The Journal of Chemical Physics

RESEARCH ARTICLE | MARCH 21 2024

CREST—A program for the exploration of low-energy molecular chemical space

Special Collection: Modular and Interoperable Software for Chemical Physics

Philipp Pracht ; Stefan Grimme ; Christoph Bannwarth ; Fabian Bohle ; Sebastian Ehlert ; Gereon Feldmann ; Johannes Gorges ; Marcel Müller ; Tim Neudecker ; Christoph Plett ; Sebastian Spicher ; Pit Steinbach ; Patryk A. Wesołowski ; Felix Zeller

J. Chem. Phys. 160, 114110 (2024)
https://doi.org/10.1063/5.0197592

The Journal of Chemical Physics ARTICLE pubs.aip.org/aip/jcp

CREST—A program for the exploration of low-energy molecular chemical space

Cite as: J. Chem. Phys. 160, 114110 (2024); doi: 10.1063/5.0197592
Submitted: 13 January 2024 • Accepted: 29 February 2024 •
Published Online: 21 March 2024

Philipp Pracht,[1,a)] Stefan Grimme,[2] Christoph Bannwarth,[3] Fabian Bohle,[2] Sebastian Ehlert,[4] Gereon Feldmann,[3] Johannes Gorges,[2] Marcel Müller,[2] Tim Neudecker,[5] Christoph Plett,[2] Sebastian Spicher,[6] Pit Steinbach,[3] Patryk A. Wesołowski,[1] and Felix Zeller[5]

AFFILIATIONS

[1] Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom
[2] Mulliken Center for Theoretical Chemistry, Institute for Physical and Theoretical Chemistry, University of Bonn, Beringstr. 4, 53115 Bonn, Germany
[3] Institute for Physical Chemistry, RWTH Aachen University, Melatener Str. 20, 52056 Aachen, Germany
[4] AI4Science, Microsoft Research, Evert van de Beekstraat 354, 1118 CZ Schiphol, The Netherlands
[5] Institute for Physical and Theoretical Chemistry, University of Bremen, 28359 Bremen, Germany
[6] BASF SE, Carl-Bosch Straße 38, 67056 Ludwigshafen am Rhein, Germany

Note: This paper is part of the JCP Special Topic on Modular and Interoperable Software for Chemical Physics.
[a)] **Author to whom correspondence should be addressed:** pp555@cam.ac.uk

ABSTRACT

Conformer–rotamer sampling tool (CREST) is an open-source program for the efficient and automated exploration of molecular chemical space. Originally developed in Pracht *et al.* [Phys. Chem. Chem. Phys. **22**, 7169 (2020)] as an automated driver for calculations at the extended tight-binding level (xTB), it offers a variety of molecular- and metadynamics simulations, geometry optimization, and molecular structure analysis capabilities. Implemented algorithms include automated procedures for conformational sampling, explicit solvation studies, the calculation of absolute molecular entropy, and the identification of molecular protonation and deprotonation sites. Calculations are set up to run concurrently, providing efficient single-node parallelization. CREST is designed to require minimal user input and comes with an implementation of the GFN*n*-xTB Hamiltonians and the GFN-FF force-field. Furthermore, interfaces to any quantum chemistry and force-field software can easily be created. In this article, we present recent developments in the CREST code and show a selection of applications for the most important features of the program. An important novelty is the refactored calculation backend, which provides significant speed-up for sampling of small or medium-sized drug molecules and allows for more sophisticated setups, for example, quantum mechanics/molecular mechanics and minimum energy crossing point calculations.

I. INTRODUCTION

Scientific software development in chemical physics has always been driven by the need for greater efficiency and adaptability in the face of evolving theoretical methods and advances in computer hardware.[1–4] The conventional paradigm of self-contained codes, while capable of reproducing specific functionalities, has inadvertently given rise to duplicated efforts and faces challenges in staying ahead of the rapid pace of change. Over the past couple of years, and in response to these challenges, software developers are increasingly embracing a "divide and conquer" strategy and creating interoperable software to advance chemical physics research.[5–8] New developments focus on creating software tailored to specific tasks within the broader field of computational chemistry, which not only streamlines development efforts but also facilitates easier adaptation to new methodologies and hardware architectures. Simultaneously, researchers are adopting a collaborative ethos by enhancing the interoperability of their software to enable the seamless

integration of features from diverse program packages. Popular examples from the wider field of chemical physics include the atomic simulation environment (ASE),[9] the large-scale atomic/molecular massively parallel simulator (LAMMPS),[10] and the cheminformatics software package RDKit.[11] The collaborative approach enables the coordinated utilization of atomic simulation capabilities across various programs, enhancing the prospects for innovation and providing researchers with a comprehensive toolkit to address complex problems in many areas of chemical physics research.

One such field of research with continued attention is molecular chemical space exploration (CSE).[12–14] Chemical space and, in particular, the low-energy molecular chemical space, e.g., featuring weak metallic interactions that require a description of the electronic structure. The need for sampling this vast space becomes quickly apparent considering that even small changes to the studied system can strongly affect the calculated properties according to the structure-activity-principle. Among the most prominent and insightful examples for members of the low-energy chemical space are molecular conformations. Exploring the conformational space is a fundamental task and often among the first steps in any computational study concerning molecules or clusters[15,16] (cf. Fig. 1).

The key problem here is the rapid growth of the conformational space of a molecule with its size, which is roughly exponential ($\approx 3^N$) with the number of "rotatable" bonds (N).[14,19–21] Associated with that is a huge computational effort, giving rise to a variety of programs and algorithms specific to the sampling task. A good review of the topic was provided by Hawkins,[20] to which we refer the reader at this point. Typical techniques include various types of molecular dynamics (MD) based sampling,[22,23] Monte Carlo based algorithms such as basin-hopping global optimization,[21,24,25] and heuristic/rule-based chemoinformatic conformer generators as implemented in Omega,[26] Molassembler,[27] ConfGen,[28] Frog2,[29] RDKit,[11] and others.[30–35] Similar challenges and algorithmic varieties are encountered for other members of the low-energy chemical space, including the tasks of determining protonation sites, finding other types of isomers such as tautomers, or describing the molecule's behavior upon (explicit) solvation.

From a physical point of view, all CSE applications are tied to analysis of the potential energy landscape (PEL),[14,21] imposing an important and often limiting condition on any kind of sampling: The exploration can only ever be as good (or bad) as the potential used to investigate the underlying energy landscape. While the best quality energy landscapes, i.e., the most physical ones, are usually obtained using first principles electronic structure methods, for example, density functional theory (DFT), even medium-sized pharmaceutical molecules require thousands to millions of energy and gradient evaluations to explore the chemical space, making large-scale high-throughput calculations impractical at such quantum mechanical levels.[15,16] Much less costly molecular mechanics (MM) calculations like classical force-field (FF) or coarse-grained methods allow exploration of substantially larger systems but often suffer in terms of accuracy and parameter availability. A balance between computational cost and accuracy is essential, which in recent years has been addressed by machine learning (ML) methods[36,37] and the revival of semiempirical quantum mechanics (SQM).[38,39]

In previous studies, we presented the CREST code[19,40] (abbreviation for Conformer–Rotamer Ensemble Sampling Tool) as a versatile program for studying the chemical space of drug sized molecules at the semiempirical extended tight-binding (xTB) level. Being based on the GFNn-xTB methods,[39] CREST offered capabilities for automated conformational sampling, conformational entropy calculations,[40,41] explicit solvation studies,[42,43] protonation site determination,[44] and more[45–48] for systems containing any element up to radon ($Z \leq 86$). The code was developed in order to provide a robust, broadly applicable, and easy-to-use tool for both experts and non-experts in the field.

This paper reports on the status of the program's latest iteration, CREST 3.0, including new implementations and a major re-factorization of the original Fortran code. In the following, Sec. II summarizes the primary algorithm implemented in CREST, followed by an overview of further technical particulars in Sec. III, which includes discussions of implemented potentials and atomic simulation methods. Information about the program distribution is provided in Sec. IV.

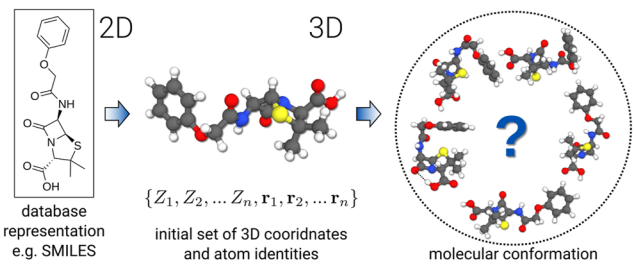

FIG. 1. The first step in many computational chemical physics studies: Conformational sampling is necessary to find the energetically favored structures of a molecule. Typically, starting from a low-dimensional representation such as the molecular Lewis structure (left) or its SMILES[17] or InChi[18] code, an initial set of three-dimensional coordinates is created (center), which can then be applied in conformational sampling applications (right).

II. ALGORITHMS AND APPLICATIONS

A. Conformational sampling

Conformational sampling is the original and eponymous application of CREST. Historically, the program was developed with a normal mode-following approach for generating conformations in the context of nuclear magnetic resonance (NMR) spectra calculation.[49] Soon after, this procedure was discarded in favor of a quicker and more robust metadynamics-based algorithm,[19,50] which is the basis for current implementations. Herein, a main design aspect of CREST is the automation of all calculations. Starting from an initial user-defined configuration of the input system (nuclear charges, Cartesian coordinates, molecular charge, multiplicity/spin), the system is analyzed, and sensible simulation parameters are determined. In particular, the overall metadynamics (MTD) run times and bias potential strengths are calculated based on the system size and a flexibility measure ξ_f (for details on this, see Refs. 19 and 40) to allow an automatic initialization of simulations. Furthermore, an initial geometry optimization is conducted to check for any occurring changes to the molecular topology that could be caused simply by the potential and, if observed, safely interrupt the sampling process.

The general workflow for CREST's default conformational sampling routine is outlined in Fig. 2.

Following the automatic setup, a block of metadynamics simulations (see Sec. III B 1) is performed using the Cartesian root-mean-square-deviation (RMSD)[51] as the collective variables.[50] The applied metadynamics bias potential ensures a reliable generation of new molecular conformations by blocking out previously visited minima on the potential energy landscape. Since automation is the key step and different molecules require different widths and strengths of the bias potential, i.e., the prefactor and exponent employed in a Gaussian function [see Eq. (15)], several MTD simulations with various bias settings are run in parallel to allow exploration for a large variety of systems. MTDs routinely cross high energy barriers during the simulation to generate new conformations; however, small lower barrier changes such as side-chain rotations are sometimes not adequately captured. For this purpose, two additional steps are implemented. First, a block of regular MD simulations is performed using two different temperatures, starting from the lowest few (typically four) conformers found up to this point. This will generate predominantly conformations that are connected via small barriers relative to those overcome by the metadynamics and target the sampling of rotamers, which are degenerate conformers featuring nuclear permutation (see Sec. II A 2). Second, the rotamer generation is further augmented by a procedure adapted from genetic crossing (GC) algorithms.[31,52] Here, the internal coordinates (Z-matrices) of two structures, systematically drawn from the current ensemble, are used to calculate a mutation for the lowest conformer. The newly generated conformations often differ primarily in the orientation of functional groups, leading to a completion of the rotamer rather than the conformer space. Furthermore, the use of internal coordinates makes the GC step less robust than the preceding molecular and metadynamics sampling. Between any of these steps, molecular geometries are optimized in a multi-level optimization approach[19] using the Approximate Normal Coordinate OPTimizer (ANCOPT) algorithm (Sec. III B 2), and the ensemble is sorted via the Conformer–Rotamer Ensemble GENeration (CREGEN) procedure (Sec. II A 2) to distinguish conformers, rotamers, and prune duplicates. If, at any point, a lower-energy conformer is found, the entire procedure is iteratively restarted with this new structure, which can be interpreted as the progression of a Markov chain. The combined workflow is hence abbreviated as the iterative metadynamics-genetic crossing (iMTD-GC) method, which was designed to both reliably reproduce the global minimum of drug-like molecules and to provide good ensemble coverage as measured by an ensemble entropy, as outlined in Ref. 50. The corresponding code is optimized for single-node (OpenMP) parallelization to allow the concurrent execution of all molecular dynamics and geometry optimization processes. In theory, the algorithm is "embarrassingly parallel" up to the point where each structure is processed by a single thread.

The development of the current program version was concerned with improving the technical performance of the code. Previously designed as an OpenMP scheduler for calculations performed with the `xtb` program,[53] the CREST 3.0 source code

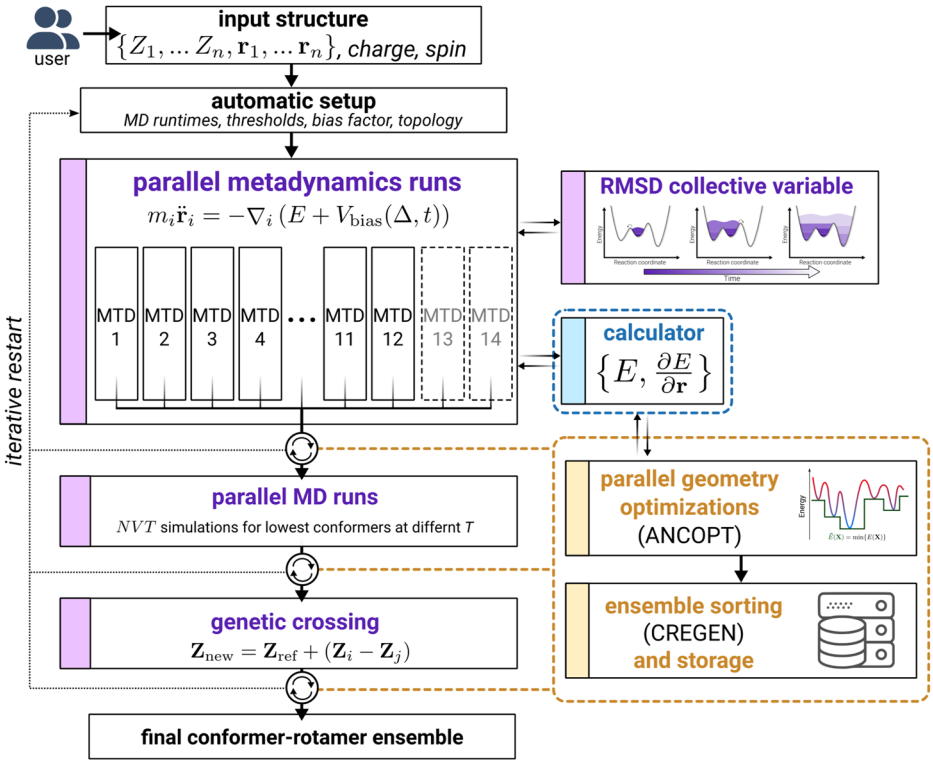

FIG. 2. Outline of the iterative metadynamics and genetic crossing (iMTD-GC) workflow to generate molecular conformations.

has since largely been refactored and extended to implement the corresponding calculators directly (cf. Sec. III). For workflows such as iMTD-GC, the change from a subprocess-based to the current calculator/API-based program provides significant computation-time improvements, as demonstrated for three typical target molecules in Fig. 3.

For the molecules provided as examples, namely *n*-octane, amoxicillin, and Tamiflu, notable reductions in computation time ranging from factors 1.6–4.7 are achieved when using CREST 3.0 as opposed to the earlier program version. Running iMTD-GC for these systems requires a number of energy and gradient evaluations on the order of $\mathcal{O}(10^5)$, making the achievable performance highly dependent on the chosen level of theory, the number of parallel running CPUs, and the system size. As a rule-of-thumb, savings compared to the `xtb` subprocess version (CREST ≤ 2.12) will be greatest for small systems where a separate process and I/O operations would produce non-negligible overhead compared to the actual energy and gradient evaluation. Unsurprisingly, the greatest computation time reduction was achieved for the flexible but small *n*-octane molecule, running iMTD-GC with GFN-FF. In conclusion, the improved efficiency of CREST 3.0 renders it more suitable for high-throughput conformational sampling of small-molecule databases than previous versions.

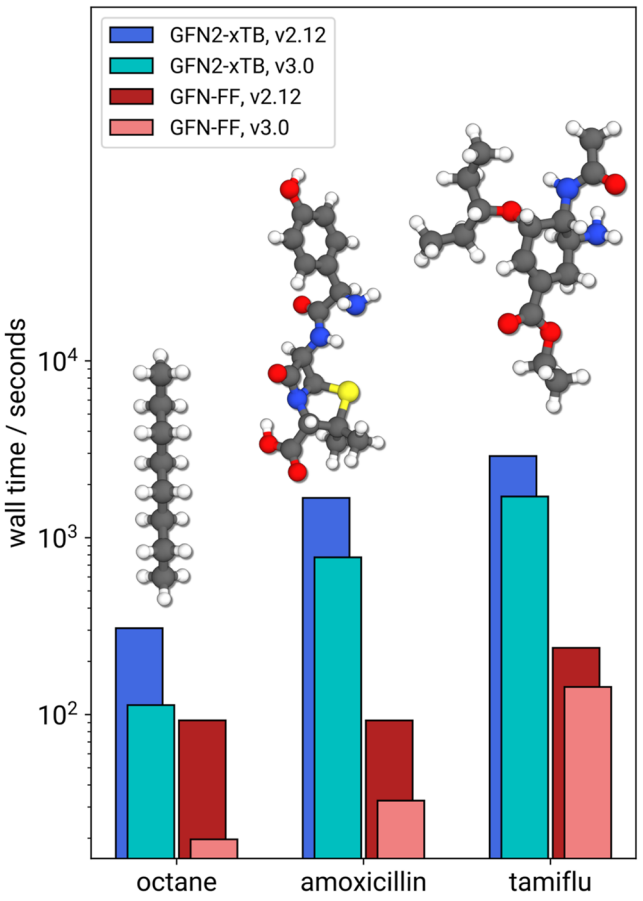

FIG. 3. Computational wall times for running iMTD-GC with CREST versions 2.12 and 3.0 for the *n*-octane, amoxicillin, and Tamiflu molecules at the GFN2-xTB (blue) and GFN-FF (red) levels of theory. All calculations were run on the same system, each using 16 threads of Intel Xeon Gold 6326 CPUs (2.90 GHz). Timings are presented on a logarithmic scale. The revised program version 3.0 (light-colored bars) shows improved efficiency for all cases.

1. Using semiempirical methods for conformational sampling

The conformational sampling of CREST was explicitly developed in combination with the (semi)empirical methods of the GFN*n*-xTB/-FF family.[39,54–57] These methods are supposed to yield reasonably accurate PELs, associated with only little computational effort compared to density functional or wave function theory. Both computational efficiency and sufficient accuracy are essential requirements for extensive conformational sampling approaches. To answer the question of how well the approximated PELs match high-level references, conformational energy benchmarks are evaluated in the following. In Fig. 4(a), GFN2-xTB and GFN-FF are compared to competing methods with a focus on conformational energies. Comparisons are grouped according to force-field (FF), SQM, and (composite) DFT and Hartree–Fock (HF)-based approaches. For a statistically proper and fair evaluation across a multitude of different benchmarks, we chose the established WTMAD-2 averaging scheme, originally introduced with the GMTKN55 benchmark database[58] (see the supplementary material for details). Besides the conformer benchmarks included in GMTKN55 (Amino20x4: amino acid conformers, BUT14DIOL: butane-1,4-diol conformers, ICONF: inorganic systems, MCONF: melatonin conformers, PCONF21: tri- and tetrapeptide conformers, SCONF: sugar conformers, UPU23: RNA-backbone conformers), further conformer benchmarks from the literature were taken into account for the illustrated WTMAD-2 values: ACONFL[61] (conformers of longer *n*-alkane chains), MPCONF196[64] (conformers of acyclic and cyclic model peptides and other macrocycles), 37CONF8[63] (conformers of organic drug-like molecules), TMCONF16 (subset of TMCONF40,[60] conformers of transition-metal complexes), and MALT222[66] (α-maltose conformers).

