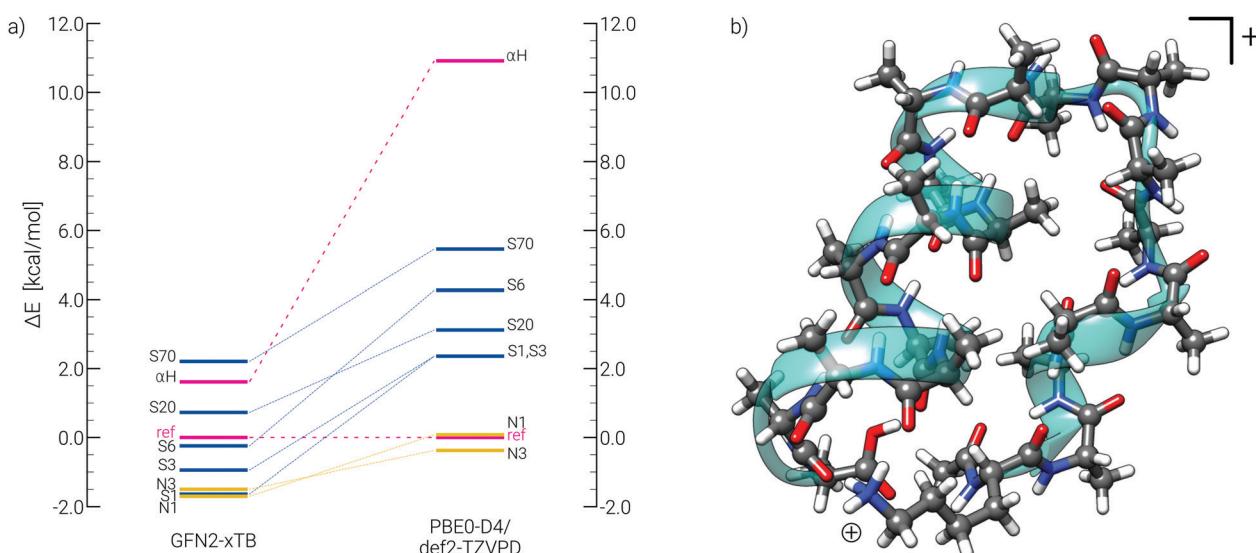


Fig. 15 (a) Comparison of relative energies for selected conformers at the GFN2-xTB and PBE0-D4/def2-TZVPD//PBEh-3c levels. Conformers labeled by “S” were generated starting from the N-protonated helical conformer (case B). Conformers labeled by “N” were generated starting from the reference structure “ref” (case C). αH is the most stable α -helical conformer. (b) Overall most stable iMTD-GC conformer at DFT level of $\text{Ac-Ala}_{19}\text{-LysH}^+$, labeled N3.

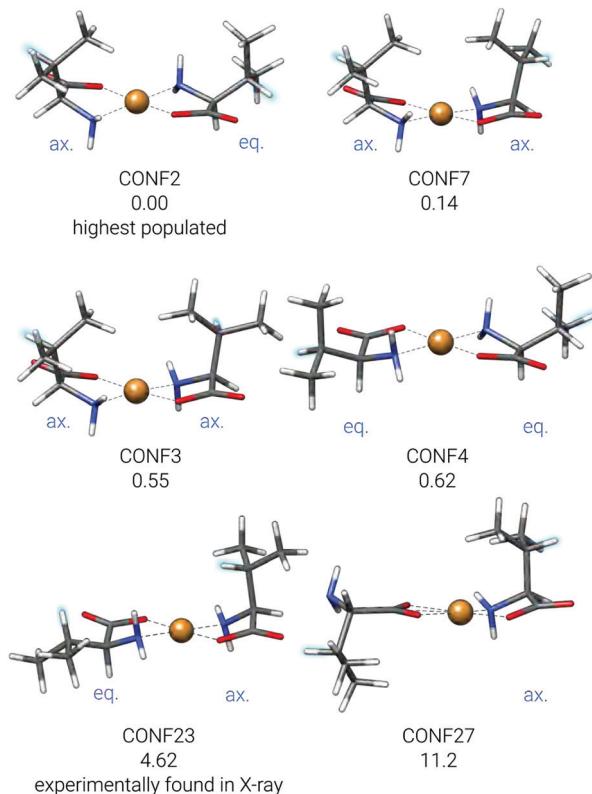


Fig. 16 Conformers of *trans*-Cu^{II}(L-valine)₂ optimized in gas-phase at B97-3c¹⁴⁵ level of theory. The free energies given below the respective conformer name are in kcal mol⁻¹ (@298.15 K).

level, where the first four conformers are populated and conformers CONF23 and CONF27 are higher lying conformations. Conformer CONF23 is not populated in the gas-phase but closely resembles the experimentally observed conformation in the crystal structure.¹⁴⁶ This indicates its stabilization in the solid phase due to packing (or other related) effects.

As a second example, the κ^2 -tris(pyridyl)methanol-diacetyl-platinum(II) complex was investigated. The chelate complex was studied in implicit methanol and the ensemble consists of 68 conformers within an energy window of 10 kcal mol⁻¹. The tris(pyridyl)methanol ligand forms a κ^2 -chelate complex with

the Pt(II)-ion and has a free coordination site at the third uncoordinated pyridyl group. In Fig. 17, a selection of conformers is shown, highlighting the flexibility and dynamic processes that may occur in solution. The acetyl groups (COMe) are almost freely rotatable and the uncoordinated pyridyl ring easily can rotate as well. The numbering at the coordinating pyridyl-nitrogens, illustrates that the chelate complex can open up and re-coordinate to the free center. This is particularly visible in conformer 59, where the hydroxy group coordinates to the Pt-ion and the pyridyl-group is facing away from the metal center. The conformer search gives valuable insight into the flexibility of this Pt(II)-chelate-complex, which has been confirmed experimentally.¹⁴⁷ The wall computation time for this calculation was 12 minutes on 40 cores (Intel Xeon Gold 6148 CPU @ 2.4 GHz).

7 Specialized applications

The general setup of the conformational search algorithm as a combination of molecular dynamics simulations and quantum chemical structure optimization enables several specialized applications. Atoms can be constrained in the input structure or removed from the bias potential. Furthermore, other potentials than V_{bias} can be included in the calculations.

7.1 Constrained conformational sampling

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

7.1.1 Tyrosine conformation on a graphene surface. Biosensing of α -amino acids using nanomaterials is a vital research for which a detailed knowledge of the amino acid conformations at the material-interface is essential.^{148,149} The conformational search of L-tyrosine at a model graphene surface is demonstrated here. The graphene sheet consists of 216 carbon atoms and has D_{6h} symmetry. For the conformational search with CREST, the

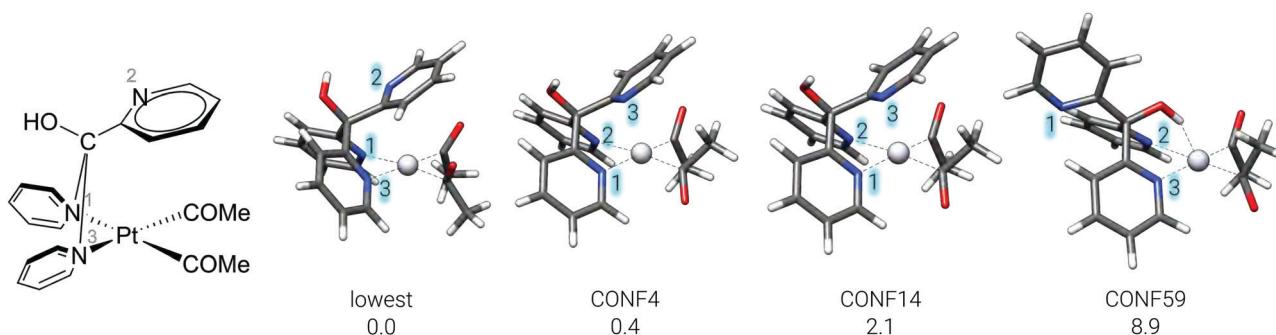


Fig. 17 Lewis structure of [Pt(COMe)₂(2-py)₃COH] and selected lowest and higher lying conformers calculated at GFN2-xTB[GBSA(MeOH)] (GBSA = Generalized Born (GB) with solvent accessible surface area (SA) level of theory (ΔE in kcal mol⁻¹) are shown. The numbering at the nitrogen atoms is aiding in the distinction of the coordination centers.