In the force-field regime (blue), two "general" force-fields were tested: the general Amber force-field[73] (GAFF) is an extension of the Amber force-field for proteins and nucleic acids, next to GFN-FF,[57] the generic polarizable force-field of the GFN family of methods utilized in CREST. Besides GFN2-xTB, the considered SQM methods comprise another tight-binding (TB) method (cyan), DFTB3-D3(BJ) in the 3ob-3-1 parameterization,[74–77] and the Hartree–Fock-based PM6-D3H4X[78–80] (red). HF-3c, a minimal basis set Hartree–Fock (HF) with specific approximations,[81] has been grouped together with the HF-based PM6-D3H4X semiempirical method. Generally, the TB methods are more accurate than the FFs, whereas PM6-D3H4X is outperformed by GFN-FF. GFN-FF and GFN2-xTB stand out with noticeably lower WTMAD-2s in comparison to GAFF on the one end and PM6-D3H4X and DFTB3-D3(BJ) on the other end. Notably, neither a "proper" QM treatment of HF-3c nor a semiempirical method lead to results superior to the respective WTMAD-2, which is slightly larger than that of GFN2-xTB. The DFT methods are grouped into two categories, composite DFT methods employing smaller basis sets (coral)

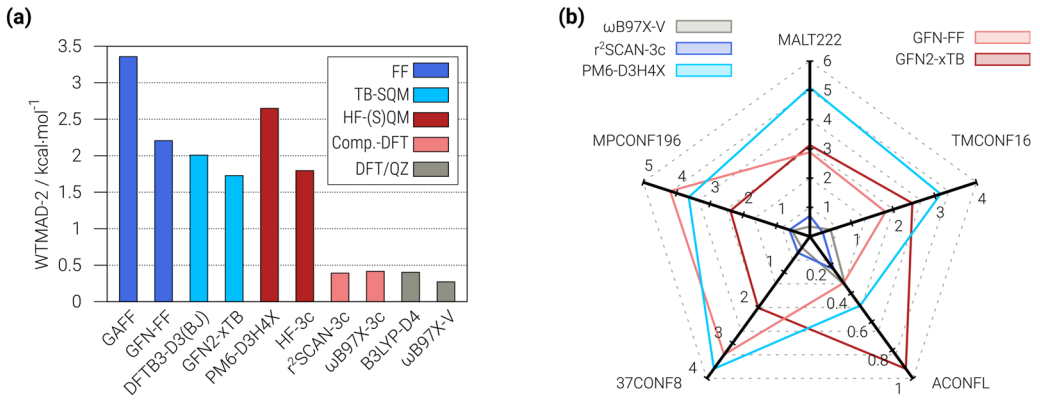

FIG. 4. (a) Weighted total mean absolute deviations (WTMAD-2, see Refs. 58 or 59 for the definition) to reference relative conformational energies for various methods in kcal mol^{-1}. The benchmark sets Amino20x4, BUT14DIOL, ICONF, MCONF, PCONF21, SCONF, UPU23 (all part of GMTKN55[58]), TMCONF16,[60] ACONFL,[61,62] 37CONF8,[63] MPCONF196,[64,65] and MALT222[66] were considered for the presented WTMAD-2. TMCONF16 (for GAFF additionally ICONF) were not taken into account for the WTMAD-2 calculation for DFTB3-D3(BJ) and GAFF since appropriate parameters were not available. (b) Mean absolute errors (MAEs, kcal mol^{-1}) for selected benchmark sets and methods. For B3LYP-D4[67–70] and ωB97X-V,[71] def2-QZVP(P) basis sets[72] were used throughout. The benchmark data employed in the figures, including the WTMAD-2 calculation, are provided in a spreadsheet in the supplementary material.

and DFT functionals in practically converged quadruple-ζ (QZ) basis sets (gray). Remarkably, the more efficient "3c" composite methods, r^{2}SCAN-3c[82] and ωB97X-3c,[59] reach virtually the same performance for conformational energies as the (range-separated) hybrid functionals B3LYP-D4[67–70] and ωB97X-V[71] in QZ bases. At present, state-of-the-art DFT methods are substantially more accurate in describing the subtle energy differences in different molecular conformations with four to five times lower mean absolute errors (MAEs) than SQM or FF methods but are at the same time much more expensive. In Fig. 4(b), MAEs of selected methods are shown for the non-GMTKN55 conformer energy benchmark sets. Consistent with Fig. 4(a), ωB97X-V/QZ and r^{2}SCAN-3c exhibit very similar MAEs throughout. PM6-D3H4X is in almost all cases inferior to GFN2-xTB except for ACONFL, for which GFN2-xTB has visibly larger errors than the competing methods. Note, however, that errors for none of the tested methods exceed the target accuracy of 1 kcal mol^{-1} for the ACONFL set. Even though the differences in overall performance are only moderate, GFN2-xTB and GFN-FF show very different strengths and weaknesses. GFN2-xTB is distinctly more accurate for organic drug-like molecules and peptides in 37CONF8 and MPCONF196, while GFN-FF is better at describing long alkane chains in ACONFL and transition metal complexes in TMCONF16, respectively. However, with regards to the latter benchmark, we note that transition metal complexes are particularly challenging, and related studies report an opposite trend for GFN-FF and GFN2-xTB.[60]

For further comparison, Fig. 5 contains relative energies with respect to the energetically lowest conformer for ensembles from the MPCONF196 set, correlated with the reference relative energies. The ωB97X-V relative energies agree excellently with the reference data and are mostly within the desired accuracy of <1 kcal mol^{-1}. GFN2-xTB and, more pronounced, GFN-FF reveal a stronger scatter of the relative energies but still show the correct qualitative trends in most cases, which is mandatory for conformational sampling. This expectation can easily be confirmed by looking at the corresponding coefficients of determination and Spearman rank correlation coefficients, which are shown in Table I.

Given that the two GFN methods are multiple orders of magnitude faster than the above-discussed DFT methods, they are reasonable choices for exploring the conformational space, which is consistent with findings in related literature.[60,61,63,65] However, the significant differences in accuracy may necessitate the re-ranking

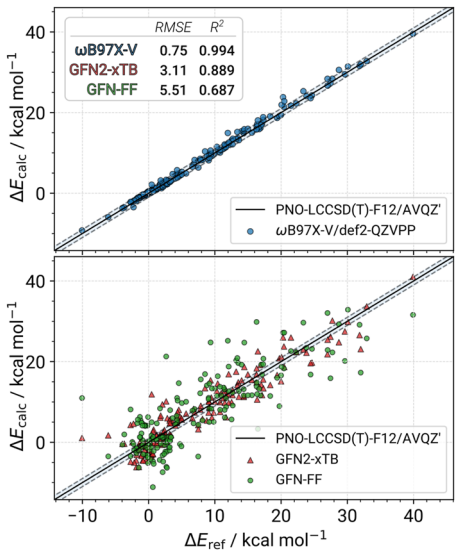

FIG. 5. Correlation of reference relative conformational energies of the MPCONF196 benchmark set[64] (reference data taken from Ref. 65) compared to ωB97X-V (top), GFN2-xTB, and GFN-FF relative energies (bottom). The dashed lines above and below the parity diagonal indicate the desired "chemical accuracy" of 1 kcal mol^{-1}. The RMSEs for the three methods are provided in kcal mol^{-1}. The coefficients of determination (R^2) are dimensionless.

TABLE I. Coefficients of determination (R^2) and Spearman rank correlation coefficients (r_s) computed for the data shown in Fig. 5. Both coefficients are dimensionless.

Method	R^2	r_s
ωB97X-V	0.994	0.996
GFN2-xTB	0.889	0.925
GFN-FF	0.687	0.787

and optimization of conformer ensembles at the DFT level if highly accurate results are required. Furthermore, while results for the ensemble entropy (see Sec. II A 4) suggest good ensemble coverage and similar sampling fitness with both SQM and FF methods, minima on the PEL might differ and occasionally collapse into the same structure upon optimization at another level of theory. This is an issue of potential rather than sampling methodology, and the user is well advised to investigate such effects on a case-by-case basis.

2. Ensemble sorting algorithm

The generation of molecular configurations in a top-down approach necessitates the implementation of bookkeeping algorithms for tracking the ensemble. Each configuration is defined as a minimum in the potential energy landscape. The energy landscape complexity grows exponentially with degrees of freedom (DOF), and even for medium-sized systems, like a non-covalently bound cluster of 55 atoms, it is estimated to be on the order of $\mathcal{O}(10^{10})$ minima.[24,25] In the special case of molecular conformations, the topology of covalent bonds between atoms may not change, which employs a constraint on the overall energy landscape complexity. Still, the number of accessible minima scales roughly as 3^n, with n being the number of rotatable (e.g., sp^3-sp^3 carbon) bonds.[14] Due to local permutational symmetry, some side chains will lead to degenerate conformations referred to as rotamers. Common examples include methyl or phenyl groups, where the dihedral rotation around a single carbon bond will generate three and two indistinguishable rotamers, respectively. Accordingly, ensembles including both types of isomers can be referred to as conformer–rotamer ensembles (CREs). The CRE within a certain energy window is typically significantly smaller than the overall conformational space and the rather crude 3^n approximation. This is particularly true for ensembles generated with CREST,[83,84] which classifies and filters the generated structures already during the iterative sampling procedure. An *a priori* prediction of the expected number of conformations within such energy windows is, therefore, hardly possible and depends on the PEL shape of the individual molecule.

A numerical classification and bookkeeping algorithm for the Conformer–Rotamer Ensemble GENeration (CREGEN) is implemented in CREST to aid in the identification of conformers, their associated rotamers, and duplicated structures. The general multistep workflow is outlined in Fig. 6. CREGEN is composed of several steps: In the initial ensemble preprocessing, first a covalent-bond topology is determined for all structures based on neighbor lists calculated from the atomic coordination number (CN) as taken from the D4 scheme.[70,85] All structures that do not match the topology of the original input are discarded by default. In the next step, an energy window is applied to discard all structures with an energy higher

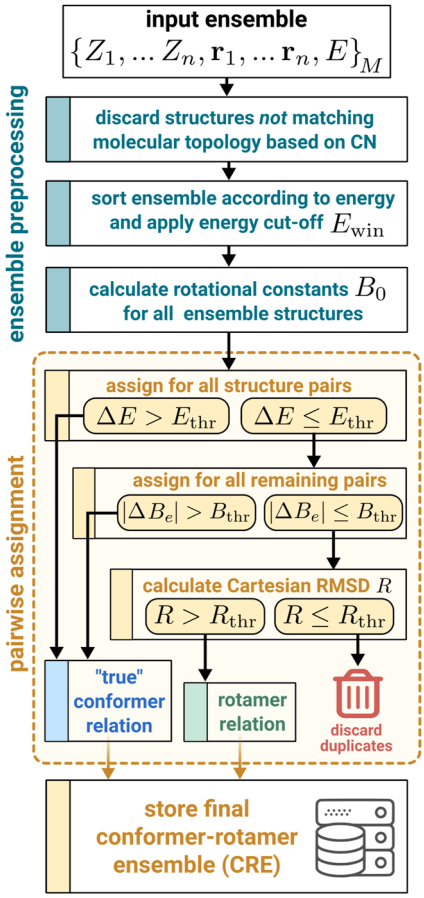

FIG. 6. Outline of the CREGEN sorting algorithm. The workflow processes an initial ensemble of *M* structures, each represented by a set of atomic coordinates and an associated energy, to eliminate duplicate structures and determine pairwise relations of "true" conformers and rotamers that are collected in the CRE.

than a given reference E_{win} relative to the lowest-energy structure in the current ensemble. Finally, rotational constants B_e are calculated for all structures, serving as a structural parameter for further comparisons.

The main pairwise classification of structures as conformers, rotamers, or duplicates is based on a sequential comparison of three criteria, namely the energy difference, the difference of rotational constants, and the Cartesian RMSD. If the energy difference (ΔE) between two structures surpasses a predetermined threshold (E_{thr}), these two candidates, defined as minima on the potential energy surface (PES), are categorized as conformers. Otherwise, a computationally expensive structure comparison becomes necessary. As an initial and cost-effective criterion, the difference between the rotational constants ($|\Delta B_e|$) can be evaluated against an empirical anisotropically corrected threshold value (B_{thr}). Since rotational constants remain unaffected by atom permutations, minimal differences in $|\Delta B_e|$ between two rotamers or duplicates of the same conformer are expected. In contrast, true conformers should exhibit notable differences in this aspect, even if they are close in energy.

TABLE II. Default thresholds for the energy window E_{win}, energy difference E_{thr}, Cartesian RMSD R_{thr}, and rotational constant B_{thr}. The latter is dynamically adjusted to the system. Systems with a higher anisotropy of B_e use a larger B_{thr}.

Parameter	Default value
E_{win}	6.0 kcal mol^{-1}
E_{thr}	0.05 kcal mol^{-1}
B_{thr}	1.0%–2.5%
R_{thr}	0.125 Å

Hence, to identify those structures and eliminate duplicates, an RMSD (R) between two potential rotamers is computed using a quaternion algorithm,[51] and the result is compared to a third threshold (R_{thr}). Only structures surpassing this threshold are retained in the final Conformer Redundancy Elimination (CRE). The default threshold values for CREGEN are detailed in Table II.

The CREGEN algorithm is optimized for processing large ensembles. Here, the most expensive step is the pairwise RMSD calculation. Due to the preceding energy and rotational constant cutoffs, the RMSD comparison has a complexity of approximately $\mathcal{O}(n \log n)$ but, at worst, will scale as $\mathcal{O}(n^2)$, with n being the ensemble size within E_{win}. Notably, the threshold-based numerical construction of the algorithm may lead to false-positive identification of true conformers as rotamers or vice versa. This problem occurs more frequently with increasing system size (N_{at}), which can only partially be compensated by an adjustment of $\{E_{thr}, R_{thr}, B_{thr}\}$. The conformer identification may, therefore, be supplemented by post-processing via permutation invariant comparisons of atomic configurations. Since the latter is accessible only via costly methods such as finding the permutational minimum RMSD via a Hungarian algorithm[86] or modern schemes such as the use of smooth atomic overlap (SOAP) kernels,[87] no implementation is currently available in CREST, and the reader is referred to the literature.[88–90]

For the reduction of ensemble sizes or the identification of representative structures, CREST additionally implements a clustering approach based on principle component analysis (PCA) of dihedral angles, combined with k-means clustering of the associated principle component space. Very briefly, the implementation follows closely the details given in Ref. 91, and ideal cluster sizes are iteratively determined through the convergence of the ratio (SSR/SST) of the sum of squares regression (SSR) and the total sum of squares (SST), which is equal to the coefficient of determination in linear regression. By default, the clustering terminates when the SSR/SST ratio exceeds a value of 0.9. Internal coordinates for the identification of suitable dihedral angles are set up in reference to the topology measure mentioned earlier to optimally represent covalent bonds. A PCA is then conducted to provide the principle components associated with the most diverse dihedral angles from this list, excluding terminal rotamers such as methyl groups. The clustering procedure can automatically be appended to the CREGEN algorithm and is used, for example, within the configurational entropy calculation (Sec. II A 4).

3. Specialized conformational sampling applications

Opposed to heuristics-based conformer generators, general and physics-based conformational sampling in CREST offers enhanced simulation capabilities with the option to define a series of geometrical constraints (see Sec. III A 5). The latter are of particular use for sampling non-covalently bound complexes. Here, an ellipsoidal potential can be used to encapsulate the studied complex and avoid problems due to the metadynamics bias which, by maximization of the RMSD due to increasing the distance between the non-covalently bound fragments, would minimize the energy and hence only generate dissociated clusters. A dedicated non-covalent interaction (NCI) runtype, called NCI-MTD, is implemented in CREST to automatically set up the ellipsoid potential and slightly modify the bias potential parameters. Several examples of this were shown in the original publication,[19] in comparison with DFT calculations and experimental references from rotational spectroscopy.

Another application is conformational sampling with constrained parts of the structure, for example, to find preferred conformations on a surface cutout. In this context, and as opposed to previous program versions, CREST 3 implements the complete freezing of atoms in all metadynamics simulations and geometry optimizations. An illustrative example combining the NCI-MTD mode with this new capability is provided in Fig. 7, which shows a cutout of the iron-based MIL53 metal organic framework (MOF) used for the drug delivery of Ibuprofen.[92,93] In the corresponding calculation, all iron atoms in the cutout were frozen to closely retain the crystal structure[92] geometry. Furthermore, the NCI-MTD ellipsoid potential encapsulating the entire cavity (not visualized in the figure) was applied only to the ibuprofen molecule to avoid its effusion from the MOF. With a calculation runtime of ~38 min on a standard commercial laptop, 122 unique conformers for the complex were found within an energy window of 6 kcal mol^{-1} at the GFN-FF level, of

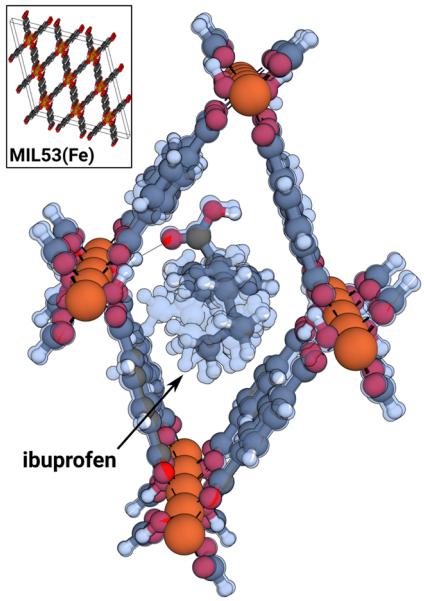

FIG. 7. Overlay of the ten most preferred conformations of the Ibuprofen molecule, embedded within one cell cutout of the MIL53(Fe) MOF (Cambridge Crystal Structure Database code HAMXEC),[92] calculated at GFN-FF level. The most stable conformation is shown in solid color; all other conformations are shown in transparent blue.

which the lowest ten are shown overlaid in Fig. 7. These conformers are particularly useful for calculating the more accurate adsorption energies of the drug-in-MOF complex. Even at the inexpensive force-field level, the lowest energy conformer was found to be stabilized by 3.8 kcal mol^{-1} compared to the initial optimized starting structure and showed a strong preference for forming a hydrogen bond to the MOF. This example clearly demonstrates the versatility and efficacy of both CREST and GFN-FF. The machinery for constrained conformational sampling is easy to set up and provides a powerful toolset for a variety of applications.[42,43,94,95] New applications, for example, conformational sampling with applied hydrostatic pressure via the extended hydrostatic compression force-field (X-HCFF) method[96] (see Sec. III A 3), are currently explored by our groups.

4. Absolute molecular entropy calculation

As an extension to the previously discussed iMTD-GC workflow, an algorithm for the calculation of absolute molecular entropy was introduced in Ref. 40. For high accuracy calculations of thermodynamic quantities, such as Gibbs free energy, precise knowledge of the systems entropy is required. The absolute molecular entropy (S°) calculation in CREST is based on the widely adopted quasi-harmonic analysis (QHA), which was pioneered by Karplus *et al.*[97–99] In general, the partition function for a canonical ensemble with independent degrees of freedom allows the separation of S° into individual contributions,

$$S^\circ = S_{\mathrm{elec}} + S_{\mathrm{trans}} + S_{\mathrm{rot}} + S_{\mathrm{vib}} + S_{\mathrm{conf}}, \tag{1}$$

where S_{elec} is the electronic entropy due to different electronic states, S_{trans}, S_{rot}, and S_{vib} are the translational, rotational, and vibrational entropy contributions, respectively, and S_{conf} is the conformational entropy. The latter can be understood as part of the vibrational entropy at the global minimum due to the "mixing" of other conformational states.[40,99–101] This separation of S_{conf} and S_{vib} motivates the use of a partition within QHA that can generally be given by

$$S^\circ \approx \sum_i^{N_{\mathrm{conf}}} p_i S^{\mathrm{trv}}_{\mathrm{approx},i} + S_{\mathrm{conf}}, \tag{2}$$

where p_i corresponds to the Boltzmann population for structure i, calculated as

$$p_i = \frac{g_i e^{-\beta E_i}}{\sum_j g_j e^{-\beta E_j}}. \tag{3}$$

The first right-hand side term in Eq. (2) is a weighted sum over all translational, rotational, and vibrational components of the entropy that is typically calculated within a rigid-rotor harmonic-oscillator (RRHO) approximation.[102] The second term, containing the—purely informational—*configurational* entropy,[99] is given by the Gibbs–Shannon equation

$$S_{\mathrm{conf}} = -R \sum_i^{N_{\mathrm{conf}}} p_i \ln(p_i). \tag{4}$$

In practice, the calculation of S° via Eq. (2) faces two main problems: First, the exact partition of S_{vib} and S_{conf} is unknown and depends on the level of theory. Second, obtaining a properly converged S_{conf} and overcoming broken ergodicity is a severe technical challenge. Since S_{conf} is a logarithmic quantity, the extraordinarily large number of high energy conformers for flexible molecules may still significantly contribute to the entropy, which makes the required sampling of configurations computationally unfeasible. Promising recent developments for the calculation of S° that nonetheless approach this sampling problem either follow the QHA[40,103] or employ so-called mutual information expansions.[104–106]

In our work,[40,41] we proposed a modification of iMTD-GC aimed at more thorough sampling, explicitly for obtaining the conformational entropy. This workflow is primarily aimed at generating complete ensembles rather than finding the global minimum and ideally should be started from the latter.[40] The respective workflow is outlined in Fig. 8. The core novelty compared to the default iMTD-GC is the replacement of the MD and GC steps after an initial block of metadynamics simulations by a batch of umbrella sampling[107] dynamics using the same type of RMSD bias potentials as introduced for the MTDs. This will effectively forbid the simulations to visit certain (known) regions of the energy landscape. However, instead of modifying the list of bias structures during the simulations as in the MTD, a fixed set of representative bias configurations is chosen in an extension to CREGEN via a PCA of intramolecular dihedral angles combined with k-means clustering.[91] Due to the RMSD-based potential and the fixed nature of the list of reference configurations, we referred to the umbrella sampling approach as static metadynamics (sMTD) and the whole workflow as iMTD-sMTD. The block of umbrella sampling simulations is repeated iteratively until convergence with regards to a conformational entropy calculated via the Gibbs–Shannon equation is reached for the CRE. In general, this workflow is more thorough than iMTD-GC and can routinely be applied for CSE, but it is also more costly. At this point, we again note that the number of generated conformers within a given energy cutoff is much smaller than the theoretically expected overall number of minima, for example, according to the 3^n approximation (see above).[83,84] While this is certainly beneficial for sampling the corresponding ensembles as completely as possible, it warrants consideration of the energy window, which must be large enough to capture the cumulative configurational entropy contributions of high-energy conformations. We generally find 6–10 kcal mol^{-1} sufficient for the energy window, depending on the molecule size.

The calculation of absolute molecular entropy with iMTD-sMTD requires further (automated) post-processing. First, the average translational, rotational, and vibrational contributions [weighted sum in Eq. (2)] are calculated using a modified and scaled RRHO approach (msRRHO), which refers to the application of a scaling factor for high and a harmonic-oscillator/free rotor interpolation for low frequency modes.[40,108] For convenience and efficiency reasons, this term is calculated at an SQM or force-field level of theory and shifted by the value of the lowest-energy conformer to obtain an additive contribution that can be added directly to a more accurate msRRHO calculation for the lowest conformer at the DFT level,

$$\bar{S}_{\mathrm{msRRHO}} = S^{\mathrm{DFT}}_{\mathrm{msRRHO},0} + \sum_i^{N_{\mathrm{conf}}} p_i S^{\mathrm{SQM}}_{\mathrm{msRRHO},i} - S^{\mathrm{SQM}}_{\mathrm{msRRHO},0}. \tag{5}$$

Second, since Eq. (4) will provide an exact entropy only in the unachievable limit of infinite sampling, we extrapolate to ensemble

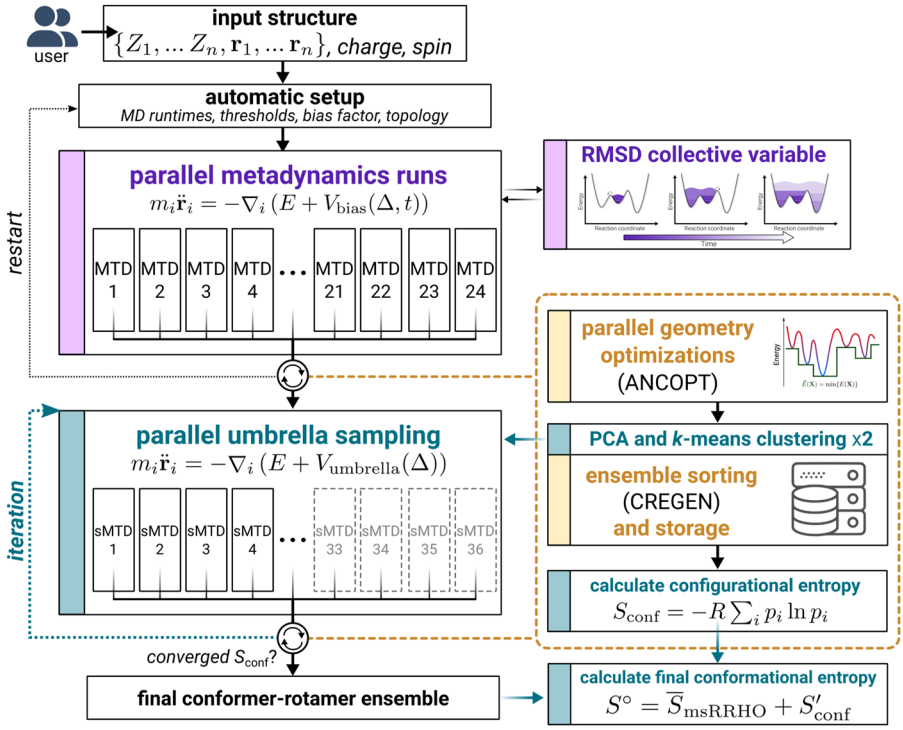

FIG. 8. Outline of the iterative metadynamics–static metadynamics (iMTD-sMTD) workflow used to calculate the configurational entropy.

completeness by using the configurational entropy documented during each of the iterations of iMTD-sMTD via

$$S'_{\mathrm{conf}} = \lim_{x\to\infty}\left(S_{\mathrm{conf}}(x_0) + \lambda_1\left(1 - \exp\left(\lambda_2\, x^{\lambda_3}\right)\right)\right), \tag{6}$$

where x refers to the sMTD iteration cycle, with x_0 being the starting point after the initial MTD block. The parameters λ_{1-3} are automatically fitted via a Levenberg–Marquardt[109,110] algorithm using the Gibbs–Shannon configurational entropies obtained for each iMTD-sMTD iteration x (cf. Fig. 8). While it was argued that a simple extrapolation formula might not reproduce the entropy variation in conformational flexibility across different molecules,[106] we note that this is not the purpose of this term. Rather, it captures the growth of entropy due to a sequential, but systematic, completion of the ensemble by a given search algorithm. If the latter steadily draws new conformations from all regions of the energy landscape and overcomes broken ergodicity, the associated growth in entropy will closely follow Eq. (6). Furthermore, the overall contribution of entropy due to the extrapolation is small and does not exceed more than a few entropy units, for example, ~1 cal mol^{-1} K^{-1} for the extremely flexible octadecane molecule, as seen in Fig. 9(a). However, we observed a significant improvement in the iMTD-sMTD workflows numerical stability and found the extrapolation scheme capable of effectively tackling the finite-sampling issue.[40] The conformational entropy approach was thoroughly benchmarked with regards to both its technical performance and the precision of the predicted absolute molecular entropy in Ref. 40. For the latter, a benchmark set containing experimentally observed absolute molecular entropies of flexible molecules was composed from the work of the Head-Gordon group[112] and extended with references for other systems, including long n-alkanes. A minimal visual summary of our findings is presented in Fig. 9(b). The comparison with the experiment reveals a general underestimation of the absolute molecular entropy if conformational contributions are neglected, exceeding a chemical accuracy of ±3 cal mol^{-1} K^{-1}, especially for larger molecules with a high number of DOF. For the most flexible molecules in the employed test set, errors easily reach 30 cal mol^{-1} K^{-1} (9 kcal mol^{-1} at room temperature), which signals the importance of S_{conf} as a potential accuracy factor. The iMTD-sMTD approach implemented in CREST is able to routinely capture these effects for drug-sized molecules, providing exceptionally low errors of just ~1 cal mol^{-1} K^{-1} for both mean absolute deviation (MAD) and RMSE when coupled to a msRRHO DFT calculation for the lowest conformer.

B. Explicit solvation via quantum cluster growth

Beyond the, so far, presented functionality, CREST is also able to model explicit and micro-solvation. For this purpose, the Quantum Cluster Growth (QCG) algorithm[42] was recently implemented, allowing the automated generation of cluster ensembles containing a solute solvated with explicit solvent molecules. Similar strategies have been published by other groups.[113–115] By using the GFNn-xTB methods, any element up to radon can be included in the calculation, making the solvation of chemically diverse solutes with almost any conceivable solvent, including organic solvents just like ionic liquids, feasible. The resulting conformers of solute–solvent clusters can

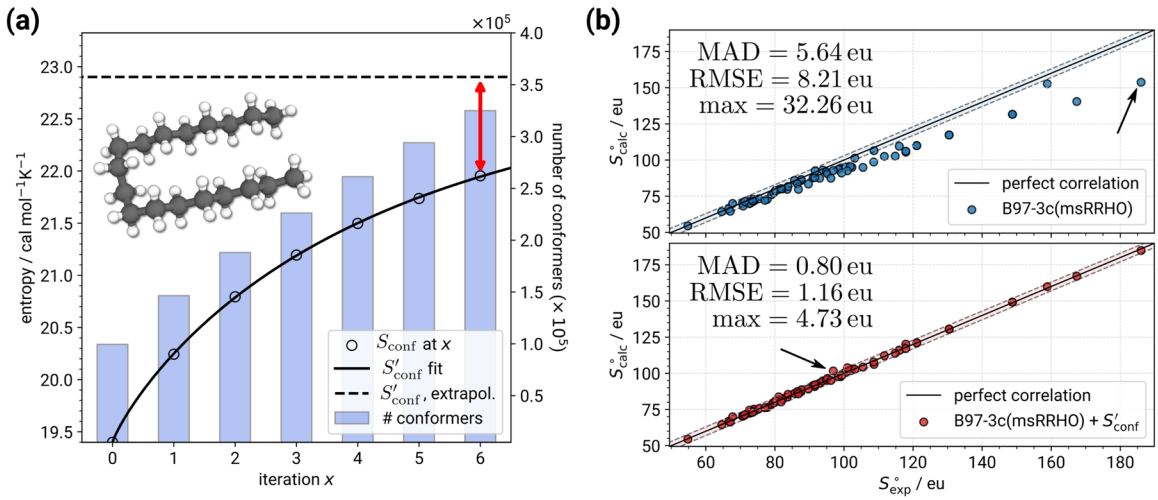

FIG. 9. (a) Extrapolation of S'_{conf} for the example of *n*-octadecane. The red double sided arrow shows the final correction to the conformational entropy, which in this case accounts for ~1 cal mol^{-1} K^{-1} (4.5% of the conformational entropy). (b) Parity comparison of calculated and experimental absolute molecular entropy employing only msRRHO at the B97-3c level of theory[111] for the most stable conformer (top, blue) and the same calculation corrected by S'_{conf} obtained with CREST and GFN2-xTB (bottom, red). "eu" refers to entropy units of cal mol^{-1} K^{-1}. The shaded area around the parity line refers to the domain of chemical accuracy (±3 eu). The black arrows show the case with the highest deviation from the experimental reference. All data shown in these figures were taken from Ref. 40, which also includes more detailed discussions.

directly be used to compute the properties of the solute in the liquid phase. Additionally, the QCG workflow is able to compute solvation-free energies with such ensembles. The QCG algorithm, outlined in Fig. 10, contains two essential steps that can be subsequently invoked.

The "Growth" part yields a single, low-energy solute–solvent cluster, while the subsequent ensemble generation generates different conformers from this cluster. To get clusters of equally distributed solvent molecules around the solute, two wall potentials are used in both steps. The inner wall potential, affecting only the solute, serves to keep the solute in the center of the cluster, while the outer one, applied to the whole cluster, suppresses the clustering of the solvent molecules to ensure an evenly distributed solvent shell. As the size of the wall potentials is crucial to accurately reflecting the real behavior in solution, it is carefully constructed using geometrical features of the solute and solvent, the solvent accessible surface area (SASA) of the solute, and the number of solvent molecules allowing a dynamical adjustment during the growth (see Ref. 95 for details). The cluster is grown step-wise by adding solvent molecules at low-energy interaction sites. Here, either a combination of the `xtbiff` program[116] and GFN*n*-xTB/GFN-FF geometry optimizations or the recently developed and for this task recommended aISS docking algorithm[43] implemented in `xtb` can be employed. The Growth algorithm stops if a certain amount of solvent molecules is added that is either determined by the user or automatically based on converging interaction energies between the solute and solvent, determined by a moving average [cf. Fig. 11(b)]. Thereby, the algorithm yields a solute–solvent cluster along with some properties like energy convergence and the final wall potentials for direct use with the `xtb` program. For the second step that generates an ensemble from this structure, the user has the choice of performing a single MD, MTD, or NCI-MTD workflow (see Sec. II A 3) for the conformer search with the wall potentials. Finally, the whole ensemble is optimized without the wall potentials and energetically ranked, including information about the Boltzmann weighting. Additionally, CREST can compute solvation free energies from this ensemble. Therefore, a reference cluster of pure solvent molecules is generated. By subtracting the Boltzmann-weighted free energies of the solvent cluster and the single solute molecule in the gas phase from the solute–solvent cluster, the energy for transferring the solute from the gas phase into the solution is obtained, which corresponds to the solvation free energy. An example of solvating the anticoagulant drug Rivaroxaban[117] with 100 water molecules using the QCG algorithm is given in Fig. 11. Additionally, the interaction energy of the solute and solvent shell ($E_{int}^{solu-solv}$) is shown during cluster growth.

When adding about 40 water molecules, the first solvent shell of Rivaroxaban is completed, which can be seen in a stagnating decrease of $E_{int}^{solu-solv}$. Afterward, the additional water molecules lead to small variations in the interaction energy as the additional water molecules influence the ones in the first water shell.

C. Protonation site sampling

With the ability to perform inexpensive electronic structure calculations at SQM levels of theory comes access to a variety of properties that cannot be obtained with a classical force-field. For instance, molecular orbitals (MOs) are readily obtained and can be used to analyze and describe the molecule. A popular use of MOs from SQM is to employ them as an orbital guess in higher-level electronic structure calculations.[118]

Localized molecular orbitals (LMOs) from GFN*n*-xTB can be used to efficiently identify possible protonation sites of a molecule as part of the CSE.[19,44] The obvious advantage of this procedure over heuristics-based or ML algorithms is its generality in that any chemical system can be treated as long as MOs can be supplied to

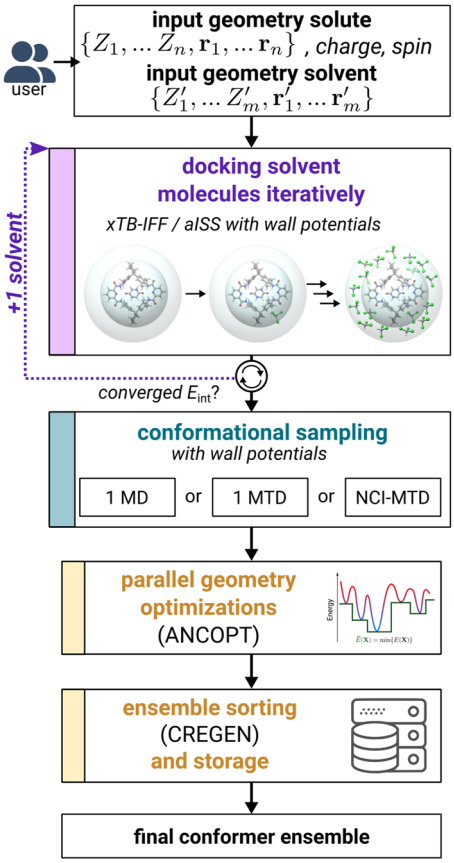

FIG. 10. Schematic illustration of the QCG workflow.

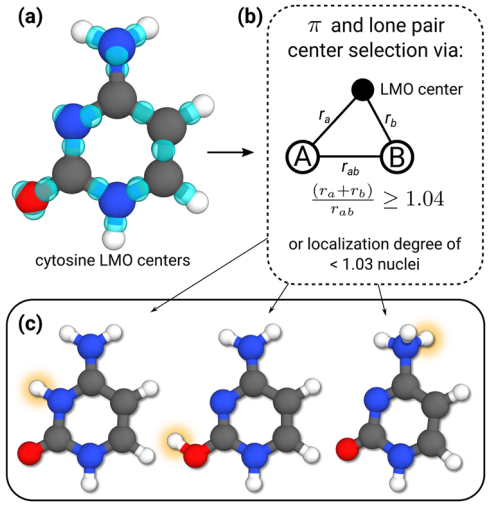

FIG. 12. (a) LMO centers of the cytosine molecule, represented by transparent blue spheres. (b) Conditions for identifying suitable π and lone pair centers for protonation. (c) Low-energy protomers of cytosine, calculated at the GFN2-xTB/ALPB(toluene) level.

the algorithm. Practically, the procedure is interfaced to the GFNn-xTB methods. The general protonation site screening procedure in CREST is outlined in Fig. 12 for the cytosine molecule. Starting from a calculation of LMOs via the `xtb` program,[53] suitable lone pairs and π-centers are identified. The two criteria employed for this are either a localization degree of less than 1.03 nuclei, i.e., the LMO belongs to a single atom, or a triangular inequality greater than 1.04 if the LMO is shared between two atoms [cf. Fig. 12(b)]. After adding a proton to the identified LMO centers and optimizing the corresponding structures, the generated protomers are ranked energetically. Similar procedures based on localization of the electrostatic potential have also been proposed.[119] A typical target size for the protonation site screening procedures is drug molecules. Larger systems are at risk of performing inefficiently, especially when a high degree of conformational flexibility is involved. In such cases, post-processing by conformational sampling is necessary to accurately rank the relative stabilities of the generated isomers.

A similar yet simpler strategy can be adapted for identifying deprotonation sites, where all possible deprotonated isomers can be generated by systematically removing the corresponding hydrogen atoms and optimizing the molecular structure.[19] Combining both protonation and deprotonation site algorithms sequentially, CREST enables the calculation of prototropic tautomers. Here, the total number of technically accessible tautomers, $n \times m$, is predetermined by the number of hydrogen atoms in the system n and accessible protonation sites m calculated from the LMOs. Furthermore, the extension to alkali ion addition in both the protonation and tautomerization screening algorithms is straightforward. The

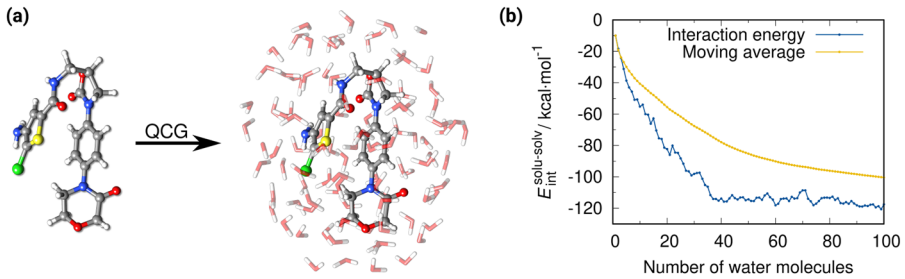

FIG. 11. (a) Solvation of Rivaroxaban with 100 water molecules employing the QCG algorithm. (b) Interaction energy of the solute and the solvent shell during cluster growth.