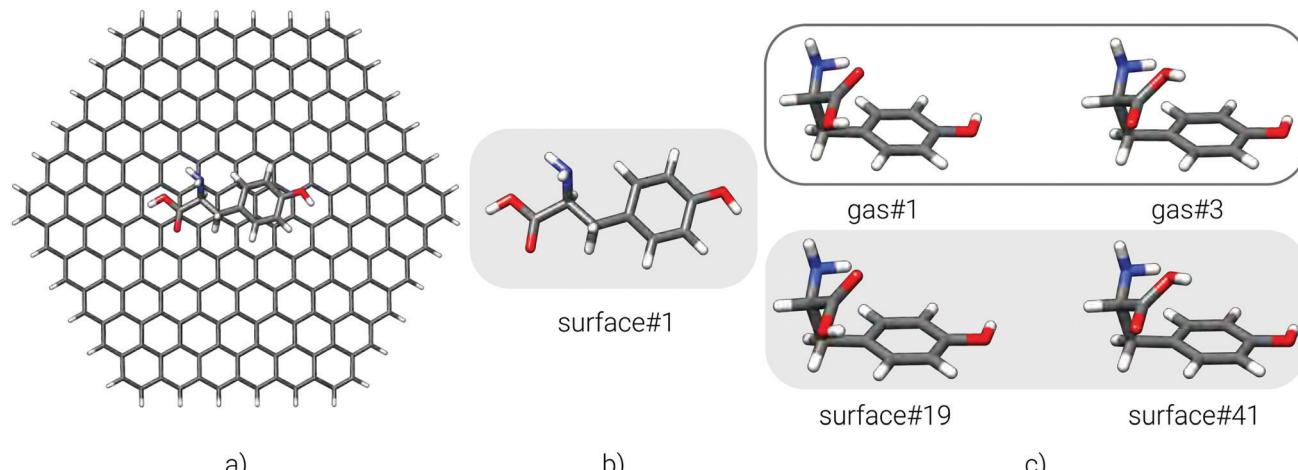


Fig. 18 Tyrosine conformations on a model graphene surface and in the gas-phase. (a) Most stable tyrosine conformer on the graphene surface. (b) Most stable tyrosine conformer depicted without graphene. (c) Comparison between gas-phase and interface conformers.

graphene layer is constrained and all graphene atoms are removed from the RMSD criterion. Otherwise, this would lead to the dissociation of the complex and a strong deformation of the graphene monolayer. Fig. 18a shows the lowest lying L-tyrosine conformation found at the graphene surface.

On the surface 108, different tyrosine conformers were found. All low lying conformers show the $\pi\cdots\pi$ interaction motif via parallel alignment of the phenyl-ring to the graph sheet. Higher-energy conformations bend away from the parallel phenyl-ring arrangement. Other observed interaction motifs are C(α/β)-H $\cdots\pi$ and N-H/O-H $\cdots\pi$ interactions. The interaction energy (E_{int}) for the most stable conformer and the graphene mono-layer amounts to $-2.8 \text{ kcal mol}^{-1}$ at the GFN2-xTB level. To highlight the difference between a constrained conformational search at an interface and in the gas-phase, tyrosine conformers were created in the gas-phase without any constraints. Here, 30 conformers were found. Comparing the geometries taken from the graphene-based ensemble and the gas-phase it is seen, that conformations are created at the graphene interface, which are not present in the gas-phase ensemble.

As depicted in Fig. 18b and c some low lying gas-phase conformations can be found on the graphene surface, but they are not the lowest populated conformers, e.g., the lowest gas-phase tyrosine conformer is equal to the 19th conformer from the graphene-tyrosine ensemble and is $1.3 \text{ kcal mol}^{-1}$ higher in energy than the lowest graphene-tyrosine conformer. The graphene potential clearly influences the tyrosine conformations and hence it is essential to create conformations in their genuine environment.

7.1.2 Conformers of transition-states. Studying reaction kinetics is usually done by investigating activation energies of rate determining steps. Reactions are analyzed with the assumption that the reaction kinetics are termed by the transition-state (TS) free energy, relative to the free energy of the reactants. It is assumed that the reaction proceeds through the TS, which is linked to the energetically most favorable substrate. If the substrate can interconvert quickly between

its low energy conformations, the reaction is governed by the Curtin–Hammett principle.¹⁵⁰ In this case, the reaction proceeds through the lowest TS which is not necessarily connected to the lowest lying substrate conformer. Hence, for accurate results it may be necessary to search the chemical reaction space and find low-lying TS conformations. Constraint conformer searches can be applied to a previously found TS. The procedure is demonstrated for the enzyme COMT (catechol-O-methyl transferase), which catalyzes the methyl group transfer from S-adenosyl-L-methionine (SAM) to a catechol ion^{151,152} (see Fig. 19a for the Lewis structure of the TS). First, the basic S_N2 reaction was modeled by taking the active site of the enzyme and saturating all capped bonds with methyl groups. A TS guess was optimized and checked for the reaction mode by performing a harmonic frequency calculation. To preserve the TS vibrational mode in the conformational search, the atoms dominantly contributing to this mode were constrained. In this system, the catechol oxygen, the carbon of the transferred methyl group, and the sulfur of the SAM group were fixed. To retain the magnesium-cation coordination the Mg-ligand distances were constrained as well (only one constraint per ligand). Additionally, the water and the amid ligand were constrained to an O_{H₂O}-Mg-O_{Amide} angle of 180°. For the TS conformational search only the constraining of the breaking and forming bonds is necessary and all other constraints are used to keep the active site of the enzyme intact. The iMTD-GC approach generates 141 conformers within 6 kcal mol⁻¹. Overall, the calculation takes 46 minutes on 40 cores (Intel Xeon Gold 6148 CPU @ 2.4 GHz). The conformers are good estimates for the further optimization into the TS at GFN2-xTB[GBSA(MeOH)] level. Here, 138 true TS are obtained. During the optimization of the TS, geometries can converge into the same TS geometry and have to be sorted out. After sorting, 91 unique conformers within 6.1 kcal mol⁻¹ remain. Overlays of the optimized TS ensemble and the lowest lying TS are depicted in Fig. 19b and c. The procedure provides a semiautomated approach of finding lower lying TS by relaxing the ligand geometries.

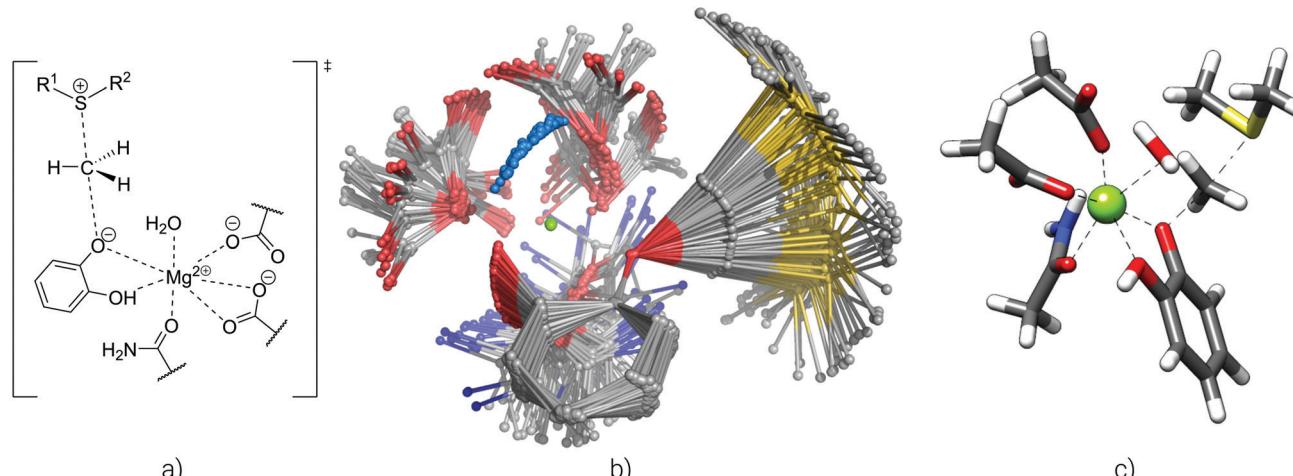


Fig. 19 TS of the active site of the COMT enzyme. (a) Lewis structure of the S_N2 methyl-group transfer reaction, (b) TS ensemble of 91 TS which are optimized after the iMTD-GC conformer search (atoms in light blue depict the water oxygen, all hydrogen atoms are omitted for clarity), (c) lowest lying TS at GFN2-xTB[GBSA(MeOH)] level.

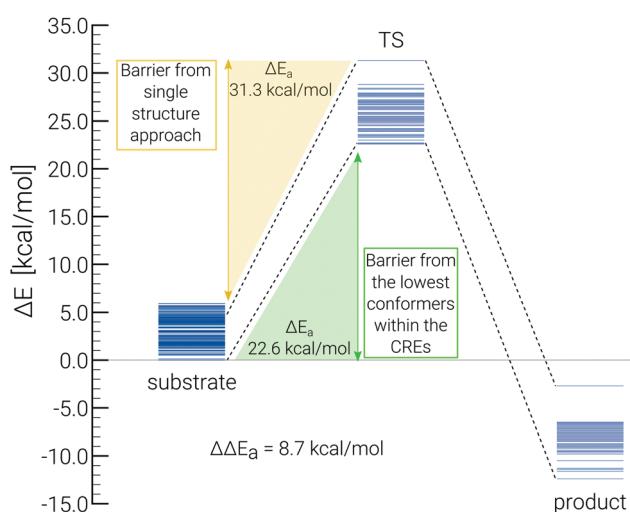


Fig. 20 Reaction barriers of the S_N2 methyl-group transfer. The reaction barrier from the single structure approach (yellow) and barrier from the lowest conformers of the CREs (green) are compared. Energies of the conformers relative to the lowest substrate conformation are illustrated by horizontal blue lines. All energies were calculated at the GFN2-xTB[GBSA(MeOH)] level.

For the COMT example the lowest-lying TS conformer is $13.5 \text{ kcal mol}^{-1}$ lower in energy than the initial TS. Comparing the barrier from the lowest substrate conformer and the lowest TS-conformation with the reaction barrier from the single structure approach (see Fig. 20) the barrier height is reduced by $8.7 \text{ kcal mol}^{-1}$ when using the ensemble based protocol. Although the work was conducted at a semiempirical level it should be stressed that the procedure is generally applicable and that it is also possible to refine the TS at DFT-level.

7.2 Aggregate sampling

With the MTD based approach it is possible to also screen for different conformations of noncovalently bound aggregates

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

$$E_{\text{pot}} = \sum_i^N \left(\frac{|\mathbf{R}_i - \mathbf{O}|}{R_{i,\text{pot}}} \right)^{10}, \quad (6)$$

where the summation runs over all atoms N . \mathbf{R}_i are the Cartesian coordinates of atom i , \mathbf{O} is the center of the potential (*i.e.*, the origin), and $R_{i,\text{pot}}$ is the radius of the potential parallel to $\mathbf{R}_i - \mathbf{O}$. If such a potential is combined with a strong RMSD bias, chemical reactions can be enforced, resulting in a mode similar to the nano-reactor presented in ref. 74. However, the aim of the NCI-iMTD procedure is to generate a low-energy ensemble for which the parameters are adjusted accordingly. In the following two examples demonstrate the application of the NCI-iMTD procedure.

7.2.1 Water hexamer. Since each noncovalently interacting fragment can also have different conformations on its own even for small systems a large number of complexes is possible. In the general case one would search first with iMTD-GC the monomer conformations before NCI-iMTD is started. Hence, as a primary example small molecular clusters such as $(H_2O)_6$ are used here, where fragment conformations are irrelevant. In the first step the ellipsoidal potential is automatically generated from scaled principal rotation axes. For the water hexamer the resulting potential is schematically shown in Fig. 21. Within a 6 kcal mol^{-1} energy window 69 different gas-phase aggregates are found for $(H_2O)_6$. Many of the generated structures are isomers that differ only by the direction of their hydrogen bonds and otherwise show a similar geometry. Hence, only a selection of six noteworthy structures is shown in Fig. 22. The structures shown

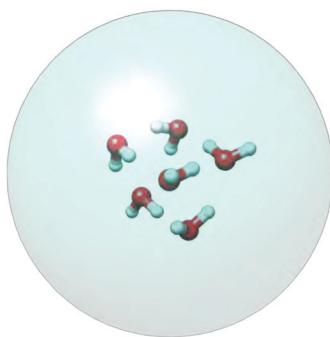


Fig. 21 Visualization of the almost spherical ellipsoidal potential around a $(\text{H}_2\text{O})_6$ cluster.

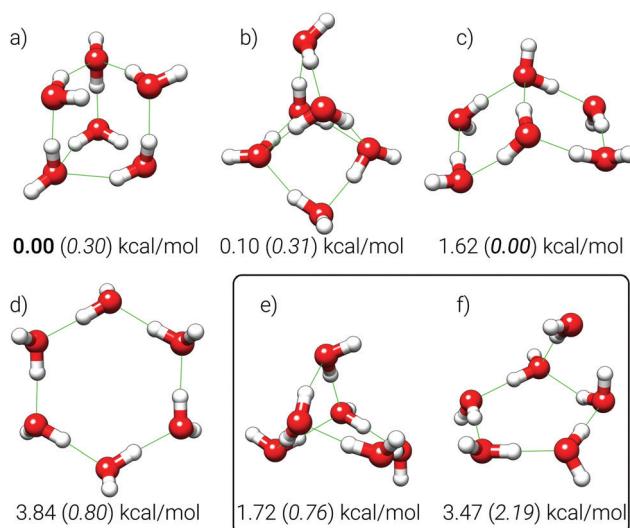


Fig. 22 Six different aggregates that are automatically generated for $(\text{H}_2\text{O})_6$. Relative energies (ΔE) at the GFN2-xTB level are given below the respective structures. Energy values in parenthesis show relative energies at the PBEh-3c level.

in Fig. 22a–d are well-known from literature (*i.e.*, the prism-, cage-, book-, and cyclic-hexamer respectively) and are for example included in the WATER27 subset of the GMTKN55 database.^{114,153} The two aggregates in Fig. 22e and f, which were chosen randomly from the ensemble, additionally highlight the structural diversity created by the NCI-iMTD procedure.

7.2.2 1-Naphthol dimer. The second example is the 1-naphthol dimer, where various π - π stacking and hydrogen bonding motifs are possible. In the literature seven conformations were proposed based on a comparison of the experimental and theoretical rotational constants.¹⁵⁴ With the default NCI-iMTD mode, 88 unique aggregates of the 1-naphthol dimer are found at the GFN2-xTB level of theory within a 6 kcal mol⁻¹ energy window. To figure out whether the seven proposed complexes are included in this ensemble, structures were pre-selected based on low mean relative deviations (MRD) of the calculated GFN2-xTB and experimental rotational constants from ref. 154. Afterwards, these structures were visually inspected. The reference rotational constants were obtained for geometries calculated at

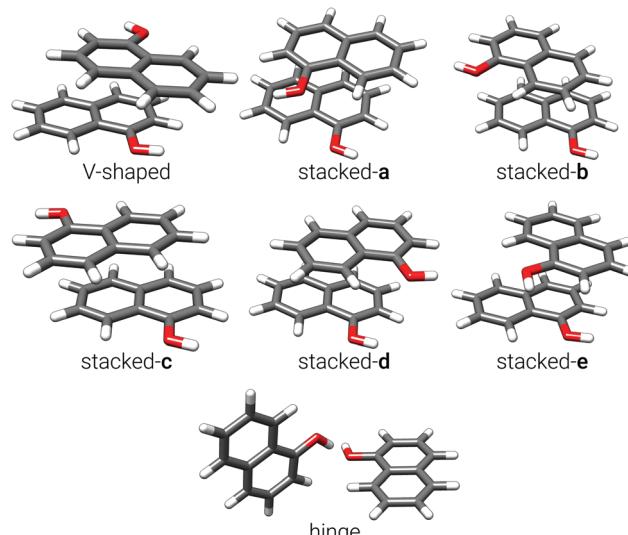


Fig. 23 Seven 1-naphthol dimer aggregates that were generated at the GFN2-xTB level of theory and match the structures proposed in the literature.