TABLE III. Geometrical constraint potentials implemented in the CREST program.

Constraint	Potential
Distance/bond	$V(r_{ij}) = \frac{1}{2}k_r(r_{ij} - r_0)^2$
Angle	$V(\theta_{ijk}) = \frac{1}{2}k_\theta\left(\frac{\mathbf{r}_{ij}\cdot\mathbf{r}_{kj}}{\lvert\mathbf{r}_{ij}\rVert\mathbf{r}_{kj}\rvert} - \theta_0\right)^2$
Dihedral angle	$V(\phi_{ijkl}) = \frac{1}{2}k_\phi\left(\frac{\mathbf{n}_{ijk}\cdot\mathbf{n}_{jkl}}{\lvert\mathbf{n}_{ijk}\rVert\mathbf{n}_{jkl}\rvert} - \phi_0\right)^2$
Bond range	$V(r_{ij}) = k_B T \ln\left((1 + e^{\beta(r_{ij} - r_0^{\text{upper}})})(1 + e^{\beta(r_0^{\text{lower}} - r_{ij})})\right)$
Ellipsoid potential	$V(\mathbf{r}) = \sum_i k_B T \ln\left(1 + \exp\left(\beta(\lvert\mathbf{r}_i - \mathbf{O}\rvert - r_0)\right)\right)$

automated protonation and tautomerization procedures of CREST were efficiently applied in the calculation of pK_a values via linear free energy relationships.[16,120,121] One significant limitation of these procedures should be highlighted: the manipulation of molecules, either by adding or removing an atom directly from the input structure, ignores the naturally occurring reaction barriers. This oversight can sometimes result in atypical protonation or deprotonation patterns,[122] and users are urged to carefully analyze the resulting ensembles with informed chemical insight. Enhancements to these algorithms are planned for future development.

D. Automated reaction discovery for mass spectra elucidation

Automatic reaction discovery has become an important field of modern computational chemistry.[123] The MSREACT tool of CREST aids reaction exploration with a special focus on creating fragments and rearrangement products (isomers) occurring in mass spectrometry (MS) experiments, such as electron ionization (EI) or Electrospray Ionization/Collision Induced Dissociation (ESI/CID). Systematic generation of possible products for the input molecule is achieved through the application of constraining potentials between atom pairs and optimization using an efficient SQM method. Interatomic distances are deliberately extended well beyond their equilibrium values, inducing dissociation and rearrangement reactions. As a default setting, atom pairs separated by up to three covalent bonds are subjected to elongation through a repulsive harmonic potential, as given in Table III. Hereby, r_0 is set to 1.5 times the sum of the covalent radii of both atoms plus the number of bonds in between, with k_r set to 0.05 E_h/bohr^2. The number of bonds between two atoms is identified from the molecular graph or topology (see Sec. II A 2) via the Floyd–Warshall algorithm.[124] Subsequent optimization at "crude" energy convergence settings (see Sec. III B 2) is performed with GFNn-xTB at an elevated finite electronic temperature of 5000 K to favor the generation of open-shell radicals and to partially account for the multi-reference nature of the open-shell systems typically occurring in (EI-)MS.[125] To produce often observed products due to hydrogen rearrangements, further optimizations are conducted with attractive potentials (k_r set to −0.05 E_h/bohr^2) applied between hydrogen atoms and potential protonation sites within a certain cutoff distance (by default 4 Å), which are determined by LMOs from GFNn-xTB as described in Sec. II C. For planar molecules, for which simple bond stretching may not be enough to get all important isomers, additionally, atomic displacements of the atom positions and subsequent optimizations can be performed. Duplicates of the resulting fragment-isomer ensemble showing the same topology are removed with the quantum chemistry-inspired molecular identifier *MolBar*.[126] Next, the remaining structures are sorted with the CREGEN algorithm to remove potential remaining duplicates at the employed GFNn-xTB level and to remove unreasonable structures above a given energy threshold (default: 200 kcal mol^{-1}). Fragmentation of a molecule is detected if the distance between all atoms of two fragments is 1.3

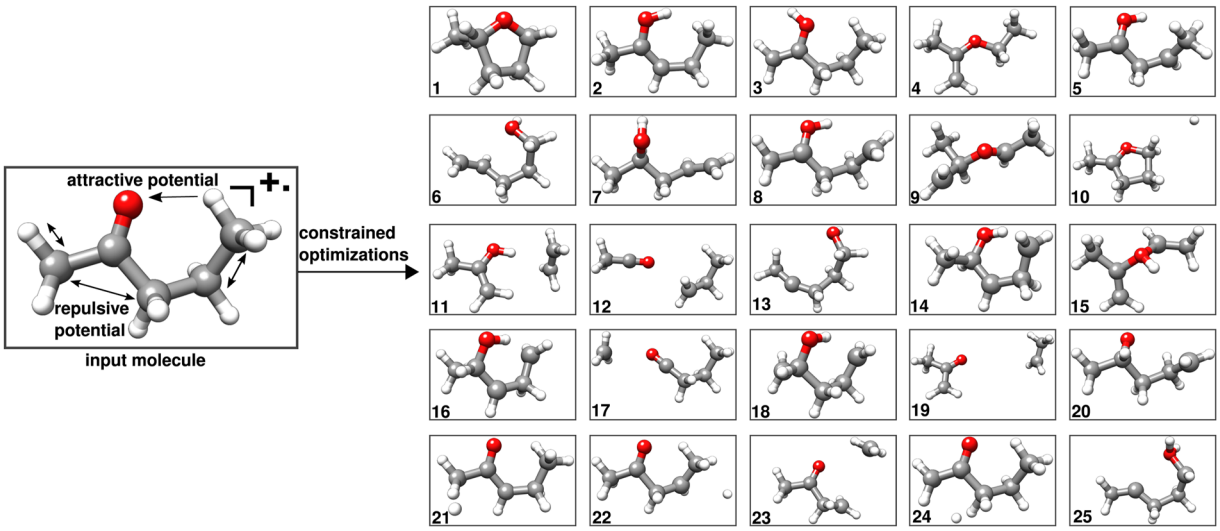

FIG. 13. Generated unique fragments of positively charged 2-pentanone at the GFN2-xTB level using the MSREACT mode with default settings. The structures are numerically labeled based on their relative energies.

times larger than the sum of the covalent radii of the respective atoms.

As an example, Fig. 13 shows the isomers and fragments obtained for 2-pentanone with default settings in the EI-mode, i.e., positively charged open shell radical, using GFN2-xTB. A total of 81 optimizations resulted in 25 unique isomers and fragment pairs. The generated products depend on the input conformer, as specific rearrangements may require spatially close atoms, as in, for example, the important fragment **8** stemming from a McLafferty type rearrangement. Here, the lowest conformer found by CREST at the GFN2-xTB level was used as the input structure.

Subsequent fragmentation via cascade reactions can be obtained by employing the MSREACT mode iteratively on the generated fragments. The structures can be used as a starting point for automated reaction network exploration by combining them with automated transition state search to predict the relative occurrence of the fragments. An automated workflow employing this approach to compute mass spectra is currently being investigated. By combining the protonation sampling tool and using the MSREACT mode for the generated protomers or deprotomers, molecular species relevant to ESI/CID-MS can also be generated.

E. Minimum energy crossing points

A recent addition to the CREST program is the adaptation of sampling workflows like iMTD-GC in context of other points of interest on the potential energy landscapes, for example non-adiabatic minimum energy crossing points (MECPs).[47,48] MECPs serve as a good approximation to minimum energy conical intersections (MECIs).[127,128] The idea here is to locate the MECP on an artificial seam PES given by

$$E_{\mathrm{tot}} = E_{\mathrm{avg}} + V_{\mathrm{gap}}, \tag{7}$$

combining two core components E_{avg} and a penalty function V_{gap}, in a so-called "derivative coupling vector-free" treatment.[129,130] Considering a crossing point between n electronic states of the molecule, the seam PES is based on the arithmetic mean,

$$E_{\mathrm{avg}} = \frac{1}{n}\sum_{i}^{n} E_i. \tag{8}$$

This energy average is combined with a pairwise penalty

$$V_{\mathrm{gap}} = \sum_{i}^{n}\sum_{j>i}^{n} \sigma\left(e^{-|\Delta E_{\mathrm{ij}}|} + k\right)\frac{\Delta E_{ij}^2}{|\Delta E_{ij}| + \alpha}, \tag{9}$$

minimizing the energy gap $|\Delta E_{ij}|$ between two diabatic states i and j. The latter potential confines the overall energy to a lower dimensional hyperline where the states are degenerate.

Approaches to exploring the MECI space by locating MECPs via a metadynamics-based strategy have recently been presented by the groups of Lindner *et al.*[131] and Pieri *et al.*[132] CREST 3 allows the specification of a gap constraint Eq. (9) and the application of Eq. (8) in the previously presented workflows, like iMTD-GC. Employing this in combination with the GFNn-xTB level of theory enables rapid screening of MECPs, as we have previously demonstrated.[47,48] A common example is the S_0/S_1 MECIs (S_0/T_1 MECPs) of benzene,[132] which are shown in Fig. 14(a).

Furthermore, employing interfaces to quantum chemistry programs like ORCA[134,135] (see Sec. III A 4), CREST provides capabilities to post-process the crossing points obtained with GFNn-xTB at higher DFT levels with the same technical machinery. Figure 14(b) demonstrates the latter for optimizing the S_0/T_1 MECP of azobenzene, using the broken-symmetry unrestricted Kohn–Sham (BS-UKS) PBEh-3c level of theory[136] to describe the S_0 and "regular" UKS PBEh-3c for the triplet state. These new capabilities of the

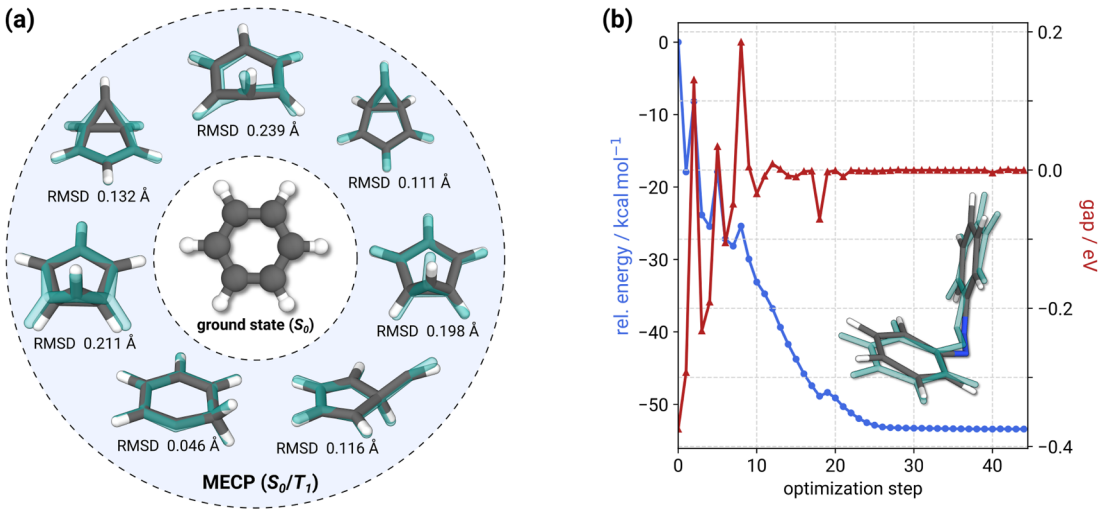

FIG. 14. (a) Minimum energy crossing point geometries found for the S_0/S_1 crossing (S_0/T_1 for xTB) of benzene, via a modified GFN0-xTB-based algorithm.[48] Reference *hh*TDA-BHLYP-D3(BJ)/def2-SV(P)[133] geometries are shown in transparent cyan, and the corresponding Cartesian RMSD is given below each structure. (b) Total energy and energy gap change of the S_0/T_1 MECP optimization of azobenzene, calculated at the PBEh-3c level of theory. The corresponding structure is shown in solid color, overlaid by the respective GFN0-xTB MECP in transparent cyan.

CREST code provide valuable tools, for example, to aid the study of photochemical reactions and have been successfully employed for the latter.[137]

III. IMPLEMENTATION DETAILS

A. Calculator and implemented potentials

By default, CREST comes with implementations of the GFN*n*-xTB Hamiltonians[39,54–56] and the GFN-FF force-field.[57] Since these methods are parametrized for all elements up to radon ($Z \leq 86$), a majority of the chemical space is accessible with the implemented algorithms discussed in Sec. II. From a computational viewpoint, the heart of the program is the calculator instance interfacing the different Hamiltonians, geometric constraint potentials, and storing parameters (cf. Fig. 15).

While general program usage is still possible via command line arguments, as in previous versions of CREST, the increasing complexity of implementations like the calculation container demands the option to define detailed calculation setups. For this purpose, CREST 3.0 implements a new TOML input reader based on the open-source `toml-f` project.[138] Listing 1 shows the new TOML input for a simple GFN1-xTB + ALPB(toluene) singlepoint calculation. More complex calculations, as, for example, required in a MC-ONIOM setup (Sec. III B 5), or for minimum energy crossing point calculations (Sec. II E), support the definition of multiple `[[calculation.level]]` instances that are processed by the calculator.

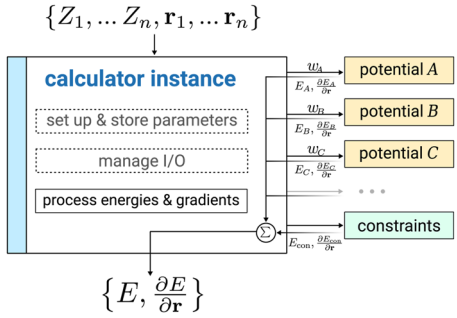

FIG. 15. Outline of the calculator setup in CREST. The calculator processes a molecular structure with n atoms, as defined by its atomic numbers $\mathbf{Z} = \{Z_1, \ldots, Z_n\}$ and Cartesian coordinates $\mathbf{r} = \{\mathbf{r}_1, \ldots, \mathbf{r}_n\}$, and returns an energy E with the corresponding gradient $\partial E/\partial \mathbf{r}$. The setup allows to define multiple potentials, for example, GFN*n*-xTB, and combine the respective energies and gradients via user-defined weights w. Additionally, a variety of structural constraints can be defined.

LISTING 1. Example of an TOML input file for setting up a GFN1-xTB + ALPB(toluene) singlepoint energy evaluation in CREST.

```
# This is a CREST input file
input = 'struc.xyz'
# Runtype
runtype='singlepoint'
# Parallelization
threads = 8
# Calculation setup
[calculation]
type = 1  # specify the use of energy & gradient
          # from the first [calculation.level]
elog="energies.log"
# Calculation level definitions
[[calculation.level]]
method = "gfn1"
chrg = 0
alpb = "toluene"
```

In Subsections III A 1–III A 5, a short overview of the most important interfaces of the CREST calculator instance is given. The calculator is, in turn, interfaced to different modules that contain typical atomic simulation procedures, as discussed in Sec. III B.

1. GFN1-xTB and GFN2-xTB Hamiltonians via `tblite`

In an effort to capture the idea of specialized code and interoperability, the `tblite`[139] project was created to unify various tight-binding methods in one codebase, among which GFN1- and GFN2-xTB.[39,55,56] `tblite` restricts its functionality to the computation of singlepoint energies and their Cartesian gradients, in contrast to a full simulation environment like `xtb`.[53] Interfaces for the programming languages Fortran, C, and Python are provided, which constitutes an advantage over the `xtb` codebase. This Fortran interface for energy and gradient calculations makes the repeated input–output operations, previously necessary[19] for various CREST runtypes, obsolete. This leads to a significant speed-up for several CREST workflows (cf. Sec. II A) and is, therefore, now set as the default when requesting GFN1- or GFN2-xTB as the electronic structure method. Furthermore, the relatively modern features of Fortran 2003 and later standards allow the use of runtime polymorphism in an object-oriented programming (OOP) style. OOP techniques are used heavily throughout the `tblite` code base to allow a high degree of flexibility in the combination of interactions constituting the tight-binding Hamiltonians. The advantage of such a datatype structure can be rationalized by examining the energy expressions of GFN1- and GFN2-xTB in Eqs. (10) and (11), respectively,

$$E^{\mathrm{GFN1}} = E_{\mathrm{rep}} + E_{\mathrm{EHT}} + E^{\mathrm{D3}}_{\mathrm{disp}} + E_{\mathrm{XB}} + \underbrace{E^{(2)}_{\gamma} + E^{(3)}_{\Gamma}}_{E_{\mathrm{Coulomb+XC}}}, \tag{10}$$

and

$$E^{\mathrm{GFN2}} = E_{\mathrm{rep}} + E_{\mathrm{EHT}} + E^{\mathrm{D4}}_{\mathrm{disp}} + \underbrace{E^{(2)}_{\gamma} + E^{(2,3)}_{\mathrm{AES}} + E^{(2,3)}_{\mathrm{AXC}} + E^{(3)}_{\Gamma}}_{E_{\mathrm{Coulomb+XC}}}. \tag{11}$$

Most energy contributions appear in both expressions, like the repulsion energy E_{rep}, the extended-Hückel-type energy E_{EHT}, the dispersion energy E_{disp}, and the Coulomb and exhange-correlation (XC) energies $E_{\mathrm{Coulomb+XC}}$. The exact formulation of these interactions and the energy expressions that follow from them might differ by method. Besides the two-body $E^{(2)}_{\gamma}$ and three-body $E^{(3)}_{\Gamma}$ Coulomb interactions, GFN2-xTB also accounts for anisotropic effects through E_{AES} and E_{AXC}, for example. However, this general reformulation in terms of polymorphic interaction containers allows for a high degree of code reusability. Additional interactions, like the halogen bond correction E_{XB}, can be added to a calculator instance and can also be manipulated through the interface. The implicit solvation model ALPB[140] is also implemented as an interaction container in `tblite` and can be added to CREST calculations, as seen in Listing 1. The integration of `tblite` is performed via meson subprojects and git-subprojects.

2. GFN0-xTB and GFN-FF

Following a similar design strategy as `tblite`, the GFN0-xTB Hamiltonian[54] and the GFN-FF method[57] were adapted from the `xtb` code[53] to provide thread-safe standalone calculators. Both projects are hosted on GitHub[141,142] and provide build options via the meson and CMake build systems. An implementation into `tblite` was not carried out because the methods differ conceptually from GFN1- and GFN2-xTB. GFN0-xTB, a non-self-consistent variant of the GFNn-xTB Hamiltonians, calculates the total energy as

$$E^{\mathrm{GFN0}} = E_{\mathrm{rep}} + E_{\mathrm{EHT}} + E^{\mathrm{D4}}_{\mathrm{disp}} + E_{\mathrm{EEQ}} + E_{\mathrm{srb}}, \tag{12}$$

which resembles the GFN1- and GFN2-xTB energy expression. However, E_{EHT} is obtained by a single diagonalization of the Hückel-type Hamiltonian, a classical short range bond correction E_{srb} is employed, and the self-consistent charge electrostatics are replaced by a charge equilibration term E_{EEQ}. The latter is also employed in the GFN-FF method, which calculates the energy via

$$E^{\mathrm{GFNFF}} = E_{\mathrm{bond}} + E_{\mathrm{angle}} + E_{\mathrm{dihedral}} + E_{\mathrm{rep}} + E^{\mathrm{D4}}_{\mathrm{disp}} + E_{\mathrm{EEQ}} + E_{\mathrm{XB}} + E_{\mathrm{HB}}, \tag{13}$$

where, alongside typical bond, angle, dihedral, and repulsion terms, E_{XB} and E_{HB} denote halogen and hydrogen bond contributions. GFN-FF is considered a partially polarized force-field following the same design principles as GFNn-xTB.[39,57]

3. X-HCFF

The extended hydrostatic compression force-field (X-HCFF) method[96] is a mechanochemical model to simulate the effect of hydrostatic pressure on molecular geometries. For this purpose, the molecular cavity is approximated by fused spheres, centered at each atom i, of the respective van der Waals radius.[143] The resulting cavity surface is discretized using a Lebedev–Laikov grid.[144] Pressure is then simulated by projecting compression forces from each tessellation point j to the corresponding atom using the classical definition of pressure, $P = F/A$, leading to

$$\boldsymbol{F}_i = \sum_{j}^{N_{\mathrm{Tess}}(i)} -P \cdot A_j \cdot \boldsymbol{n}_j, \tag{14}$$

where F_i is the overall compression force projected onto atom i, P is the pressure, A_j is the area of j, and $\boldsymbol{n}_j$ is the surface normal pointing toward the cavity. The resulting forces are added to the nuclear gradient, leading to a distortion of the pressure-free equilibrium geometry.

To be used within CREST, X-HCFF was implemented, similar to `tblite`, in the external standalone library `xhcfflib`, calculating the X-HCFF gradient contribution via a singlepoint call. The cavity discretization is carried out following the tessellation scheme used by ALPB.[140] The details of the implementation of `xhcfflib` together with a comprehensive overview of application examples will be given elsewhere; however, in the following, we show the decrease in volume of small methane clusters during compression as a first example application of the use of GFN tight binding methods together with X-HCFF. The respective volumes at a fixed pressure were calculated as the average volumes during a 30 ps MD simulation using GFN2-xTB/X-HCFF for methane clusters consisting of 12, 24, 36, and 64 molecules at 300 K. It should be noted that the pressure is fixed during the X-HCFF calculation while the cavity volume changes with interatomic distances; therefore, the simulation resembles an *NPT* ensemble. As can be seen in Fig. 16, the cluster compressibility

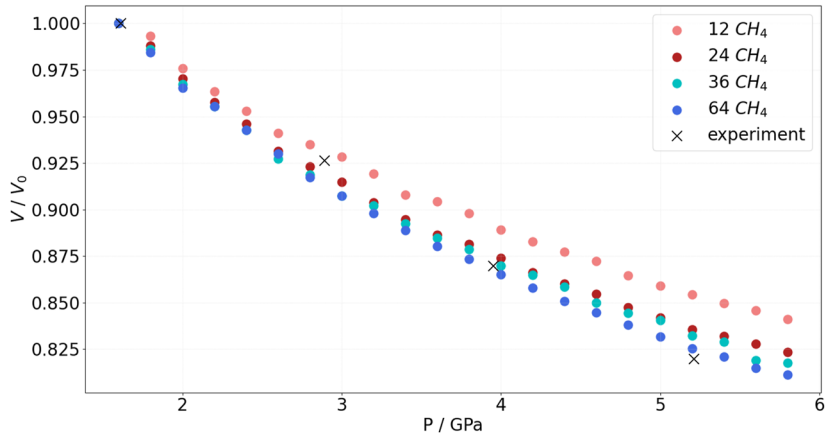

FIG. 16. Comparison of the compression of methane clusters using X-HCFF and GFN2-xTB with experimental[145] values. V_0 corresponds to the volume at 1.6 GPa.

converges rapidly with increasing cluster size, and the model is able to reproduce experimental XRD data obtained with a diamond-anvil cell[145] quite well.

4. Subprocesses

Subprocesses allow the creation of interfaces to other atomic simulation programs. Essentially, *any* potential can be interfaced to the calculator instance as long as energies and gradients are written in a readable format (e.g., a file containing the energy and $3N_{\mathrm{at}}$ entries for the gradient in atomic units). By default, CREST provides a subprocess interface to the ORCA program[134,135] as well as fallback subprocess calculators to the original `xtb` code.[53] Furthermore, an interface to generic shell scripts is implemented. The corresponding calculations can be defined via `[[calculation.level]]` instances that are processed by the calculator, as shown in Listing 2 for the ORCA code.

Subprocess calls need to be treated with caution. For rapid energy and gradient evaluations, for example, via `xtb`, hosting the subprocess creates computational overhead and should generally be avoided in procedures with a high number of calculator evaluations, in particular conformational sampling.

5. Constraints

A series of geometrical constraints are available in CREST to enable targeted sampling (see Sec. II A 3). The corresponding potentials allow user-defined fixing of substructure parts and are listed in Table III. Interatomic distances r_{ij}, angles θ_{ijk}, and dihedral angles ϕ_{ijkl} are constrained via a simple harmonic potential that, aside from the selection of atoms i,j,k,l, only needs the definition of a reference value and a force constant k, which can be individually defined for each set of constrained atoms. The bond range constraint allows the definition of an upper and lower distance limit for a pair of atoms and employs a "logfermi" type potential. For the latter, $k_{\mathrm{B}}T$ has the function of a force constant, and the parameter β is used to control the exponential strength. The ellipsoid potential, which is also of the logfermi type, is used to construct a confining ellipsoid around the entire system (relative to the coordinate origin O), which is particularly useful to avoid dissociation of

LISTING 2. `[[calculation.level]]` setup for a subprocess call to the ORCA program.

```
# ORCA subprocess setup
[[calculation.level]]
chrg = 0
multiplicity = 1
orca_template = 'ORCA.in' # ORCA input file template
orca_cmd = '/full/path/to/orca'
```

non-covalently bound complexes in CREST's NCI mode (see Sec. II A 3). Additionally, the complete freezing of atomic coordinates was implemented into CREST 3.0. The corresponding DOF of the gradient is set to zero and projected out during geometry optimization and molecular dynamics simulations.

B. Implemented atomic simulation methods

1. Molecular and metadynamics

MD is the main tool for the generation of new molecular configurations within the CSE of CREST. In this context, no properties or observables, aside from the molecular geometries, are evaluated from the MD trajectories. However, by basing the CSE on the Newtonian equations of motion, the program promotes sampling based on physical principles rather than a heuristic or chemoinformatic generation of configurations.[19,20] The main advantage of this setup is that algorithms are, in principle, universally applicable to any atomistic system and level of theory, limited only by the available potentials implemented in the calculation container and the associated computational cost.

CREST employs a leap-frog (Verlet-type) algorithm[146] with a simple Berendsen-type thermostat[147] for *NVT* molecular and metadynamics simulations. Longer time steps for extended simulations are enabled by two features: first, an implementation of a SHAKE algorithm[148] introducing a bond constraint into the equations of motion, and second, a modification (increase) of atomic masses, in particular hydrogen, for the calculation of the MD velocities.[149]

Several excellent reviews of metadynamics as a tool for rare event sampling can be found in the literature.[150–152] The metadynamics implemented in CREST is based on the Cartesian RMSD of atomic positions relative to a reference geometry, which provides a single scalar quantity as a collective variable (CV) from which a bias potential is constructed. Bias potentials based on the RMSD as a path collective variable were first reported in 2007 by Branduardi *et al.*[153] The present implementation is based on more recent efforts targeted explicitly at conformational sampling.[19,50] Each corresponding bias potential $V_{\mathrm{bias},i}$ takes the form of a Gaussian function and depends on the RMSD as collective variables Δ_i for a documented reference structure i. The total bias at time t is finally calculated for a set of bias structures $\mathcal{B}$ as

$$V_{\mathrm{bias}}(\Delta, t) = \sum_{i}^{|\mathcal{B}(t)|} k_i \exp\left(-\alpha_i \Delta_i^2\right), \tag{15}$$

where k_i is an empirical force constant and α_i is the Gaussian function width associated with the ith bias structure. In practice, all force constants and Gaussian widths are chosen the same for a given bias potential rather than choosing them individually for each bias configuration. The Cartesian RMSD Δ_i, after optimal alignment between the current MD snapshot at time t and the bias structure i via a quaternion algorithm,[51] is calculated as

$$\Delta_i = \sqrt{\frac{1}{N_{\mathrm{at}}} \sum_{j}^{N_{\mathrm{at}}} (\delta \mathbf{r}_j)^2}. \tag{16}$$

Here, N_{at} is the number of atoms considered for the collective variable, which can refer to either the full structure or a subsystem, and $\delta \mathbf{r}_j = |\mathbf{r}_j^{(\mathrm{t})} - \mathbf{r}_j^{(\mathrm{i})}|$ denotes the Euclidean distance of atom j in the current and the bias configuration. The list of bias configurations $\mathcal{B}$ is dynamically appended with structure snapshots taken during the MTD simulation, i.e., dependent on t. Accordingly, the dump frequency for taking those snapshots is an empirical parameter that influences performance.

Adding new bias structures to the list $\mathcal{B}$ in Eq. (15) will, over time, "fill" the energy landscape [cf. Fig. 17(a)] and, therefore, ensure the exploration of new regions of the conformational space of the molecule.[50] The addition of a single bias configuration typically leads to an instantaneous increase in total energy and temperature during the MTD, as shown in Fig. 17(b). This change is quickly compensated for by the thermostat. For applications within conformational sampling, the bias parameters k and α are selected to maintain the molecular topology, i.e., small enough to avoid the breaking of covalent bonds or the inversion of stereo centers but still able to cross other high-energy barriers.[19]

2. Geometry optimization

The default optimization engine in CREST is ANCOPT (Approximate Normal Coordinate OPTimizer), which was originally developed as a standalone code[154] and later adapted from the `xtb` software.[53] ANCOPT is based on rational function optimization (RFO), originally introduced by Banerjee *et al.*[155–157] Like many state-of-the-art optimization algorithms,[158] RFO employs the augmented Newton–Raphson step,

$$\Delta \mathbf{r}_k = -(\mathbf{H}_k - \lambda_k \mathbf{I})^{-1} \mathbf{g}_k, \tag{17}$$

where $\Delta \mathbf{r}_k$ is the coordinate update at optimization step k, $\mathbf{H}_k$ and $\mathbf{g}_k$ are the Hessian and gradient, respectively, and the scalar λ_k determines the step length. According to the RFO formalism, λ_k and $\Delta \mathbf{r}_k$ can be obtained from the set of linear equations

$$\begin{pmatrix} \mathbf{H}_k & \mathbf{g}_k \\ \mathbf{g}_k^{\top} & 0 \end{pmatrix} \begin{pmatrix} \Delta \mathbf{r}_k \\ 1 \end{pmatrix} = \lambda_k \begin{pmatrix} \Delta \mathbf{r}_k \\ 1 \end{pmatrix}, \tag{18}$$

where λ_k is chosen for minimizations as the lowest eigenvalue of the augmented Hessian matrix [the left-hand side of Eq. (18)]. Since only the lowest eigenvalue is required, ANCOPT employs the Davidson method for subspace diagonalization[159] to quickly estimate λ_k for systems with a high number of DOF. The RFO step may become very large and "overshoot" the minimum if the augmented Hessian has small eigenvalues and is, therefore, dynamically rescaled to a predefined maximum step length.[157] Further efficiency considerations concern the computational cost of calculating numerical Hessian matrices that scale linearly [$\sim\mathcal{O}(6N)$] with the system size (see Sec. III B 3). Hence, to avoid costly calculations of $\mathbf{H}_k$ at each step, the well-known Broyden–Fletcher–Goldfarb–Shanno (BFGS) Hessian update algorithm[160–163]

$$\mathbf{H}_k = \mathbf{H}_{k-1} + \Delta \mathbf{H}_{\mathrm{BFGS}}, \tag{19}$$

is employed to estimate the Hessian from previous iterations, $\mathbf{H}_{k-1}$, and an update factor $\Delta \mathbf{H}_{\mathrm{BFGS}}$ calculated from the gradient.

A further addition extends the RFO to form the basis of ANCOPT: The entire procedure is conducted in approximate normal coordinates (ANC), which are based on the initial $\mathbf{H}_k$ computed

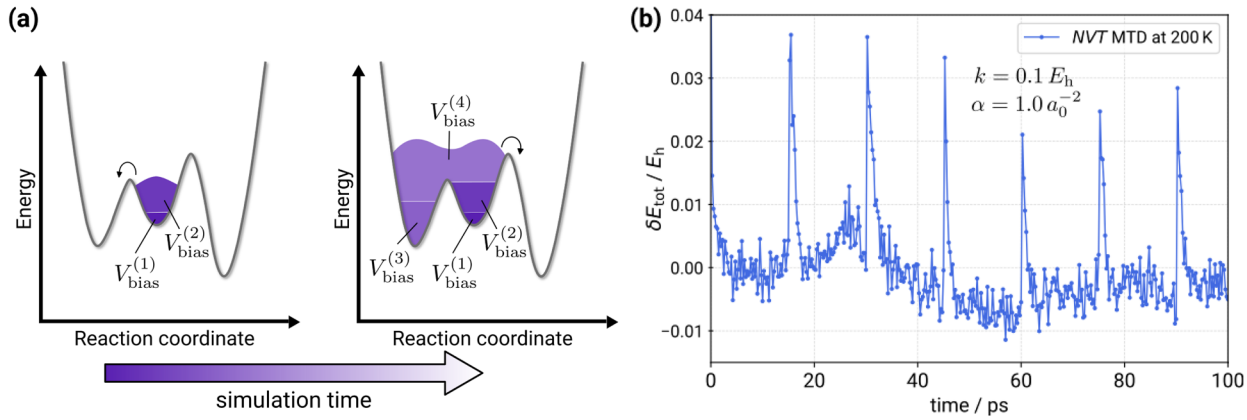

FIG. 17. Effect of the RMSD bias potential. (a) Schematic representation of sequentially adding bias configurations to the metadynamics simulation. (b) 100 ps MTD of the Tamiflu molecule at 200 K, calculated at the GFN-FF level. A new bias configuration is added to $\mathcal{B}$ every 15 ps, resulting in a brief, sharp increase in the total energy. The total energy is plotted relative to its average value.

from a model Hessian.[164] For molecular geometries, internal coordinates are often a preferred choice.[165] This is demonstrated for the taxol molecule in Fig. 18, where ANCOPT is compared to L-BFGS[166,167] employing Cartesian coordinates, as implemented in the `OPTIM` program.[168] For comparability, both programs use the same `tblite` backend and employ the GFN2-xTB Hamiltonian.

Clearly, ANCOPT shows superior performance, requiring about 200 steps less than the L-BFGS minimization to converge into the closest minimum satisfying the same energy ($\Delta E = 0.5 \times 10^{-4}$) and root-mean-square (rms) force ($\|\nabla E\| = 0.5 \times 10^{-4}$) convergence criteria. The overall algorithm of ANCOPT is highly optimized for the rapid geometry optimization of molecular geometries that are needed during the demanding task of CSE, where often thousands of such evaluations are required.

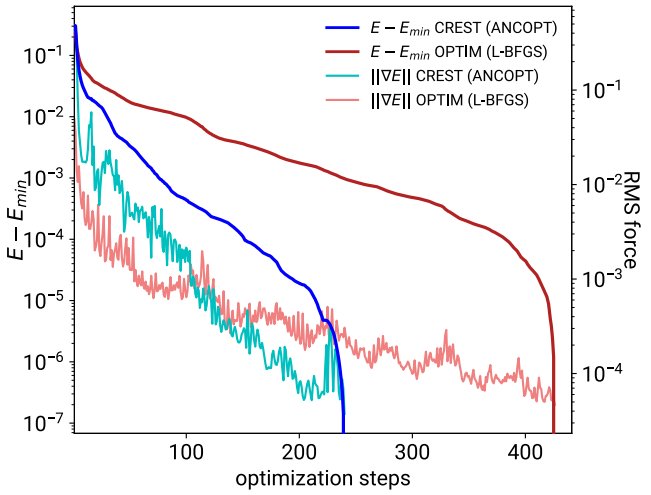

FIG. 18. Comparison of the ANCOPT and L-BFGS geometry optimization algorithms for the taxol molecule (113 atoms) in terms of energy and rms force $\|\nabla E\|$ convergence, calculated at the GFN2-xTB level. The energy convergence is shown relative to the energy of the converged minimum ($E - E_{min}$ in E_h). Both quantities are plotted on a logarithmic scale.

For CSE, CREST utilizes a multi-level optimization strategy. Here, during the sampling iterations, molecular geometries are optimized in (typically) three steps with increasingly tight convergence criteria. In between these steps, CREGEN is used to filter out the high-energy intermediates that would require a large number of optimization cycles in order to converge tightly and duplicated structures that clearly would converge to the same minimum. Different convergence criteria are predefined to simplify the selection. The corresponding settings are shown in Table IV.

3. Numerical Hessian calculation

The architecture of CREST allows for the extraction of energies and gradients from every method that is accessible through the new calculator APIs for other software. Based on this, we implemented a procedure for the computation of a (semi–)numerical Hessian independent of the provided method. The derivatives of the gradients (hence the second derivatives of the energy) are computed by the symmetric two-point finite difference quotient,

$$\frac{\partial \mathbf{g}}{\partial q_i} = \mathbf{h}_{q_i} \approx \frac{\mathbf{g}_{q_i+d} - \mathbf{g}_{q_i-d}}{2d}, \tag{20}$$

TABLE IV. Predefined optimization levels of ANCOPT in terms of the change in energy (in Hartree), rms force ($\|\nabla E\|$ in Hartree E_h per ANC α), and the maximum number of optimization cycles, which defaults to 200 in case N_{at} is small.

Setting	$\Delta E_{\mathrm{conv}}/E_\mathrm{h}$	$\|\nabla E\|/E_\mathrm{h}\alpha^{-1}$	Max. Cycles
Crude	5×10^{-4}	1×10^{-2}	N_{at}
Vloose	1×10^{-4}	6×10^{-3}	N_{at}
Loose	5×10^{-5}	4×10^{-3}	$2N_{\mathrm{at}}$
Default	5×10^{-6}	1×10^{-3}	$3N_{\mathrm{at}}$
Tight	1×10^{-6}	8×10^{-4}	$5N_{\mathrm{at}}$
Vtight	1×10^{-7}	2×10^{-4}	$20N_{\mathrm{at}}$
Extreme	5×10^{-8}	5×10^{-5}	$20N_{\mathrm{at}}$

where $\mathbf{g}_{q_i \pm d}$ are the gradients of the geometry with the displaced (Cartesian) coordinate q_i of atom i, and d is the size of the corresponding displacement. $\mathbf{h}_{q_i}$ provides one column of the complete Hessian matrix $\mathbf{H}$ in Cartesian coordinates, requiring $3N$ two point finite difference calculations for the complete Hessian. The associated gradient calculations hence lead to a formal computational cost of $\mathcal{O}(6N)$.