B3LYP-D3(BJ)/6-31+g(d,p) level of theory taken from the original publication (see ref. 154). For all seven complexes a corresponding (or at least closely related) structure was found with the NCI-iMTD mode at the GFN2-xTB level of theory. The geometries are shown in Fig. 23. From benchmark studies it is known that rotational constants for geometries calculated at a semi-empirical level have larger relative deviations compared with geometries calculated by DFT.¹⁵⁵ Therefore, the GFN2-xTB geometries are not expected to reproduce the hybrid DFT rotational constants very well but rather serve as a guideline for the structure evaluation. The rotational constants for the seven predicted complexes of the 1-naphthol dimer are given in Table 3. The rotational constants calculated for the GFN2-xTB geometries deviate from the reference by 4–5% on average. Many more aggregates are generated, and none of the seven isomers is the most energetically favored complex of the ensemble at the semiempirical level. However, the trend in the aggregate stability identified by Jäger *et al.*,¹⁵⁴ that the V-shaped forms are more stable compared to the stacked

Table 3 Rotational constants B_e for the seven predicted aggregates shown in Fig. 23

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

^a Calculated for GFN2-xTB geometries. ^b Taken from ref. 154.

Table 4 Rotational constants B_e for the four new predicted 1-naphthol aggregates shown in Fig. 24

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

^a Taken from ref. 154.

and hinge forms is reproduced for the GFN2-xTB ensemble. Comparing directly to the experimental rotational constants (see Table 4), a total of five structures from the NCI-iMTD ensemble fit to the experimental values. All five geometries have rotational constants matching the experiment but a clear identification based only on the very similar rotational constants is not possible. Four of these structures are V-shaped homologues. In fact, the most stable of them is the V-shaped complex that was identified by the authors ref. 154 and is shown in Fig. 23. Only one of the structures is not V-shaped but appears to be a more symmetric form of the stacked **b**-complex. The predicted complexes (except the already known V-shaped form) are shown in Fig. 24. The new generated aggregates support the original conclusion that the true conformation of the 1-naphthol dimer is a V-shaped type complex.

7.3 Automated protonation/cationization

The procedure for the automated protonation was already discussed in ref. 29 and is only briefly outlined here. The computational protocol has since been optimized and was implemented into CREST. The workflow is similar to the concept used within the chemical reaction networks,^{62,64} where

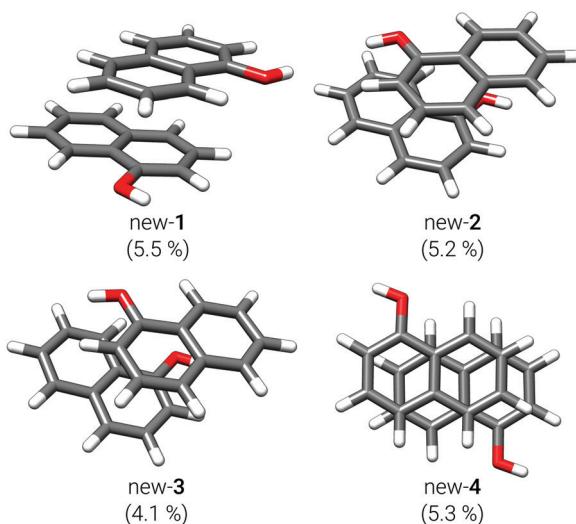


Fig. 24 Four new aggregates are predicted for the 1-naphthol dimer generated at the GFN2-xTB level of theory. The MRD of the calculated rotational constant from the respective experimental value is given in parenthesis.

the reactive sites are determined first and then a reactive species is placed in close vicinity during a geometry optimization. Here, reactive sites are π - and lone-pair (LP)-centers that are obtained from localized molecular orbitals (LMOs) and the reactive species is a proton (H^+). The geometry optimization leads to a set of different protomers. Sorting the protomer ensemble is based on relative proton affinities, in the same way that the CRE sorting depends on conformational energies. GFN *n*-xTB is able to describe these proton affinities sufficiently accurate.²⁹ The procedure is schematically outlined in Fig. 1 and in Fig. 25 for the benzocaine molecule, which is a prominent example for protomers in the literature.^{156–159} In the gas-phase, the O-protonated benzocaine molecule is favored over the N-protonated species (at GFN2-xTB level). All possible protonation sites in the aromatic ring are also obtained with the automated procedure but are still not populated. However, if the calculation is performed with implicit water solvation, both the N- and O-protonated species are obtained and populated. This corresponds to the experimental finding that N-protonation can (only) occur in the gas-phase under the influence of microsolvation.¹⁵⁹ In a modified version of the procedure also the cationization of molecules, for example with a sodium ion Na^+ , is possible. The setup is the same as for the protonation site screening, but the corresponding ion is placed at the π - or LP-center, instead of a proton. An example from the literature¹⁶⁰ is the cationization of adenosine to $[\text{Ade} + \text{Na}]^+$. At the GFN2-xTB level ten different cationized structures are generated within a conservative 30 kcal mol^{−1} energy window. The two energetically lowest structures are found within 10 kcal mol^{−1} and agree with the predicted relative stabilities in the literature.¹⁶⁰ Also reasonable higher-energetic coordination motives are generated by the automatized screening procedure. Gas-phase structures of the two lowest and one higher energetic $[\text{Ade} + \text{Na}]^+$ complex are given in Fig. 26.

The preferred coordination sites found, *i.e.*, 1-[Ade + Na]⁺ and 2-[Ade + Na]⁺ in Fig. 26, are correct and were already discussed in the literature.¹⁶⁰ By default the screening procedure only yields topologically unique structures without specific attention being paid to the conformation and hence may not lead to the global minimum. Small conformational differences can, *e.g.*, result from a different conformation of the input structure

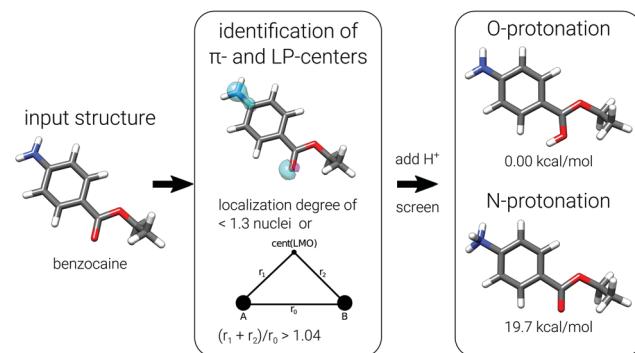


Fig. 25 Schematic procedure of an automatized protonation of the benzocaine molecule.

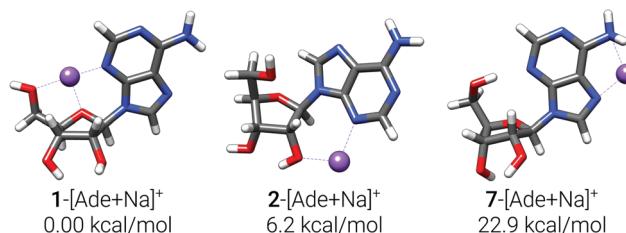


Fig. 26 Three of overall ten $[Ade + Na]^+$ structures generated at the GFN2-xTB level within a 30 kcal mol^{-1} energy window. Relative energies at GFN2-xTB level are given below the corresponding structure.

(here adenine). Additional conformational searches for the products were not performed but in real applications should be conducted.

The trivial counterpart to the automated protonation is the automated deprotonation. Since the screening only involves the removal of protons to generate input deprotonomers and the results are comparable to the automated protonation, it will not be discussed here. Overall the protonation/cationization procedure provides an automated approach of finding relevant protomers and coordination sites which can be used in further computational studies.