Due to the numerical differentiation of the (analytic) gradients and deviations originating from this, $\mathbf{H}$ is not exactly symmetric, and a symmetrization is performed afterward by taking the mean of the Hessian and its transpose. Furthermore, the symmetric difference quotient approach for the computation of the Hessian in Cartesian coordinates inherits the impurity of translational and rotational contributions, which need to be projected out. The corresponding projection matrix is constructed in two steps. First, the coordinates are shifted to the geometrical center via

$$\mathbf{q}_j = \mathbf{q}'_j - \frac{1}{N_{\mathrm{at}}} \sum_i^{N_{\mathrm{at}}} \mathbf{q}'_i. \tag{21}$$

Second, a matrix $\mathbf{F} \in \mathbb{R}^{3N \times 6}$ is constructed, where each row represents either a translation or rotation of the entire system. The orthonormalization of $\mathbf{F}$ provides the projection matrix $\mathbf{F}^{\mathrm{p}}$, which is used for the projection

$$\mathbf{H}^{\mathrm{p}} = \left(\mathbf{I} - \mathbf{F}^{\mathrm{p}}\mathbf{F}^{\mathrm{p}\top}\right)\mathbf{H}\left(\mathbf{I} - \mathbf{F}^{\mathrm{p}}\mathbf{F}^{\mathrm{p}\top}\right)^{\top}. \tag{22}$$

The projected Hessian $\mathbf{H}^{\mathrm{p}}$ does not include any translational or rotational contributions, i.e., the corresponding eigenvalues are zero. For the computation of the frequencies in the harmonic oscillator approximation, the Hessian is mass weighted, as shown in the following equation:

$$h^{\mathrm{p,m}}_{q_i q_j} = \frac{1}{\sqrt{m_{q_i} m_{q_j}}} h^{\mathrm{p}}_{q_i q_j}, \tag{23}$$

where $h^{\mathrm{p,m}}_{q_i q_j}$ is one element of $\mathbf{H}^{\mathrm{p,m}}$. Solving the eigenvalue problem for $\mathbf{H}^{\mathrm{p,m}}$ provides the force constants and the normal modes of the corresponding vibrations. As an extension to this procedure, CREST implements the so-called `thermo` routine originating from the `xtb` code[53] which, given the molecular structure and corresponding frequencies, calculates thermostatistical quantities (enthalpy, entropy, ZPVE, heat capacity) at different temperatures for the translational, rotational, and vibrational DOF. The program herein applies the msRRHO approach.[40,108]

4. Effective Hessian

At an MECP (see Sec. II E), the PESs of both electronic states are not at a stationary point along all $3N - 6$ dimensions. Therefore, the computation of nuclear quantum effects cannot be performed in the typical way of computing the Hessian. Harvey *et al.* proposed an effective Hessian,[169,170] which follows from the second-order Taylor expansion along the crossing hyperline according to

$$E = E_{\mathrm{MECP}} + \frac{1}{2}\Delta\mathbf{q}^{\top}\mathbf{H}_{\mathrm{eff}}\Delta\mathbf{q}, \tag{24}$$

where $\Delta\mathbf{q}$ is the displacement along the crossing hyperline, thus in the dimensions, where the PESs are degenerate. The displacement along the hyperline is orthogonal to the gradient difference of the two PESs $\Delta\mathbf{g}$. The effective Hessian $\mathbf{H}_{\mathrm{eff}} \in \mathbb{R}^{3N \times 3N}$ [Eq. (25)] requires knowledge about the Hessians $\mathbf{H}_1, \mathbf{H}_2$ and gradients $\mathbf{g}_1, \mathbf{g}_2 \in \mathbb{R}^{3N}$ from the two crossing PESs,

$$\mathbf{H}_{\mathrm{eff}} = \frac{\|\mathbf{g}_2\|\mathbf{H}_1 \pm \|\mathbf{g}_1\|\mathbf{H}_2}{\|\Delta\mathbf{g}\|}. \tag{25}$$

Here, $\Delta\mathbf{g} = \mathbf{g}_1 - \mathbf{g}_2$ is the gradient difference. Analogous to the classification of conical intersections (CIs) into peaked and sloped, Lykhin *et al.* suggested an approach for the differentiation of these two classes based on the vector product of the gradients $\mathbf{g}_1\mathbf{g}_2$ for MECPs.[171,172] If the vector product $\mathbf{g}_1\mathbf{g}_2 > 0$, the intersection is considered sloped, and a minus sign must be used. In the other case, $\mathbf{g}_1\mathbf{g}_2 < 0$, the intersection is peaked, and a plus sign is used. Additionally, since the MECP is a minimum along $3N$-7 modes, in addition to the projection shown for the seminumerical Hessian [Eq. (22)], a projection of the normal mode orthogonal to the seam is performed. The corresponding mode describes the reaction path between the minima that are connected by the MECP and corresponds to the gradient difference $\Delta\mathbf{g}$. The latter is used to create the projection[172] according to

$$\mathbf{H}^{\mathrm{p}}_{\mathrm{eff}} = \left(\mathbf{I} - \frac{\Delta\mathbf{g}\Delta\mathbf{g}^{\top}}{\|\Delta\mathbf{g}\|^2}\right)\mathbf{H}_{\mathrm{eff}}\left(\mathbf{I} - \frac{\Delta\mathbf{g}\Delta\mathbf{g}^{\top}}{\|\Delta\mathbf{g}\|^2}\right). \tag{26}$$

After mass weighting and diagonalizing the Hessian, as described earlier, the corresponding frequencies of the vibrations are obtained.

To show the influence of the nuclear quantum effects on the rate of state crossing, the allene molecule is shown as an example in Fig. 19. The effective Hessian was computed using the newly implemented procedures of CREST with the composite density functional theory method PBEh-3c[136] via the ORCA subprocess interface. Here, restricted Kohn–Sham formalism was used for the S_0 minimum and unrestricted formalism was used for the MECP and T_1 geometries, and thermodynamic corrections were computed using the RRHO approximation.[102]

This case study clearly reveals that thermodynamic corrections computed via the effective Hessian at the MECP and via the Hessian at the T_1 geometry lead to a near-degeneracy of the two states.

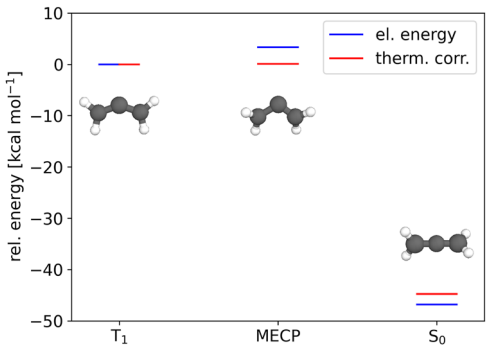

FIG. 19. Relative energies of the minima of T_1 and S_0 and the MECP between the states are shown. The energies are shown relative to the respective triplet energy. The chosen level of theory for the geometry optimization and the computation of the Hessian is PBEh-3c.

An inclusion of the thermodynamic contributions would, therefore, show an increase in the intersystem crossing rate, proving their importance for describing non-radiative relaxation processes in an excited state.

It should be mentioned that the GFN*n*-xTB methods, due to the lack of spin-discriminating energy terms, lead to fully adiabatic states. Hence, only avoided crossings are obtained. Therefore, there is no need for the use of an effective Hessian with these methods.

5. Multicenter n-layer ONIOM

CREST implements a new standalone Fortran library called lwONIOM[172,173] to allow for subtractive QM/MM (i.e., ONIOM[174,175]) calculations with, in principle, an arbitrary number of levels and multiple center definitions. The QM/MM methodology is especially valuable for investigating macroscopic and biomolecular systems.[176–178] Computational efficiency is achieved by selectively employing high-level QM or SQM calculations in a specific region of interest while resorting to lower-level MM calculations for the rest of the system.

The multi-layer and multi-center ONIOM (denoted as MC-ONIOM*n*, with *n* being the number of layers) implementation in lwONIOM follows the recent work by Seeber *et al.*[179] Here, the ONIOM layer-dependencies are easiest represented as a tree graph, as shown in Fig. 20. Each node represents a substructure of the original system that, with the exception of the initial structure, is tied to a parent node depending on its layer. Atoms from the initial system must be assigned to various nodes by the user, where truncated bonds are automatically saturated by linking atoms. The final dependency tree allows a recursive construction of the total ONIOM property. The general formulation for constructing the energy, gradient, or Hessian of a given node, denoted as $\mathcal{F}_i$ follows the equation

$$\mathcal{F}_i = \mathcal{F}_i^h + \sum_j (\mathcal{F}_j - \mathcal{F}_j^l). \tag{27}$$

Here, $\mathcal{F}_i^h$ represents the high-level energy/gradient construction of the parent node, $\mathcal{F}_j$ signifies the recursively constructed property of the children nodes, and $\mathcal{F}_j^l$ denotes their low-level energy/gradient/Hessian contribution. It is crucial to emphasize that the high-level calculations of each parent node correspond to the same level of theory as the low-level calculations of its associated children nodes. The recursion terminates when a node has no further child nodes, in which case $\mathcal{F}_i = \mathcal{F}_i^h$. The gradient of the *i*th model system $\mathbf{g}_i'$ can be projected into the basis of the real system $\mathbf{g}_i$ via

$$\mathbf{g}_i = \mathbf{g}_i' \mathbf{J}_i. \tag{28}$$

In a similar fashion, the Hessian matrix of the individual fragments can be projected onto the basis of the real system.[179] The Jacobian employed in Eq. (28) is given by

$$\mathbf{J}_i = \begin{pmatrix} \mathbf{J}_{11} & \cdots & \mathbf{J}_{1n} \\ \vdots & \ddots & \vdots \\ \mathbf{J}_{m1} & \cdots & \mathbf{J}_{mn} \end{pmatrix}, \tag{29}$$

where m is the dimension of the *i*th subsystem or fragment and n is the dimension of the real system. The corresponding matrix elements $\mathbf{J}_{nm}$ are given by

$$\mathbf{J}_{mn} = \mathbf{E} \cdot \frac{\partial \mathbf{r}_m'}{\partial \mathbf{r}_n}, \tag{30}$$

where $\mathbf{E}$ is a 3×3 identity matrix. The derivatives $\frac{\partial \mathbf{r}_m'}{\partial \mathbf{r}_n}$ are either 1, 0, or a value depending on the linking atom position.[174,179] Once all gradients (or Hessians) of the subsystems have been projected into the basis of the real system, constructing the full MC-ONIOM*n* gradient is straightforward via the recursive algorithm given by Eq. (27).

In a benchmarking example, MC-ONIOM*n* calculations ($n = 2$ layers) were performed for a 5,15-linked porphyrin nanoring C–C coupling reaction [Figs. 21(a) and 21(b)].[180] Due to the size of the system, the nanoring allows the definition of six bis-porhyrin units that can either be set up as independent high-level ONIOM region or merged into a single such layer. Similar to previous studies in the literature, in particular Ref. 181, calculations were performed employing GFN1-xTB for the high SQM layer, together with GFN-FF for the low MM layer. Both layers employed ALPB implicit solvation[140] for toluene. The conventional two-level SQM/MM ONIOM approach [ONIOM2(GFN1-xTB:GFN-FF)] was compared with the six-centered MC-ONIOM2(GFN1-xTB:GFN-FF) variant. Timings are shown in Fig. 21(c), reaction energies (per C–C bond formed), and Cartesian RMSDs compared to the GFN2-xTB starting structure are provided in Table V.

MC-ONIOM2 results quite clearly provide a considerably greater trade-off between calculation cost and accuracy compared to the conventional ONIOM2 setup. Calculated coupling energies differ by only 0.03 kcal mol^{-1} between the two ONIOM variants, and both are within less than 1 kcal mol^{-1} of the reference GFN1-xTB value. However, the multi-center treatment speeds up the calculation by an additional factor >2 on the same machine [cf. Fig. 21(c)] compared to the "standard" two-center ONIOM2. Expectantly, coupling energies calculated only at the GFN-FF level are not plausible due to the classical treatment.

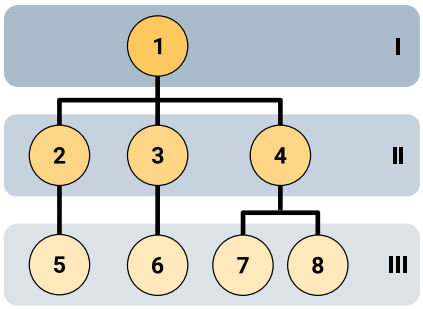

FIG. 20. Schematic construction of the MC-ONIOM*n* dependency tree for a setup with three layers. Fragments are represented as nodes. The highest layer (I) must include a single node containing all the atoms of the system. All other layers (II, III) can include multiple non-overlapping subsystems, treated as children nodes of the original node or nodes of subsequently higher layers.

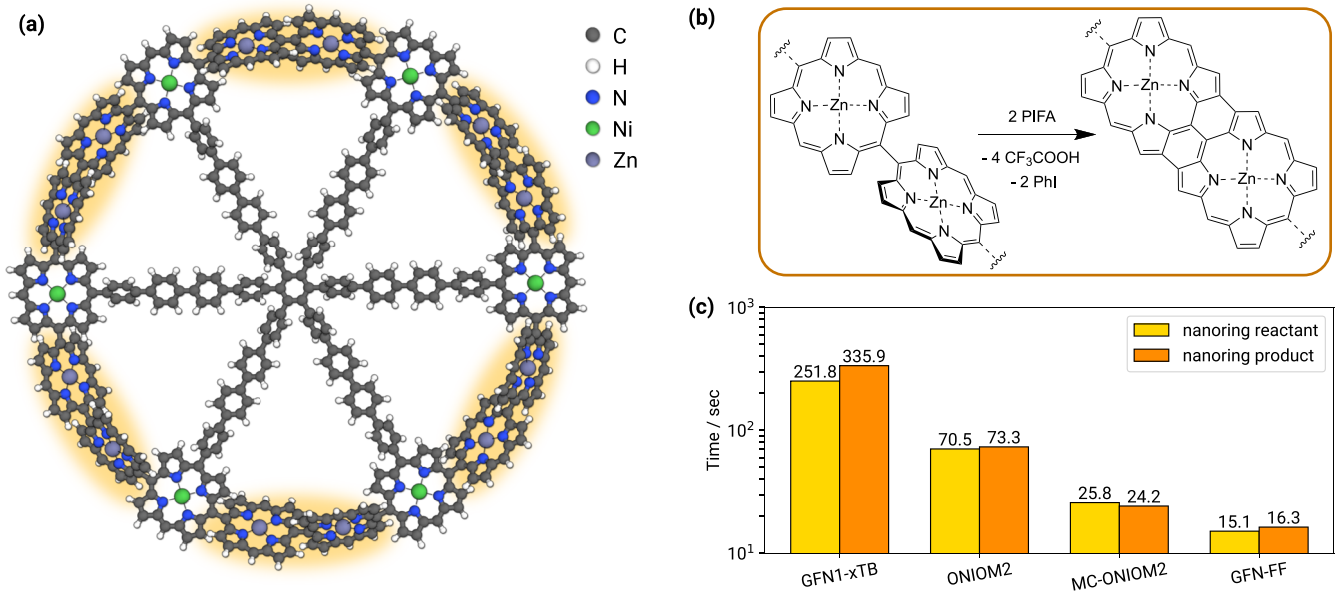

FIG. 21. (a) Nanoring structure adapted from Refs. 180 and 181. Yellow-shaded Zn-functionalized porphyrin units are used for the high-level QM calculations within the ONIOM setup. (b) 5,15-linked porphyrin nanoring C–C coupling reaction. (c) Computation times referring to the full geometry optimization (starting from the GFN2-xTB optimized structures taken from Ref. 181) wall time on 8 CPUs. ONIOM2 refers to a conventional two-layer GFN1-xTB:GFN-FF setup,[181] while MC-ONIOM2 refers to the multicenter setup with six independent bis-porphyrin units forming a single GFN1-xTB "high-level" layer.

The ONIOM model setup within the CREST input files follows conventions, as demonstrated in Listing 3. Fragments (nodes) are defined on a by-atom basis; layers are defined by fragments. Finally, each layer is tied to a level of theory from the `[[calculation.level]]` TOML list (cf. Listing 1).

While this setup technically allows for the use of ONIOM energy and gradient in the sampling algorithms of CREST, we recommend avoiding using it for RMSD-based metadynamics applications. Due to the strong bias potential and the weak points introduced by the mechanically embedded layers, undesirable breaking of bonds can frequently occur. In those instances, it is recommended to first perform the sampling at a cost-efficient level of theory, for example, with the force-field employed for the outer ONIOM layer, and in a second step, evaluate the generated structures within the ONIOM framework. Similar built-in strategies are discussed in Sec. III B 6.

TABLE V. C–C coupling reaction energies of the 5,15-linked porphyrin nanoring. Values were calculated for the corresponding molecular geometries optimized at the specified levels of theory. Atomic RMSDs relative to the GFN2-xTB optimized structure (taken from Ref. 181) are given for the more flexible nanoring reactant.

Method	ΔE (kcal mol^{-1})	RMSD (Å)
GFN1-xTB	−36.89	0.070
ONIOM2(GFN1-xTB:GFN-FF)	−36.74	0.212
MC-ONIOM2(GFN1-xTB:GFN-FF)	−36.71	0.213
GFN-FF	155.20	0.392

6. "On-the-fly" multiscale modeling

In consideration of computational efficiency, it is sometimes not desirable to run certain parts of the sampling algorithms, in particular metadynamics simulations, at the more expensive SQM level rather than at a force-field level. Likewise, the post-processing of structures by singlepoint energy calculations on the optimized structures might be feasible and useful. The modular construction of CREST 3.0's calculator allows the definition of such multiscale setups and can be used in algorithms like iMTD-GC and iMTD-sMTD. A minimal example is shown in Listing 4.

This example input will automatically run iMTD-GC in a three-stage multiscale approach: All molecular dynamics and metadynamics simulations will be conducted at the GFN-FF level, geometries will then be optimized at the GFN0-xTB level, and finally, a singlepoint energy on the optimized geometries will be calculated at the GFN2-xTB level. Only the latter will be used for the CREGEN evaluation. This framework enables sophisticated calculation setups; however, it is currently incompatible with the other multiscale models described earlier (ONIOM and MECPs). Furthermore, the user is responsible to judge whether the intended setup is sensible. For example, while a DFT singlepoint refinement via ORCA could easily be defined within the multiscale setup, integrating this into iMTD-GC would still lead to unfavorably high computation times and is advised against.

IV. PROGRAM DISTRIBUTION

CREST is developed with an open-source focus under the LGPL-3.0 license. All code is publicly accessible on GitHub at

LISTING 3. ONIOM setup example for the nanoring system shown in Fig. 21 in TOML format.

```
[lwoniom]
natoms = 846
# Fragments in ascending order
fragment.1 = 'all'   # fragment 1 contains all atoms (may be omitted)
fragment.2 = '401-445,839,840,8-10,766-780,584'
fragment.3 = '446-493,841,842,781-796'
fragment.4 = '494-539,6,7,797-812,843,844'
fragment.5 = '540-582,1-5,845,846,813-827'
fragment.6 = '308-353,12,13,585,735-749,835,836'
fragment.7 = '354-400,11,837,838,750-764'
# Define layers on an by-fragment basis
layer.1 = [1]  # layer 1 contains only fragment 1
layer.2 = [2,3,4,5,6,7] # layer 2 constains 6 fragments
# Tie each layer to a calculation level
layerlevel.1 = 2  # tie layer 1 to the 2nd [[calculation.level]]
layerlevel.2 = 1  # tie layer 2 to the 1st [[calculation.level]]
```

LISTING 4. Minimal TOML input example for setting up a multiscale run of iMTD-GC.

```
input = 'struc.xyz'
runtype = 'imtd-gc'
[[calculation.level]] # 1st calculation level
method = 'gfnff'
[[calculation.level]] # 2nd calculation level
method = 'gfn0'
[[calculation.level]] # 3rd calculation level
method = 'gfn2'
refine = 'singlepoint' # only used for a singlepoint re-ranking
[dynamics] # MD/MTD options block
active = [1] # the first [[calculation.level]] is active only in MD/MTD
```

https://github.com/crest-lab/crest. The repository includes basic build instructions and provides a continuous release binary that is automatically compiled from the most recent commit on the main branch using a `meson/ifort` environment. A pre-built binary is also supported by the conda-forge infrastructure. Build options are available with the CMake and `meson` build systems for Intel and GNU compilers. We generally recommend using the Intel architecture.

Developing scientific software as open-source not only accelerates scientific progress through global collaboration but also enhances transparency, allowing fellow researchers to validate and reproduce findings.[5,7] Moreover, open-source software grants the advantage of customization, enabling others to adapt the software to their specific research needs, often leading to novel applications. The long-term sustainability of open-source projects ensures that software remains relevant and useful over time, contributing to a lasting impact on the scientific community. At its core, open-source development embodies the principle of shared knowledge and aligns with the ethical responsibility of advancing science for the greater good.

The choice of the GNU Lesser General Public License version 3.0 (LGPL-3.0) reflects our commitment to preserving the open nature of the project and its benefits for both individuals and organizations. The LGPL-3.0 license permits the use of CREST in both open-source and proprietary projects, enabling wider adoption without compromising the freedom to inspect, modify, and distribute improvements to the core software. Furthermore, it ensures that any modifications made to the CREST codebase remain open and accessible to the community, promoting continued growth and improvement of the toolkit.