7.4 Automatized tautomerization and isomerization

Tautomerism is a widespread phenomenon that influences the chemistry of molecules with readily interchangeable isomers. In the most common type of this inter-conversion the isomers only differ in the position of a proton, which is called prototropy.¹⁵⁰ Knowledge about a molecules possible tautomeric behavior is of great importance, since the isomers can strongly differ in their physical and chemical properties.^{35,161} It is also highly important in structural databases and hence, many chemoinformatic approaches exist for the identification of tautomers.^{36,162,163} The automated quantum chemical tautomerization (or isomerization) is a standalone feature implemented in CREST. The protocol involves a sequence of protonating and deprotonating steps and was first applied in ref. 43 for the calculation of pK_a values in water. If larger topological changes are induced by this procedure, the obtained structures have to be considered structural isomers instead of prototropic tautomers. A starting geometry is protonated as described in Section 7.3 and each resulting protomer is deprotonated at every position. With this procedure all stable structures containing a single permutation of proton positions with respect to the input are obtained. Typically, these “first order” prototropic tautomers are already a good estimate of a molecules’ tautomerism. However, even for simple molecules such as guanine shown in Fig. 27, further permutation of protonation sites can lead to additional tautomers. Hence, the screening procedure is an iterative sequence of protonation and deprotonation which is performed twice or more. Fig. 27 shows automatically generated low-energy tautomers of guanine. The tautomerization was started from the keto form of guanine, dubbed **1-gua** in Fig. 27. In the first protonation/deprotonation iteration, the tautomers **2-gua**, **3-gua** and **4-gua** are obtained, which differ only by the position of a single

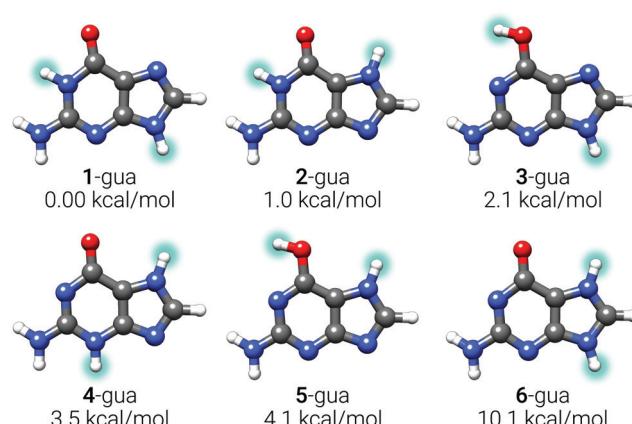


Fig. 27 Six lowest prototropic tautomers of guanine generated at the GFN2-xTB[GBSA(H_2O)] level within a 30 kcal mol^{-1} energy window. Relative energies at the GFN2-xTB[GBSA(H_2O)] level are given below the corresponding structure. Proton positions that are changed during the tautomerization are highlighted.

hydrogen atom (relative to **1-gua**). Further iterations also yield the structures **4-gua** and **5-gua** with two permuted hydrogen positions. Typically, two iterations of protonation and deprotonation are sufficient to recover relevant low-energy tautomers. In case of the guanine ensemble at the GFN2-xTB[GBSA(H_2O)] level all experimentally known low-energy tautomers (Fig. 27) as well as higher-energetic structures discussed in the literature were recovered.^{164–168}

As already mentioned the QM based protocol enables its application also to metal containing molecules. In terms of tautomerism this includes the typical prototropic case (*e.g.*, at the ligands), as well as less common phenomena such as proton-hydride tautomerism. An example for the latter is the tautomerism of $[(Cp^*)Rh(bpy)H]^+$ as shown in Fig. 28, which is part of an experimentally suggested catalytic cycle.¹⁶⁹ The GBSA implicit solvation model for acetonitrile was employed to resemble the experimental conditions. Within a 10 kcal mol^{-1} window there are three distinct proton-hydride tautomers of $[(Cp^*)Rh(bpy)H]^+$ at the GFN2-xTB[GBSA(MeCN)] level (some broken structures were discarded). Rh-**a** is taken as input, where the proton is bound to the Cp^* ligand, facing downwards

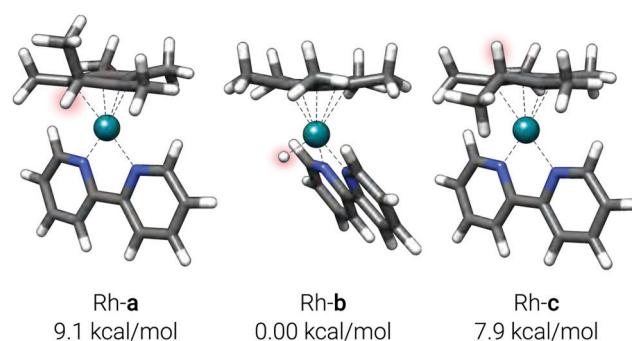


Fig. 28 Selected proton-hydride tautomers of $[(Cp^*)Rh(bpy)H]^+$ obtained automatically by using Rh-**a** as an input. Relative energies at the GFN2-xTB[GBSA(MeCN)] level are given below the structures.

to the rhodium atom. The procedure automatically recovers the hydride structure Rh-**[(Cp*)RhH(bpy)]⁺**, which is in fact the most stable tautomer and another prototropic tautomer Rh-**c**, in which also the Cp* ligand is protonated, but the proton is facing away from the metal center. While Rh-**c** is approximately 1.2 kcal mol⁻¹ more stable than Rh-**a**, only the latter is a reactive species in the catalytic cycle.¹⁶⁹ However, since there are reactive centers (LMOs) also on the up-facing side of Cp*, the procedure generates Rh-**c**.

In general, the automated sequence of protonation/deprotonation allows to access any tautomer that converges to a local minimum during the geometry optimization. The reason for this is that no barriers for the addition or secession of protons are taken into account and hence thermodynamically unfavorable (*i.e.*, high energetic) isomers can be produced. Therefore, the procedure recovers typically not only prototropic tautomers but also other structural isomers. Due to the exploratory nature of the approach it is currently up to the user to decide which of the generated structures shall be considered as tautomers and which as structural isomers. The exploration has, however, the advantage that more complicated types of tautomerism, such as ring-chain-tautomers, are often found by the same workflow. In contrast, many chemoinformatic tautomerization tools require additional heuristics for the treatment of ring-chain-tautomerism.^{35,36}

8 Troubleshooting

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

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

If deviations occur in “fair” comparisons, *i.e.*, the simulated conditions match the experimental ones, their cause is mostly

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

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

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

9 Conclusion

We presented a variety of automated quantum chemical screening procedures for the efficient exploration of the low-energy chemical space. The main focus herein is the generation and separation of the different isomers that are referred to as conformers. Different conformations contribute to various physical observables, such as NMR shifts and coupling constants, reaction barriers or pK_a values. Therefore, the knowledge about a molecules’ ensemble of different conformations is a valuable information required for accurate computational modeling. We have shown a computational workflow for the generation of conformers based on a metadynamics (MTD) approach with a self-similarity energy penalty V_{bias} that utilizes the atomic Cartesian RMSD as a collective variable. Furthermore, the procedure includes a genetic structure crossing (GC) step and was implemented in an iterative algorithm, which is conveyed in the abbreviation iMTD-GC. The induced directionality of the chemical space exploration due to the RMSD bias allows for shorter simulation times compared with conventional MD based sampling approaches. Savings in the computational cost are in turn invested for the reoptimization of geometries at a low-cost QM level. The iterative ansatz hereby helps to explore the conformational space with respect to the

global minimum structure, as it only terminates if no lower energy conformers are found. The algorithm is general in the sense that it will work on every level of theory for the underlying PES. However, the meta-dynamics sampling shows its potential especially in combination with SQM methods that make the evaluation of thousands of molecular geometries feasible. For this purpose, the robust and reliable GFN2-xTB method was employed. Even at a relatively cheap level of quantum chemistry structures are generated that often match quite well with the experimentally observed conformations. The performance was assessed for several systems with up to 220 atoms in direct comparison with either experiment or high level theoretical data. The investigated systems include typical organic compounds, organo-metallic complexes, and noncovalently bound clusters. The conformational search algorithm can also be modified by applying additional constraints, *e.g.*, fixing of different bond lengths. This constrained conformational sampling can be a valuable tool for obtaining better TS geometries and barriers.