Besides the main repository, online documentation with example show-cases is maintained at https://crest-lab.github.io/crest-docs.

V. CONCLUSION

In this paper, we present the latest developments of the chemical space exploration software CREST. The program is designed as a straightforward solution to the automated sampling problem of molecular systems up to a size of a few hundred atoms. With a focus on efficiency and single-node parallelization, several workflows based on semiempirical electronic structure calculations are implemented and can be used even on standard desktop computers. The main features of CREST 3.0 include

- implementation of the GFNn-xTB and GFN-FF methods via `tblite`[139] and standalone calculators,[141,142] providing efficient semiempirical calculations for systems containing all elements up to radon ($Z \leq 86$) in standard applications such as energy and gradient calculations, geometry optimization, and molecular dynamics simulations,
- capabilities to interface any atomistic simulation program providing gradients. The current code comes with interfaces to the ORCA code[134,135] and an implementation of the X-HCFF method,[96]
- automated algorithms for conformational sampling and the calculation of configurational entropy based on the aforementioned levels of theory, including automated multiscale modeling variants,
- automated algorithms for the efficient generation of explicit solvation shells and the determination of molecular protonation sites and,
- calculators for simultaneously handling multiple levels of theory, in particular for the calculation of MECPs and for the setup of multi-center and multi-layer ONIOM calculations.

In conclusion, the enduring relevance of CREST lies in the robust versatility of its algorithms and general backend implementation. Newly developed potentials or methods can be quickly interfaced with the existing infrastructure. Hence, as technology continues to advance, CREST's adaptability ensures its continued efficacy, making it a reliable tool for molecular CSE, ready to meet the challenges of an evolving scientific landscape.

Future development efforts will be focused on three primary directions. First, the integration of algorithmic advances such as rapid coarse-grained generation of molecular conformations and global optimization techniques. Second, there are technical improvements to the code, for example, the inclusion of periodic boundary conditions for xTB methods,[182] further acceleration of calculations through message passing interface (MPI) parallelization for high-performance computing (HPC) facilities, and the creation of a Python frontend. Finally, generally more accurate CSE will be achieved by next-generation SQM methods, which are currently explored through new tight-binding approaches[183] and the novel q-vSZP basis.[184] A new method (GP3-xTB) is currently being developed and will be published separately soon.

SUPPLEMENTARY MATERIAL

The supplementary material provides a spreadsheet containing detailed benchmarking results for Sec. II A 1.

ACKNOWLEDGMENTS

P.P. gratefully acknowledges support from the Alexander von Humboldt Foundation for a Feodor Lynen Research Fellowship. S.G. acknowledges the Deutsche Forschungsgemeinschaft for general funding, mainly in the framework of the Leibniz price. C.B. is grateful for funding from the Ministry of Culture and Science of the German State of North Rhine-Westphalia (MKW) via the NRW Rückkehrprogramm. G.F. acknowledges funding from RWTH Graduate Support. P.A.W. acknowledges the Engineering and Physical Sciences Research Council (EPSRC) for funding his studentship through Doctoral Training Partnership Grant No. EP/W524633/1. M.M. and S.S. acknowledge the Fonds der Chemischen Industrie (FCI) for funding under a Kekulé scholarship. The authors acknowledge David J. Wales, Andreas Hansen, Markus Bursch, and Jan-Michael Mewes for valuable discussions.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Philipp Pracht: Conceptualization (lead); Data curation (equal); Methodology (lead); Software (lead); Writing – original draft

(equal); Writing – review & editing (equal). **Stefan Grimme**: Conceptualization (lead); Methodology (lead); Software (equal); Writing – original draft (equal); Writing – review & editing (equal). **Christoph Bannwarth**: Conceptualization (equal); Methodology (equal); Software (equal); Writing – original draft (equal); Writing – review & editing (equal). **Fabian Bohle**: Data curation (equal); Writing – original draft (equal); Writing – review & editing (equal). **Sebastian Ehlert**: Software (equal); Writing – original draft (equal); Writing – review & editing (equal). **Gereon Feldmann**: Software (equal); Writing – original draft (equal); Writing – review & editing (equal). **Johannes Gorges**: Software (equal); Writing – original draft (equal); Writing – review & editing (equal). **Marcel Müller**: Data curation (equal); Writing – original draft (equal); Writing – review & editing (equal). **Tim Neudecker**: Software (equal); Writing – original draft (equal); Writing – review & editing (equal). **Christoph Plett**: Software (equal); Writing – original draft (equal); Writing – review & editing (equal). **Sebastian Spicher**: Software (equal); Writing – original draft (equal); Writing – review & editing (equal). **Pit Steinbach**: Software (equal); Writing – original draft (equal); Writing – review & editing (equal). **Patryk A. Wesołowski**: Software (equal); Writing – original draft (equal); Writing – review & editing (equal). **Felix Zeller**: Software (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request. The software is freely available on GitHub.

NOMENCLATURE

ALPB	analytical linearized Poisson–Boltzmann (implicit solvation model)
ANC	approximate normal coordinate
ANCOPT	approximate normal coordinate optimization algorithm
BFGS	Broyden–Fletcher–Goldfarb–Shanno
CN	coordination number
CRE	conformer–rotamer ensemble
CREGEN	conformer–rotamer ensemble generation algorithm
CREST	conformer–rotamer sampling tool
CSE	chemical space exploration
DFT	density functional theory
DOF	degrees of freedom
Eu	entropy units (cal/mol K)
FF	force-field
GC	genetic crossing
GFN	geometries, frequencies, non-covalent interactions
HPC	Hartree–Fock
HF	high-performance computing
iMTD-GC	iterative MTD + GC conformer sampling algorithm
iMTD-sMTD	iterative metadynamics-static metadynamics algorithm
LMO	localized molecular orbital
MAD	mean absolute deviation
MC-ONIOM	multi-center ONIOM
MD	molecular dynamics
MECI	minimum energy conical intersection
ML	machine learning
MM	molecular mechanics
MO	molecular orbital
MOF	metal-organic framework
MPI	message passing interface
msRRHO	modified and scaled rigid-rotor harmonic-oscillator approximation
MTD	metadynamics
NCI	non-covalent interaction
NCI-MTD	NCI-focused metadynamics sampling algorithm
PCA	principal component analysis
PEL	potential energy landscape (synonym to PES)
PES	potential energy surface (synonym to PEL)
QCG	quantum cluster growth algorithm
QHA	quasi harmonic analysis
RFO	rational function optimization
RMSD	root-mean-square-deviation (of atomic positions)
RMSE	root-mean-square-error
SASA	solvent accessible surface area
sMTD	"static" metadynamics (≈umbrella sampling)
SQM	semiempirical quantum mechanics
SSR	sum of squares regression
SST	total sum of squares
WTMAD	weighted mean absolute deviation
XC	exchange-correlation
X-HCFF	extended hydrostatic compression force-field
xTB	extended tight-binding
ZPVE	zero-point vibrational energy

REFERENCES

[1] F. Neese, M. Atanasov, G. Bistoni, D. Maganas, and S. Ye, "Chemistry and quantum mechanics in 2019: Give us insight and numbers," J. Am. Chem. Soc. **141**, 2814–2824 (2019).

[2] J. Gasteiger, "Chemistry in times of artificial intelligence," ChemPhysChem **21**, 2233–2242 (2020).

[3] K. N. Houk and F. Liu, "Holy grails for computational organic chemistry and biochemistry," Acc. Chem. Res. **50**, 539–543 (2017).

[4] S. Grimme and P. R. Schreiner, "Computational chemistry: The fate of current methods and future challenges," Angew. Chem., Int. Ed. **57**, 4170–4176 (2017).

[5] R. Di Felice, M. L. Mayes, R. M. Richard, D. B. Williams-Young, G. K.-L. Chan, W. A. de Jong, N. Govind, M. Head-Gordon, M. R. Hermes, K. Kowalski, X. Li, H. Lischka, K. T. Mueller, E. Mutlu, A. M. N. Niklasson, M. R. Pederson, B. Peng, R. Shepard, E. F. Valeev, M. van Schilfgaarde, B. Vlaisavljevich, T. L. Windus, S. S. Xantheas, X. Zhang, and P. M. Zimmerman, "A perspective on sustainable computational chemistry software development and integration," J. Chem. Theory Comput. **19**, 7056–7076 (2023).

[6] C. D. Sherrill, D. E. Manolopoulos, T. J. Martínez, and A. Michaelides, "Electronic structure software," J. Chem. Phys. **153**, 070401 (2020).

[7] A. I. Krylov, J. M. Herbert, F. Furche, M. Head-Gordon, P. J. Knowles, R. Lindh, F. R. Manby, P. Pulay, C.-K. Skylaris, and H.-J. Werner, "What is the price of open-source software?," J. Phys. Chem. Lett. **6**, 2751–2754 (2015).

[8] A. Krylov, T. L. Windus, T. Barnes, E. Marin-Rimoldi, J. A. Nash, B. Pritchard, D. G. A. Smith, D. Altarawy, P. Saxe, C. Clementi, T. D. Crawford, R. J. Harrison, S. Jha, V. S. Pande, and T. Head-Gordon, "Perspective: Computational chemistry software and its advancement as illustrated through three grand challenge cases for molecular science," J. Chem. Phys. **149**, 180901 (2018).

[9] 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, "The atomic simulation environment–A Python library for working with atoms," J. Phys.: Condens. Matter **29**, 273002 (2017).
[10]A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, and S. J. Plimpton, "LAMMPS—A flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales," Comput. Phys. Commun. **271**, 108171 (2022).
[11]RDKit: Open-source cheminformatics, https://github.com/rdkit/rdkit, 2024.
[12]C. Lipinski and A. Hopkins, "Navigating chemical space for biology and medicine," Nature **432**, 855–861 (2004).
[13]P. S. Gromski, A. B. Henson, J. M. Granda, and L. Cronin, "How to explore chemical space using algorithms and automation," Nat. Rev. Chem. **3**, 119–128 (2019).
[14]D. J. Wales, *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses* (Cambridge University Press, 2003).
[15]M. Bursch, J.-M. Mewes, A. Hansen, and S. Grimme, "Best-practice DFT protocols for basic molecular computational chemistry," Angew. Chem., Int. Ed. **61**, e202205735 (2022).
[16]S. Grimme, F. Bohle, A. Hansen, P. Pracht, S. Spicher, and M. Stahn, "Efficient quantum chemical calculation of structure ensembles and free energies for nonrigid molecules," J. Phys. Chem. A **125**, 4039–4054 (2021).
[17]D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," J. Chem. Inf. Comput. Sci. **28**, 31–36 (1988).
[18]A. McNaught, "The IUPAC international chemical identifier: InChl–A new standard for molecular informatics," Chem. Int. **28**, 12–15 (2006).
[19]P. Pracht, F. Bohle, and S. Grimme, "Automated exploration of the low-energy chemical space with fast quantum chemical methods," Phys. Chem. Chem. Phys. **22**, 7169–7192 (2020).
[20]P. C. D. Hawkins, "Conformation generation: The state of the art," J. Chem. Inf. Model. **57**, 1747–1756 (2017).
[21]D. J. Wales, "Exploring energy landscapes," Annu. Rev. Phys. Chem. **69**, 401–425 (2018).
[22]V. Leone, F. Marinelli, P. Carloni, and M. Parrinello, "Targeting biomolecular flexibility with metadynamics," Curr. Opin. Struct. Biol. **20**, 148–154 (2010).
[23]J. Vymětal and J. Vondrášek, "Metadynamics as a tool for mapping the conformational and free-energy space of peptides—The alanine dipeptide case study," J. Phys. Chem. B **114**, 5632–5642 (2010).
[24]D. J. Wales and H. A. Scheraga, "Global optimization of clusters, crystals, and biomolecules," Science **285**, 1368–1372 (1999).
[25]D. J. Wales and J. P. K. Doye, "Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms," J. Phys. Chem. A **101**, 5111–5116 (1997).
[26]P. C. D. Hawkins, A. G. Skillman, G. L. Warren, B. A. Ellingson, and M. T. Stahl, "Conformer generation with OMEGA: Algorithm and validation using high quality structures from the protein databank and cambridge structural database," J. Chem. Inf. Model. **50**, 572–584 (2010).
[27]J.-G. Sobez and M. Reiher, "Molassembler: Molecular graph construction, modification, and conformer generation for inorganic and organic molecules," J. Chem. Inf. Model. **60**, 3884–3900 (2020).
[28]K. S. Watts, P. Dalal, R. B. Murphy, W. Sherman, R. A. Friesner, and J. C. Shelley, "ConfGen: A conformational search method for efficient generation of bioactive conformers," J. Chem. Inf. Model. **50**, 534–546 (2010).
[29]M. A. Miteva, F. Guyon, and P. Tufféry, "Frog2: Efficient 3D conformation ensemble generator for small compounds," Nucleic Acids Res. **38**, W622–W627 (2010).
[30]N. Sauton, D. Lagorce, B. O. Villoutreix, and M. A. Miteva, "MS-DOCK: Accurate multiple conformation generator and rigid docking protocol for multi-step virtual ligand screening," BMC Bioinf. **9**, 184 (2008).
[31]M. J. Vainio and M. S. Johnson, "Generating conformer ensembles using a multiobjective genetic algorithm," J. Chem. Inf. Model. **47**, 2462–2474 (2007).
[32]S. Riniker and G. A. Landrum, "Better informed distance geometry: Using what we know to improve conformation generation," J. Chem. Inf. Model. **55**, 2562–2574 (2015).
[33]E. A. Coutsias, K. W. Lexa, M. J. Wester, S. N. Pollock, and M. P. Jacobson, "Exhaustive conformational sampling of complex fused ring macrocycles using inverse kinematics," J. Chem. Theory Comput. **12**, 4674–4687 (2016).
[34]S. Axelrod and R. Gómez-Bombarelli, "GEOM, energy-annotated molecular conformations for property prediction and molecular generation," Sci. Data **9**, 185 (2022).
[35]S. Axelrod and R. Gómez-Bombarelli, "Molecular machine learning with conformer ensembles," Mach. Learn.: Sci. Technol. **4**, 035025 (2023).
[36]K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," Nature **559**, 547–555 (2018).
[37]O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, "Machine learning force fields," Chem. Rev. **121**, 10142–10186 (2021).
[38]A. S. Christensen, T. Kubař, Q. Cui, and M. Elstner, "Semiempirical quantum mechanical methods for noncovalent interactions for chemical and biochemical applications," Chem. Rev. **116**, 5301–5337 (2016).
[39]C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher, and S. Grimme, "Extended tight-binding quantum chemistry methods," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **11**, e1493 (2021).
[40]P. Pracht and S. Grimme, "Calculation of absolute molecular entropies and heat capacities made simple," Chem. Sci. **12**, 6551–6568 (2021).
[41]J. Gorges, S. Grimme, A. Hansen, and P. Pracht, "Towards understanding solvation effects on the conformational entropy of non-rigid molecules," Phys. Chem. Chem. Phys. **24**, 12249–12259 (2022).
[42]S. Spicher, C. Plett, P. Pracht, A. Hansen, and S. Grimme, "Automated molecular cluster growing for explicit solvation by efficient force field and tight binding methods," J. Chem. Theory Comput. **18**, 3174–3189 (2022).
[43]C. Plett and S. Grimme, "Automated and efficient generation of general molecular aggregate structures," Angew. Chem., Int. Ed. **62**, e202214477 (2023).
[44]P. Pracht, C. A. Bauer, and S. Grimme, "Automated and efficient quantum chemical determination and energetic ranking of molecular protonation sites," J. Comput. Chem. **38**, 2618–2631 (2017).
[45]P. Pracht, D. F. Grant, and S. Grimme, "Comprehensive assessment of GFN tight-binding and composite density functional theory methods for calculating gas-phase infrared spectra," J. Chem. Theory Comput. **16**, 7044–7060 (2020).
[46]E. Karunaratne, D. W. Hill, P. Pracht, J. A. Gascón, S. Grimme, and D. F. Grant, "High-throughput non-targeted chemical structure identification using gas-phase infrared spectra," Anal. Chem. **93**, 10688–10696 (2021).
[47]P. Pracht and C. Bannwarth, "Fast screening of minimum energy crossing points with semiempirical tight-binding methods," J. Chem. Theory Comput. **18**, 6370–6385 (2022).
[48]P. Pracht and C. Bannwarth, "Finding excited-state minimum energy crossing points on a budget: Non-self-consistent tight-binding methods," J. Phys. Chem. Lett. **14**, 4440–4448 (2023).
[49]S. Grimme, C. Bannwarth, S. Dohm, A. Hansen, J. Pisarek, P. Pracht, J. Seibert, and F. Neese, "Fully automated quantum-chemistry-based computation of spin–spin-coupled nuclear magnetic resonance spectra," Angew. Chem., Int. Ed. **56**, 14763–14769 (2017).
[50]S. Grimme, "Exploration of chemical compound, conformer, and reaction space with meta-dynamics simulations based on tight-binding quantum chemical calculations," J. Chem. Theory Comput. **15**, 2847–2862 (2019).
[51]E. A. Coutsias, C. Seok, and K. A. Dill, "Using quaternions to calculate RMSD," J. Comput. Chem. **25**, 1849–1857 (2004).
[52]R. Leardi, "Genetic algorithms in chemometrics and chemistry: A review," J. Chemom. **15**, 559–569 (2001).
[53]xtb–Semiempirical Extended Tight-Binding Program Package, https://github.com/grimme-lab/xtb, 2024.
[54]P. Pracht, E. Caldeweyher, S. Ehlert, and S. Grimme, "A robust non-self-consistent tight-binding quantum chemistry method for large molecules," chemRxiv (2019).
[55]S. Grimme, C. Bannwarth, and P. Shushkov, "A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements ($Z = 1 - 86$)," J. Chem. Theory Comput. **13**, 1989–2009 (2017).