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

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

Program availability

The xtb and crest programs (Linux binary only) can be obtained free of charge from GitHub (<https://github.com/grimme-lab/xtb/releases>). Instructions for both programs can be found under <https://xtb-docs.readthedocs.io/en/latest/contents.html> (last accessed 10, December, 2019). Instead of an ESI, the examples are shown in the online documentation.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

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

References

- S. Grimme and P. R. Schreiner, *Angew. Chem., Int. Ed.*, 2017, **57**, 4170–4176.
- K. N. Houk and F. Liu, *Acc. Chem. Res.*, 2017, **50**, 539–543.
- F. R. Salsbury, *Curr. Opin. Pharmacol.*, 2010, **10**, 738–744.
- J. Xu and A. Hagler, *Molecules*, 2002, **7**, 566–600.
- A. K. Ghose, T. Herbertz, J. M. Salvino and J. P. Mallamo, *Drug Discovery Today*, 2006, **11**, 1107–1114.
- Y.-C. Lo, S. E. Rensi, W. Tornig and R. B. Altman, *Drug Discovery Today*, 2018, **23**, 1538–1546.
- A. Lavecchia, *Drug Discovery Today*, 2015, **20**, 318–331.
- B. Kuhn, W. Guba, J. Hert, D. Banner, C. Bissantz, S. Ceccarelli, W. Haap, M. Körner, A. Kuglstatter, C. Lerner, P. Mattei, W. Neidhart, E. Pinard, M. G. Rudolph, T. Schulz-Gasch, T. Woltering and M. Stahl, *J. Med. Chem.*, 2016, **59**, 4087–4102.
- P. C. D. Hawkins, *J. Chem. Inf. Model.*, 2017, **57**, 1747–1756.
- P. C. D. Hawkins, A. G. Skillman, G. L. Warren, B. A. Ellingson and M. T. Stahl, *J. Chem. Inf. Model.*, 2010, **50**, 572–584.
- M. J. Vainio and M. S. Johnson, *J. Chem. Inf. Model.*, 2007, **47**, 2462–2474.
- S. Riniker and G. A. Landrum, *J. Chem. Inf. Model.*, 2015, **55**, 2562–2574.
- M. A. Miteva, F. Guyon and P. Tufféry, *Nucleic Acids Res.*, 2010, **38**, W622–W627.
- K. Tai, *Biophys. Chem.*, 2004, **107**, 213–220.
- R. J. Dorfman, K. M. Smith, B. B. Masek and R. D. Clark, *J. Comput.-Aided Mol. Des.*, 2008, **22**, 681–691.
- N. Sauton, D. Lagorce, B. O. Villoutreix and M. A. Miteva, *BMC Bioinf.*, 2008, **9**, 184.
- K. S. Watts, P. Dalal, R. B. Murphy, W. Sherman, R. A. Friesner and J. C. Shelley, *J. Chem. Inf. Model.*, 2010, **50**, 534–546.
- I. Kolossaváry and W. C. Guida, *J. Am. Chem. Soc.*, 1996, **118**, 5011–5019.
- N.-O. Friedrich, F. Flachsenberg, A. Meyder, K. Sommer, J. Kirchmair and M. Rarey, *J. Chem. Inf. Model.*, 2019, **59**, 731–742.
- P. Bonnet, D. K. Agrafiotis, F. Zhu and E. Martin, *J. Chem. Inf. Model.*, 2009, **49**, 2242–2259.
- K. S. Watts, P. Dalal, A. J. Tebben, D. L. Cheney and J. C. Shelley, *J. Chem. Inf. Model.*, 2014, **54**, 2680–2696.
- E. A. Coutsias, K. W. Lexa, M. J. Wester, S. N. Pollock and M. P. Jacobson, *J. Chem. Theory Comput.*, 2016, **12**, 4674–4687.
- A. E. Cleves and A. N. Jain, *J. Comput.-Aided Mol. Des.*, 2017, **31**, 419–439.

- 24 F. Fuster and B. Silvi, *Chem. Phys.*, 2000, **252**, 279–287.
- 25 J. C. Shelley, A. Cholleti, L. L. Frye, J. R. Greenwood, M. R. Timlin and M. Uchimaya, *J. Comput.-Aided Mol. Des.*, 2007, **21**, 681–691.
- 26 C. Liao and M. C. Nicklaus, *J. Chem. Inf. Model.*, 2009, **49**, 2801–2812.
- 27 J. F. Sanz, J. Anguiano and J. Vilarrasa, *J. Comput. Chem.*, 1988, **9**, 784–789.
- 28 J. El Yazal, F. G. Prendergast, D. E. Shaw and Y.-P. Pang, *J. Am. Chem. Soc.*, 2000, **122**, 11411–11415.
- 29 P. Pracht, C. A. Bauer and S. Grimme, *J. Comput. Chem.*, 2017, **38**, 2618–2631.
- 30 M. Namazian and H. Heidary, *THEOCHEM*, 2003, **620**, 257–263.
- 31 J. C. Kromann, F. Larsen, H. Moustafa and J. H. Jensen, *PeerJ*, 2016, **4**, e2335.
- 32 J. J. Klícić, R. A. Friesner, S.-Y. Liu and W. C. Guida, *J. Phys. Chem. A*, 2002, **106**, 1327–1335.
- 33 G. Cruciani, F. Milletti, L. Storchi, G. Sforna and L. Goracci, *Chem. Biodiversity*, 2009, **6**, 1812–1821.
- 34 A. Klamt, F. Eckert, M. Diedenhofen and M. E. Beck, *J. Phys. Chem. A*, 2003, **107**, 9380–9386.
- 35 Y. C. Martin, *J. Comput.-Aided Mol. Des.*, 2009, **23**, 693.
- 36 M. Sitzmann, W.-D. Ihlenfeldt and M. C. Nicklaus, *J. Comput.-Aided Mol. Des.*, 2010, **24**, 521–551.
- 37 S. Grimme, C. Bannwarth and P. Shushkov, *J. Chem. Theory Comput.*, 2017, **13**, 1989–2009.
- 38 C. Bannwarth, S. Ehlert and S. Grimme, *J. Chem. Theory Comput.*, 2019, **15**, 1652–1671.
- 39 P. Pracht, E. Caldeweyher, S. Ehlert and S. Grimme, *ChemRxiv*, 2019, DOI: 10.26434/chemrxiv.8326202.v1.
- 40 V. Asgeirsson, C. A. Bauer and S. Grimme, *Chem. Sci.*, 2017, **8**, 4879–4895.
- 41 S. Grimme, *Angew. Chem., Int. Ed.*, 2013, **52**, 6306–6312.
- 42 C. A. Bauer and S. Grimme, *J. Phys. Chem. A*, 2016, **120**, 3755–3766.
- 43 P. Pracht, R. Wilcken, A. Udvarhelyi, S. Rodde and S. Grimme, *J. Comput.-Aided Mol. Des.*, 2018, **32**, 1139–1149.
- 44 S. Grimme, C. Bannwarth, S. Dohm, A. Hansen, J. Pisarek, P. Pracht, J. Seibert and F. Neese, *Angew. Chem., Int. Ed.*, 2017, **56**, 14763–14769.
- 45 S. Maeda and K. Morokuma, *J. Chem. Phys.*, 2010, **132**, 241102.
- 46 S. Maeda, Y. Harabuchi, M. Takagi, T. Taketsugu and K. Morokuma, *Chem. Rec.*, 2016, **16**, 2232–2248.
- 47 M. P. Haag, A. C. Vaucher, M. Bosson, S. Redon and M. Reiher, *Chem. Phys. Chem.*, 2014, **15**, 3301–3319.
- 48 C. J. Stein and M. Reiher, *J. Comput. Chem.*, 2019, **40**, 2216–2226.
- 49 W.-J. V. Zeist, C. F. Guerra and F. M. Bickelhaupt, *J. Comput. Chem.*, 2008, **29**, 312–315.
- 50 L.-P. Wang, R. T. McGibbon, V. S. Pande and T. J. Martinez, *J. Chem. Theory Comput.*, 2016, **12**, 638–649.
- 51 D. Rappoport, C. J. Galvin, D. Y. Zubarev and A. Aspuru-Guzik, *J. Chem. Theory Comput.*, 2014, **10**, 897–907.
- 52 P. M. Zimmerman, *J. Comput. Chem.*, 2013, **34**, 1385–1392.
- 53 C. R. Jacob, S. M. Beyhan, R. E. Bulo, A. S. P. Gomes, A. W. Götz, K. Kiewisch, J. Sikkema and L. Visscher, *J. Comput. Chem.*, 2011, **32**, 2328–2338.
- 54 Y. Guan and S. E. Wheeler, *Angew. Chem., Int. Ed.*, 2017, **56**, 9101–9105.
- 55 F. Zapata, L. Ridder, J. Hidding, C. R. Jacob, I. Infante and L. Visscher, *J. Chem. Inf. Model.*, 2019, **59**, 3191–3197.
- 56 K. Ohno and S. Maeda, *J. Phys. Chem. A*, 2006, **110**, 8933–8941.
- 57 S. Maeda and K. Morokuma, *J. Chem. Theory Comput.*, 2011, **7**, 2335–2345.
- 58 S. Maeda, K. Ohno and K. Morokuma, *Phys. Chem. Chem. Phys.*, 2013, **15**, 3683–3701.
- 59 S. Maeda, T. Taketsugu and K. Morokuma, *J. Comput. Chem.*, 2014, **35**, 166–173.
- 60 S. Maeda, Y. Harabuchi, Y. Ono, T. Taketsugu and K. Morokuma, *Int. J. Quantum Chem.*, 2015, **115**, 258–269.
- 61 Y. Nagahata, S. Maeda, H. Teramoto, T. Horiyama, T. Taketsugu and T. Komatsuzaki, *J. Phys. Chem. B*, 2016, **120**, 1961–1971.
- 62 M. Bergeler, G. N. Simm, J. Proppe and M. Reiher, *J. Chem. Theory Comput.*, 2015, **11**, 5712–5722.
- 63 J. Proppe, T. Husch, G. N. Simm and M. Reiher, *Faraday Discuss.*, 2016, **195**, 497–520.
- 64 G. N. Simm, A. C. Vaucher and M. Reiher, *J. Phys. Chem. A*, 2018, **123**, 385–399.
- 65 G. N. Simm and M. Reiher, *J. Chem. Theory Comput.*, 2018, **14**, 5238–5248.
- 66 G. N. Simm, A. C. Vaucher and M. Reiher, *J. Phys. Chem. A*, 2019, **123**, 385–399.
- 67 J. Proppe and M. Reiher, *J. Chem. Theory Comput.*, 2019, **15**, 357–370.
- 68 H. B. Schlegel, *J. Comput. Chem.*, 2003, **24**, 1514–1527.
- 69 A. L. Dewyer and P. M. Zimmerman, *Org. Biomol. Chem.*, 2017, **15**, 501–504.
- 70 A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng and K. W. Jacobsen, *J. Phys.: Condens. Matter*, 2017, **29**, 273002.
- 71 G. te Velde, F. M. Bickelhaupt, E. J. Baerends, C. Fonseca Guerra, S. J. A. van Gisbergen, J. G. Snijders and T. Ziegler, *J. Comput. Chem.*, 2001, **22**, 931–967.
- 72 E. J. Baerends, et al., ADF2017, SCM, Theoretical Chemistry, Vrije Universiteit, Amsterdam, The Netherlands, <https://www.scm.com>.
- 73 J. Seibert, C. Bannwarth and S. Grimme, *J. Am. Chem. Soc.*, 2017, **139**, 11682–11685.
- 74 S. Grimme, *J. Chem. Theory Comput.*, 2019, **15**, 2847–2862.
- 75 D. Ben-Efraim and B. Green, *Tetrahedron*, 1974, **30**, 2357–2364.
- 76 H. Saito, *Magn. Reson. Chem.*, 1986, **24**, 835–852.

- 77 G. Barone, D. Duca, A. Silvestri, L. Gomez-Paloma, R. Riccio and G. Bifulco, *Chem. – Eur. J.*, 2002, **8**, 3240–3245.
- 78 A. Glättli, X. Daura, D. Seebach and W. F. van Gunsteren, *J. Am. Chem. Soc.*, 2002, **124**, 12972–12978.
- 79 Z. Brklača, M. Mališ, D. M. Smith and A.-S. Smith, *J. Chem. Theory Comput.*, 2014, **10**, 3270–3279.
- 80 T. Mori, S. Grimme and Y. Inoue, *J. Org. Chem.*, 2007, **72**, 6998–7010.
- 81 H. W. van Vlijmen, M. Schaefer and M. Karplus, *Proteins*, 1998, **33**, 145–158.
- 82 A. T. Cavasin, A. Hillisch, F. Uellendahl, S. Schneckener and A. H. Göller, *J. Chem. Inf. Model.*, 2018, **58**, 1005–1020.
- 83 M. P. Seddon, D. A. Cosgrove, M. J. Packer and V. J. Gillet, *J. Chem. Inf. Model.*, 2019, **59**, 98–116.
- 84 N.-O. Friedrich, C. de Bruyn Kops, F. Flachsenberg, K. Sommer, M. Rarey and J. Kirchmair, *J. Chem. Inf. Model.*, 2017, **57**, 2719–2728.
- 85 N. W. Mitzel and D. W. H. Rankin, *Dalton Trans.*, 2003, 3650–3662.
- 86 S. Blomeyer, M. Linnemannstöns, J. H. Nissen, J. Paulus, B. Neumann, H.-G. Stammller and N. W. Mitzel, *Angew. Chem., Int. Ed.*, 2017, **56**, 13259–13263.
- 87 Z. Feng, L. Chen, H. Maddula, O. Akcan, R. Oughtred, H. M. Berman and J. Westbrook, *Bioinformatics*, 2004, **20**, 2153–2155.
- 88 N.-O. Friedrich, A. Meyder, C. de Bruyn Kops, K. Sommer, F. Flachsenberg, M. Rarey and J. Kirchmair, *J. Chem. Inf. Model.*, 2017, **57**, 529–539.
- 89 T. Pavlov, M. Todorov, G. Stoyanova, P. Schmieder, H. Aladjov, R. Serafimova and O. Mekenyany, *J. Chem. Inf. Model.*, 2007, **47**, 851–863.
- 90 O. Gutten, D. Bím, J. Řezáč and L. Rulíšek, *J. Chem. Inf. Model.*, 2018, **58**, 48–60.
- 91 E. A. Coutsias, C. Seok and K. A. Dill, *J. Comput. Chem.*, 2004, **25**, 1849–1857.
- 92 A. Laio and M. Parrinello, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 12562–12566.
- 93 A. Barducci, M. Bonomi and M. Parrinello, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2011, **1**, 826–843.
- 94 V. Spiwok, P. Lipovová and B. Králová, *J. Phys. Chem. B*, 2007, **111**, 3073–3076.
- 95 E. Z. Wen, M.-J. Hsieh, P. A. Kollman and R. Luo, *J. Mol. Graphics Modell.*, 2004, **22**, 415–424.
- 96 V. Leone, F. Marinelli, P. Carloni and M. Parrinello, *Curr. Opin. Struct. Biol.*, 2010, **20**, 148–154.
- 97 M. A. Deriu, G. Grasso, J. A. Tuszyński, D. Gallo, U. Morbiducci and A. Danani, *PLoS Comput. Biol.*, 2016, **12**, 1–14.
- 98 J. Vymětal and J. Vondrášek, *J. Phys. Chem. B*, 2010, **114**, 5632–5642.
- 99 D. Hamelberg, J. Mongan and J. A. McCammon, *J. Chem. Phys.*, 2004, **120**, 11919–11929.
- 100 G. Torrie and J. Valleau, *J. Comput. Phys.*, 1977, **23**, 187–199.
- 101 X. Wu and S. Wang, *J. Chem. Phys.*, 1999, **110**, 9401–9410.
- 102 J. A. Rahman and J. C. Tully, *J. Chem. Phys.*, 2002, **116**, 8750–8760.
- 103 B. J. Berne and J. E. Straub, *Curr. Opin. Struct. Biol.*, 1997, **7**, 181–189.
- 104 M. M. Steiner, P.-A. Genilloud and J. W. Wilkins, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1998, **57**, 10236–10239.
- 105 S. Pal and K. Fichthorn, *Chem. Eng. J.*, 1999, **74**, 77–83.
- 106 X. G. Gong and J. W. Wilkins, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1999, **59**, 54–57.
- 107 C. A. F. de Oliveira, D. Hamelberg and J. A. McCammon, *J. Chem. Theory Comput.*, 2008, **4**, 1516–1525.
- 108 A. S. Kamenik, U. Lessel, J. E. Fuchs, T. Fox and K. R. Liedl, *J. Chem. Inf. Model.*, 2018, **58**, 982–992.
- 109 Y. Miao, W. Sinko, L. Pierce, D. Bucher, R. C. Walker and J. A. McCammon, *J. Chem. Theory Comput.*, 2014, **10**, 2677–2689.
- 110 D. B. Hibbert, *Chemom. Intell. Lab. Syst.*, 1993, **19**, 277–293.
- 111 R. Leardi, *J. Chemom.*, 2001, **15**, 559–569.
- 112 H. Valdes, K. Pluhackova, M. Pitonák, J. Řezáč and P. Hobza, *Phys. Chem. Chem. Phys.*, 2008, **10**, 2747–2757.
- 113 M. Bursch, H. Neugebauer and S. Grimme, *Angew. Chem., Int. Ed.*, 2019, **58**, 11078–11087.
- 114 L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi and S. Grimme, *Phys. Chem. Chem. Phys.*, 2017, **19**, 32184–32215.
- 115 M. Marianski, A. Supady, T. Ingram, M. Schneider and C. Baldauf, *J. Chem. Theory Comput.*, 2016, **12**, 6157–6168.
- 116 P. S. Brahmkhatriya, P. Dobeš, J. Fanfrlík, J. Řezáč, K. Paruch, A. Bronowska, M. Lepšík and P. Hobza, *Curr. Comput.-Aided Drug Des.*, 2013, **9**, 118–129.
- 117 J. Řezáč and P. Hobza, *J. Chem. Theory Comput.*, 2012, **8**, 141–151.
- 118 J. J. P. Stewart, *J. Mol. Model.*, 2007, **13**, 1173.
- 119 J. Řezáč, D. Bím, O. Gutten and L. Rulíšek, *J. Chem. Theory Comput.*, 2018, **14**, 1254–1266.
- 120 D. I. Sharapa, A. Genaev, L. Cavallo and Y. Minenkov, *Chem. Phys. Chem.*, 2019, **20**, 92–102.
- 121 K. B. Wiberg, *Tetrahedron*, 1968, **24**, 1083–1096.
- 122 I. Mayer, *J. Comput. Chem.*, 2007, **28**, 204–221.
- 123 L. B. Kier, *Quant. Struct.-Act. Relat.*, 1989, **8**, 221–224.
- 124 W. Fisanick, K. P. Cross and A. Rusinko, *Tetrahedron Comput. Methodol.*, 1990, **3**, 635–652.
- 125 K. Eichkorn, O. Treutler, H. Öhm, M. Häser and R. Ahlrichs, *Chem. Phys. Lett.*, 1995, **242**, 652–660.
- 126 F. Weigend, *Phys. Chem. Chem. Phys.*, 2006, **8**, 1057.
- 127 S. Grimme, J. G. Brandenburg, C. Bannwarth and A. Hansen, *J. Chem. Phys.*, 2015, **143**, 054107.
- 128 S. R. Domingos, C. Pérez, C. Medcraft, P. Pinacho and M. Schnell, *Phys. Chem. Chem. Phys.*, 2016, **18**, 16682–16689.
- 129 National Institute of Advanced Industrial Science and Technology, SDBSWeb, 2019, <https://sdbs.db.aist.go.jp>.
- 130 C. Adamo and V. Barone, *J. Chem. Phys.*, 1999, **110**, 6158–6170.
- 131 F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297–3305.
- 132 A. Klamt and G. Schüürmann, *J. Chem. Soc., Perkin Trans. 2*, 1993, 799–805.
- 133 A. Klamt and M. Diedenhofen, *J. Phys. Chem. A*, 2015, **119**, 5439–5445.
- 134 I.-J. Chen and N. Foloppe, *Bioorg. Med. Chem.*, 2013, **21**, 7898–7920.

- 135 C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.*, 2016, **72**, 171–179.
- 136 R. R. Hudgins and M. F. Jarrold, *J. Am. Chem. Soc.*, 1999, **121**, 3494–3501.
- 137 M. F. Jarrold, *Phys. Chem. Chem. Phys.*, 2007, **9**, 1659–1671.
- 138 F. Schubert, M. Rossi, C. Baldauf, K. Pagel, S. Warnke, G. von Helden, F. Filsinger, P. Kupser, G. Meijer, M. Salwiczek, B. Koksich, M. Scheffler and V. Blum, *Phys. Chem. Chem. Phys.*, 2015, **17**, 7373–7385.
- 139 E. Caldeweyher, S. Ehlert, A. Hansen, H. Neugebauer, S. Spicher, C. Bannwarth and S. Grimme, *J. Chem. Phys.*, 2019, **150**, 154122.
- 140 K. Hukushima and K. Nemoto, *J. Phys. Soc. Jpn.*, 1996, **65**, 1604–1608.
- 141 U. H. Hansmann, *Chem. Phys. Lett.*, 1997, **281**, 140–150.
- 142 E. Marinari and G. Parisi, *Europhys. Lett.*, 1992, **19**, 451–458.
- 143 Y. Sugita and Y. Okamoto, *Chem. Phys. Lett.*, 1999, **314**, 141–151.
- 144 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.
- 145 J. G. Brandenburg, C. Bannwarth, A. Hansen and S. Grimme, *J. Chem. Phys.*, 2018, **148**, 064104.
- 146 M. Marković, N. Judaš and J. Sabolović, *Inorg. Chem.*, 2011, **50**, 3632–3644.
- 147 M. Bette, T. Kluge, J. Schmidt and D. Steinborn, *Organometallics*, 2013, **32**, 2216–2227.
- 148 S. Pandit and M. De, *J. Phys. Chem. C*, 2017, **121**, 600–608.
- 149 S. Goenka, V. Sant and S. Sant, *J. Controlled Release*, 2014, **173**, 75–88.
- 150 A. D. McNaught and A. Wilkinson, *IUPAC Compendium of Chemical Terminology*, (the “Gold Book”), Blackwell Scientific Publications, Oxford, 2nd edn, 1997.
- 151 M. G. Medvedev, M. V. Panova, G. G. Chilov, I. S. Bushmarinov, F. N. Novikov, O. V. Stroganov, A. A. Zeifman and I. V. Svitanko, *Mendeleev Commun.*, 2017, **27**, 224–227.
- 152 G. Jindal and A. Warshel, *J. Phys. Chem. B*, 2016, **120**, 9913–9921.
- 153 V. S. Bryantsev, M. S. Diallo, A. C. T. van Duin and W. A. Goddard, *J. Chem. Theory Comput.*, 2009, **5**, 1016–1026.
- 154 N. A. Seifert, A. S. Hazrah and W. Jäger, *J. Phys. Chem. Lett.*, 2019, **10**, 2836–2841.
- 155 T. Risthaus, M. Steinmetz and S. Grimme, *J. Comput. Chem.*, 2014, **35**, 1509–1516.
- 156 P. M. Lalli, B. A. Iglesias, H. E. Toma, G. F. de Sa, R. J. Daroda, J. C. Silva Filho, J. E. Szulejko, K. Araki and M. N. Eberlin, *J. Mass Spectrom.*, 2012, **47**, 712–719.
- 157 S. Warnke, J. Seo, J. Boschmans, F. Sobott, J. H. Scrivens, C. Bleiholder, M. T. Bowers, S. Gewinner, W. Schöllkopf, K. Pagel and G. von Helden, *J. Am. Chem. Soc.*, 2015, **137**, 4236–4242.
- 158 J. Seo, S. Warnke, S. Gewinner, W. Schöllkopf, M. T. Bowers, K. Pagel and G. von Helden, *Phys. Chem. Chem. Phys.*, 2016, **18**, 25474–25482.
- 159 H. Xia and A. B. Attygalle, *J. Am. Soc. Mass Spectrom.*, 2017, **28**, 2580–2587.
- 160 Y. Zhu, L. A. Hamlow, C. C. He, S. F. Strobel, J. K. Lee, J. Gao, G. Berden, J. Oomens and M. T. Rodgers, *J. Phys. Chem. B*, 2016, **120**, 8892–8904.
- 161 L. Antonov, V. Deneva, S. Simeonov, V. Kurteva, D. Nedeltcheva and J. Wirz, *Angew. Chem., Int. Ed.*, 2009, **48**, 7875–7878.
- 162 P. Pospisil, P. Ballmer, L. Scapozza and G. Folkers, *J. Recept. Signal Transduction*, 2003, **23**, 361–371.
- 163 L. Guasch, W. Yapamudiyansel, M. L. Peach, J. A. Kelley, J. J. Barchi and M. C. Nicklaus, *J. Chem. Inf. Model.*, 2016, **56**, 2149–2161.
- 164 P. G. Mezey and J. J. Ladik, *Theor. Chim. Acta*, 1979, **52**, 129–145.
- 165 E. Nir, C. Janzen, P. Imhof, K. Kleinermanns and M. S. de Vries, *J. Chem. Phys.*, 2001, **115**, 4604–4611.
- 166 H. Langer and N. L. Doltsinis, *J. Chem. Phys.*, 2003, **118**, 5400–5407.
- 167 H. Chen and S. Li, *J. Phys. Chem. A*, 2006, **110**, 12360–12362.
- 168 C. M. Marian, *J. Phys. Chem. A*, 2007, **111**, 1545–1553.
- 169 L. M. A. Quintana, S. I. Johnson, S. L. Corona, W. Villatoro, W. A. Goddard, M. K. Takase, D. G. VanderVelde, J. R. Winkler, H. B. Gray and J. D. Blakemore, *Proc. Natl. Acad. Sci. U. S. A.*, 2016, **113**, 6409–6414.