[56]C. Bannwarth, S. Ehlert, and S. Grimme, "GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions," J. Chem. Theory Comput. **15**, 1652–1671 (2019).
[57]S. Spicher and S. Grimme, "Robust atomistic modeling of materials, organometallic, and biochemical systems," Angew. Chem., Int. Ed. **132**, 15795–15803 (2020).
[58]L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi, and S. Grimme, "A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions," Phys. Chem. Chem. Phys. **19**, 32184–32215 (2017).
[59]M. Müller, A. Hansen, and S. Grimme, "ωB97X-3c: A composite range-separated hybrid DFT method with a molecule-optimized polarized valence double-ζ basis set," J. Chem. Phys. **158**, 014103 (2023).
[60]M. Bursch, A. Hansen, P. Pracht, J. T. Kohn, and S. Grimme, "Theoretical study on conformational energies of transition metal complexes," Phys. Chem. Chem. Phys. **23**, 287–299 (2021).
[61]S. Ehlert, S. Grimme, and A. Hansen, "Conformational energy benchmark for longer *n*-alkane chains," J. Phys. Chem. A **126**, 3521 (2022).
[62]H.-J. Werner and A. Hansen, "Accurate calculation of isomerization and conformational energies of larger molecules using explicitly correlated local coupled cluster methods in molpro and ORCA," J. Chem. Theory Comput. **19**, 7007–7030 (2023).
[63]D. I. Sharapa, A. Genaev, L. Cavallo, and Y. Minenkov, "A robust and cost-efficient scheme for accurate conformational energies of organic molecules," ChemPhysChem **20**, 92–102 (2019).
[64]J. Řezáč, D. Bím, O. Gutten, and L. Rulíšek, "Toward accurate conformational energies of smaller peptides and medium-sized macrocycles: MPCONF196 benchmark energy data set," J. Chem. Theory Comput. **14**, 1254–1266 (2018).
[65]C. Plett, S. Grimme, and A. Hansen, "Conformational energies of biomolecules in solution: Extending the MPCONF196 benchmark with explicit water molecules," J. Comput. Chem. **45**, 419 (2023).
[66]M. Marianski, A. Supady, T. Ingram, M. Schneider, and C. Baldauf, "Assessing the accuracy of across-the-scale methods for predicting carbohydrate conformational energies for the examples of glucose and α-maltose," J. Chem. Theory Comput. **12**, 6157–6168 (2016).
[67]C. Lee, W. Yang, and R. G. Parr, "Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density," Phys. Rev. B **37**, 785–789 (1988).
[68]A. D. Becke, "Density-functional thermochemistry. III. The role of exact exchange," J. Chem. Phys. **98**, 5648–5652 (1993).
[69]P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch, "Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields," J. Phys. Chem. **98**, 11623–11627 (1994).
[70]E. Caldeweyher, S. Ehlert, A. Hansen, H. Neugebauer, S. Spicher, C. Bannwarth, and S. Grimme, "A generally applicable atomic-charge dependent London dispersion correction," J. Chem. Phys. **150**, 154122 (2019).
[71]N. Mardirossian and M. Head-Gordon, "ωB97X-V: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy," Phys. Chem. Chem. Phys. **16**, 9904–9924 (2014).
[72]F. Weigend and R. Ahlrichs, "Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy," Phys. Chem. Chem. Phys. **7**, 3297–3305 (2005).
[73]J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general amber force field," J. Comput. Chem. **25**, 1157–1174 (2004).
[74]M. Gaus, Q. Cui, and M. Elstner, "DFTB3: Extension of the self-consistent-charge density-functional tight-binding method (SCC-DFTB)," J. Chem. Theory Comput. **7**, 931–948 (2011).
[75]M. Gaus, A. Goez, and M. Elstner, "Parametrization and benchmark of DFTB3 for organic molecules," J. Chem. Theory Comput. **9**, 338–354 (2013).
[76]M. Gaus, X. Lu, M. Elstner, and Q. Cui, "Parameterization of DFTB3/3OB for sulfur and phosphorus for chemical and biological applications," J. Chem. Theory Comput. **10**, 1518–1537 (2014).
[77]M. Kubillus, T. Kubař, M. Gaus, J. Řezáč, and M. Elstner, "Parameterization of the DFTB3 method for Br, Ca, Cl, F, I, K, and Na in organic and biological systems," J. Chem. Theory Comput. **11**, 332–342 (2015).
[78]J. J. P. Stewart, "Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements," J. Mol. Model. **13**, 1173 (2007).
[79]J. Řezáč and P. Hobza, "Advanced corrections of hydrogen bonding and dispersion for semiempirical quantum mechanical methods," J. Chem. Theory Comput. **8**, 141–151 (2012).
[80]P. S Brahmkshatriya, P. Dobes, J. Fanfrlik, J. Rezac, K. Paruch, A. Bronowska, M. Lepsík, P. Hobza, and P. Hobza, "Quantum mechanical scoring: Structural and energetic insights into cyclin-dependent kinase 2 inhibition by pyrazolo[1,5-a]pyrimidines," Curr. Comput.-Aided Drug Des. **9**, 118–129 (2013).
[81]R. Sure and S. Grimme, "Corrected small basis set Hartree-Fock method for large systems," J. Comput. Chem. **34**, 1672–1685 (2013).
[82]S. Grimme, A. Hansen, S. Ehlert, and J.-M. Mewes, "r2SCAN-3c: A 'Swiss army knife' composite electronic-structure method," J. Chem. Phys. **154**, 064103 (2021).
[83]L. Chan, G. Morris, and G. Hutchison, "Understanding conformational entropy in small molecules," J. Chem. Theory Comput. **17**, 2099–2106 (2021).
[84]D. L. Folmsbee, D. R. Koes, and G. R. Hutchison, "Systematic comparison of experimental crystallographic geometries and gas-phase computed conformers for torsion preferences," J. Chem. Inf. Model. **63**, 7401–7411 (2023).
[85]M. Bursch, E. Caldeweyher, A. Hansen, H. Neugebauer, S. Ehlert, and S. Grimme, "Understanding and quantifying London dispersion effects in organometallic complexes," Acc. Chem. Res. **52**, 258–266 (2019).
[86]H. W. Kuhn, "The Hungarian method for the assignment problem," Nav. Res. Logist. Q. **2**, 83–97 (1955).
[87]A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," Phys. Rev. B **87**, 184115 (2013).
[88]G. Ferré, J.-B. Maillet, and G. Stoltz, "Permutation-invariant distance between atomic configurations," J. Chem. Phys. **143**, 104114 (2015).
[89]M. Griffiths, S. P. Niblett, and D. J. Wales, "Optimal alignment of structures for finite and periodic systems," J. Chem. Theory Comput. **13**, 4914–4931 (2017).
[90]S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, "Comparing molecules and solids across structural and alchemical space," Phys. Chem. Chem. Phys. **18**, 13754–13769 (2016).
[91]J. Shao, S. W. Tanner, N. Thompson, and T. E. Cheatham, "Clustering molecular dynamics trajectories: 1. Characterizing the performance of different clustering algorithms," J. Chem. Theory Comput. **3**, 2312–2334 (2007).
[92]D. Bara, E. G. Meekel, I. Pakamorė, C. Wilson, S. Ling, and R. S. Forgan, "Exploring and expanding the Fe-terephthalate metal–organic framework phase space by coordination and oxidation modulation," Mater. Horiz. **8**, 3377–3386 (2021).
[93]P. Horcajada, T. Chalati, C. Serre, B. Gillet, C. Sebrie, T. Baati, J. F. Eubank, D. Heurtaux, P. Clayette, C. Kreuz, J.-S. Chang, Y. K. Hwang, V. Marsaud, P.-N. Bories, L. Cynober, S. Gil, G. Férey, P. Couvreur, and R. Gref, "Porous metal–organic-framework nanoscale carriers as a potential platform for drug delivery and imaging," Nat. Mater. **9**, 172–178 (2010).
[94]S. Spicher, D. Abdullin, S. Grimme, and O. Schiemann, "Modeling of spin–spin distance distributions for nitroxide labeled biomacromolecules," Phys. Chem. Chem. Phys. **22**, 24282–24290 (2020).
[95]S. Spicher, M. Bursch, and S. Grimme, "Efficient calculation of small molecule binding in metal–organic frameworks and porous organic cages," J. Phys. Chem. C **124**, 27529–27541 (2020).
[96]T. Stauch, "A mechanochemical model for the simulation of molecules and molecular crystals under hydrostatic pressure," J. Chem. Phys. **153**, 134503 (2020).
[97]M. Karplus and J. N. Kushick, "Method for estimating the configurational entropy of macromolecules," Macromolecules **14**, 325–332 (1981).
[98]H. W. T. van Vlijmen, M. Schaefer, and M. Karplus, "Improving the accuracy of protein pKa calculations: Conformational averaging versus the average structure," Proteins: Struct., Funct., Bioinf. **33**, 145–158 (1998).
[99]D. Suárez and N. Díaz, "Direct methods for computing single-molecule entropies from molecular simulations," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **5**, 1–26 (2015).

[100]M. S. Head, J. A. Given, and M. K. Gilson, "'Mining minima': Direct computation of conformational free energy," J. Phys. Chem. A **101**, 1609–1618 (1997).
[101]C.-E. Chang, M. J. Potter, and M. K. Gilson, "Calculation of molecular configuration integrals," J. Phys. Chem. B **107**, 1048–1055 (2003).
[102]F. Jensen, *Introduction to Computational Chemistry* (Wiley, Chichester, UK, 2017).
[103]G. P. Pereira and M. Cecchini, "Multibasin quasi-harmonic approach for the calculation of the configurational entropy of small molecules in solution," J. Chem. Theory Comput. **17**, 1133–1142 (2021).
[104]E. Suárez, N. Díaz, and D. Suárez, "Entropy calculations of single molecules by combining the rigid-rotor and harmonic-oscillator approximations with conformational entropy estimations from molecular dynamics simulations," J. Chem. Theory Comput. **7**, 2638–2653 (2011).
[105]E. Suárez, N. Díaz, J. Méndez, and D. Suárez, "CENCALC: A computational tool for conformational entropy calculations from molecular simulations," J. Comput. Chem. **34**, 2041–2054 (2013).
[106]N. Díaz and D. Suárez, "Toward reliable and insightful entropy calculations on flexible molecules," J. Chem. Theory Comput. **18**, 7166–7178 (2022).
[107]J. Kästner, "Umbrella sampling," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **1**, 932–942 (2011).
[108]S. Grimme, "Supramolecular binding thermodynamics by dispersion-corrected density functional theory," Chem. - Eur. J. **18**, 9955–9964 (2012).
[109]K. A. Levenberg, "A method for the solution of certain non-linear problems in least squares," Q. Appl. Math. **2**, 164–168 (1944).
[110]D. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," J. Soc. Ind. Appl. Math. **11**, 431–441 (1963).
[111]J. G. Brandenburg, C. Bannwarth, A. Hansen, and S. Grimme, "B97-3c: A revised low-cost variant of the B97-D density functional method," J. Chem. Phys. **148**, 064104 (2018).
[112]Y.-P. Li, A. T. Bell, and M. Head-Gordon, "Thermodynamics of anharmonic systems: Uncoupled mode approximations for molecules," J. Chem. Theory Comput. **12**, 2861–2870 (2016).
[113]M. Steiner, T. Holzknecht, M. Schauperl, and M. Podewitz, "Quantum chemical microsolvation by automated water placement," Molecules **26**, 1793 (2021).
[114]G. N. Simm, P. L. Türtscher, and M. Reiher, "Systematic microsolvation approach with a cluster-continuum scheme and conformational sampling," J. Comput. Chem. **41**, 1144–1155 (2020).
[115]J. R. Pliego, Jr. and J. M. Riveros, "Hybrid discrete-continuum solvation methods," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **10**, e1440 (2020).
[116]S. Grimme, C. Bannwarth, E. Caldeweyher, J. Pisarek, and A. Hansen, "A general intermolecular force field based on tight-binding quantum chemical calculations," J. Chem. Phys. **147**, 161708 (2017).
[117]S. Roehrig, A. Straub, J. Pohlmann, T. Lampe, J. Pernerstorfer, K.-H. Schlemmer, P. Reinemer, and E. Perzborn, "Discovery of the novel antithrombotic agent 5-chloro-*N*-({(5*S*)-2-oxo-3- [4-(3-oxomorpholin-4-yl)phenyl]-1,3-oxazolidin-5-yl}methyl)thiophene- 2-carboxamide (BAY 59-7939): An oral, direct factor Xa inhibitor," J. Med. Chem. **48**, 5900–5908 (2005).
[118]B. Hégely and M. Kállay, "Multilevel approach to the initial guess for self-consistent field calculations," Int. J. Quantum Chem. **122**, e26782 (2022).
[119]S. A. Grimmel and M. Reiher, "The electrostatic potential as a descriptor for the protonation propensity in automated exploration of reaction mechanisms," Faraday Discuss. **220**, 443–463 (2019).
[120]P. Pracht, R. Wilcken, A. Udvarhelyi, S. Rodde, and S. Grimme, "High accuracy quantum-chemistry-based calculation and blind prediction of macroscopic pKa values in the context of the SAMPL6 challenge," J. Comput.-Aided Mol. Des. **32**, 1139–1149 (2018).
[121]P. Pracht and S. Grimme, "Efficient quantum-chemical calculations of acid dissociation constants from free-energy relationships," J. Phys. Chem. A **125**, 5681–5692 (2021).
[122]O. Abarbanel and G. Hutchison, "QupKake: Integrating machine learning and quantum chemistry for micro-pKa predictions," chemRxiv (2023).
[123]J. P. Unsleber and M. Reiher, "The exploration of chemical reaction networks," Annu. Rev. Phys. Chem. **71**, 121–142 (2020).
[124]R. W. Floyd, "Algorithm 97: Shortest path," Commun. ACM **5**, 345 (1962).
[125]S. Grimme, "Towards first principles calculation of electron impact mass spectra of molecules," Angew. Chem., Int. Ed. **52**, 6306–6312 (2013).
[126]N. van Staalduinen and C. Bannwarth, ChemRxiv (2024).
[127]A. O. Lykhin, D. S. Kaliakin, G. E. dePolo, A. A. Kuzubov, and S. A. Varganov, "Nonadiabatic transition state theory: Application to intersystem crossings in the active sites of metal-sulfur proteins," Int. J. Quantum Chem. **116**, 750–761 (2016).
[128]M. Inamori, Y. Ikabata, T. Yoshikawa, and H. Nakai, "Unveiling controlling factors of the S/S_1 minimum energy conical intersection (2): Application to penalty function method," J. Chem. Phys. **152**, 144108 (2020).
[129]D. R. Yarkony, "Marching along ridges. Efficient location of energy-minimized conical intersections of two states using extrapolatable functions," J. Phys. Chem. A **108**, 3200–3205 (2004).
[130]B. G. Levine, J. D. Coe, and T. J. Martínez, "Optimizing conical intersections without derivative coupling vectors: Application to multistate multireference second-order perturbation theory (MS-CASPT2)," J. Phys. Chem. B **112**, 405–413 (2008).
[131]J. O. Lindner, K. Sultangaleeva, M. I. S. Röhr, and R. Mitrić, "metaFALCON: A program package for automatic sampling of conical intersection seams using multistate metadynamics," J. Chem. Theory Comput. **15**, 3450–3460 (2019).
[132]E. Pieri, D. Lahana, A. M. Chang, C. R. Aldaz, K. C. Thompson, and T. J. Martínez, "The non-adiabatic nanoreactor: Towards the automated discovery of photochemistry," Chem. Sci. **12**, 7294–7307 (2021).
[133]C. Bannwarth, J. K. Yu, E. G. Hohenstein, and T. J. Martínez, "Hole–hole Tamm–Dancoff-approximated density functional theory: A highly efficient electronic structure method incorporating dynamic and static correlation," J. Chem. Phys. **153**, 024110 (2020).
[134]F. Neese, "The ORCA program system," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **2**, 73–78 (2012).
[135]F. Neese, F. Wennmohs, U. Becker, and C. Riplinger, "The ORCA quantum chemistry program package," J. Chem. Phys. **152**, 224108 (2020).
[136]S. Grimme, J. G. Brandenburg, C. Bannwarth, and A. Hansen, "Consistent structures and interactions by density functional theory with small atomic orbital basis sets," J. Chem. Phys. **143**, 054107 (2015).
[137]R. J. Kutta, J. Großkopf, N. van Staalduinen, A. Seitz, P. Pracht, S. Breitenlechner, C. Bannwarth, P. Nuernberger, and T. Bach, "Multifaceted view on the mechanism of a photochemical deracemization reaction," J. Am. Chem. Soc. **145**, 2354–2363 (2023).
[138]toml-f–TOML parser implementation for data serialization and deserialization in Fortran, https://github.com/toml-f/toml-f, 2024.
[139]tblite–A light-weight tight-binding framework, https://github.com/tblite/tblite, 2024.
[140]S. Ehlert, M. Stahn, S. Spicher, and S. Grimme, "Robust and efficient implicit solvation model for fast semiempirical methods," J. Chem. Theory Comput. **17**, 4250–4261 (2021).
[141]GFN0-xTB standalone implementation, https://github.com/pprcht/gfn0, 2024.
[142]GFN-FF standalone implementation, https://github.com/pprcht/gfnff, 2024.
[143]P. Pyykkö and M. Atsumi, "Molecular single-bond covalent radii for elements 1–118," Chem. - Eur. J. **15**, 186–197 (2009).
[144]V. I. Lebedev and D. Laikov, "A quadrature formula for the sphere of the 131st algebraic order of accuracy," Dokl. Math. **59**, 477–481 (1999).
[145]R. M. Hazen, H. K. Mao, L. W. Finger, and P. M. Bell, "Structure and compression of crystalline methane at high pressure and room temperature," Appl. Phys. Lett. **37**, 288–289 (1980).
[146]D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications* (Elsevier, 2023).
[147]H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," J. Chem. Phys. **81**, 3684–3690 (1984).
[148]H. C. Andersen, "Rattle: A 'velocity' version of the shake algorithm for molecular dynamics calculations," J. Comput. Phys. **52**, 24–34 (1983).
[149]C. W. Hopkins, S. Le Grand, R. C. Walker, and A. E. Roitberg, "Long-time-step molecular dynamics through hydrogen mass repartitioning," J. Chem. Theory Comput. **11**, 1864–1874 (2015).

[150]A. Laio and F. L. Gervasio, "Metadynamics: A method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science," Rep. Prog. Phys. **71**, 126601 (2008).
[151]A. Barducci, M. Bonomi, and M. Parrinello, "Metadynamics," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **1**, 826–843 (2011).
[152]L. Sutto, S. Marsili, and F. L. Gervasio, "New advances in metadynamics," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **2**, 771–779 (2012).
[153]D. Branduardi, F. L. Gervasio, and M. Parrinello, "From A to B in free energy space," J. Chem. Phys. **126**, 054103 (2007).
[154]S. Grimme, "ANCOPT: Approximate Normal Coordinate Rational Function Optimization Program," Universität Bonn, Bonn, Germany (2013).
[155]A. Banerjee, N. Adams, J. Simons, and R. Shepard, "Search for stationary points on surfaces," J. Phys. Chem. **89**, 52–57 (1985).
[156]J. Baker, "An algorithm for the location of transition states," J. Comput. Chem. **7**, 385–395 (1986).
[157]F. Eckert, P. Pulay, and H.-J. Werner, "*Ab initio* geometry optimization for large molecules," J. Comput. Chem. **18**, 1473–1483 (1997).
[158]H. B. Schlegel, "Geometry optimization," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **1**, 790–809 (2011).
[159]E. R. Davidson, "The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices," J. Comput. Phys. **17**, 87–94 (1975).
[160]C. G. Broyden, "The convergence of a class of double-rank minimization algorithms 1. General considerations," IMA J. Appl. Math. **6**, 76–90 (1970).
[161]R. Fletcher, "A new approach to variable metric algorithms," Comput. J. **13**, 317–322 (1970).
[162]D. Goldfarb, "A family of variable-metric methods derived by variational means," Math. Comput. **24**, 23–26 (1970).
[163]D. F. Shanno, "Conditioning of quasi-Newton methods for function minimization," Math. Comput. **24**, 647–656 (1970).
[164]R. Lindh, A. Bernhardsson, G. Karlström, and P.-Å. Malmqvist, "On the use of a Hessian model function in molecular geometry optimizations," Chem. Phys. Lett. **241**, 423–428 (1995).
[165]J. Baker, "Techniques for geometry optimization: A comparison of Cartesian and natural internal coordinates," J. Comput. Chem. **14**, 1085–1100 (1993).
[166]J. Nocedal, "Updating quasi-Newton matrices with limited storage," Math. Comput. **35**, 773–782 (1980).
[167]D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," Math. Program. **45**, 503–528 (1989).
[168]Software developments by D. J. Wales *et al.*, http://www-wales.ch.cam.ac.uk/software.html, 2024.
[169]J. N. Harvey, M. Aschi, H. Schwarz, and W. Koch, "The singlet and triplet states of phenyl cation. A hybrid approach for locating minimum energy crossing points between non-interacting potential energy surfaces," Theor. Chem. Acc. **99**, 95–99 (1998).
[170]J. N. Harvey and M. Aschi, "Spin-forbidden dehydrogenation of methoxy cation: A statistical view," Phys. Chem. Chem. Phys. **1**, 5555–5563 (1999).
[171]J. P. Malhado and J. T. Hynes, "Non-adiabatic transition probability dependence on conical intersection topography," J. Chem. Phys. **145**, 194104 (2016).
[172]P. Wesołowski, D. J. Wales, and P. Pracht, "A multilevel framework for analysis of protein folding involving disulphide bond formation," J. Phys. Chem. B (to be published).
[173]lwONIOM, A light-weight multi-center *n*-level ONIOM interface, https://github.com/crest-lab/lwoniom, 2024.
[174]L. W. Chung, W. M. C. Sameera, R. Ramozzi, A. J. Page, M. Hatanaka, G. P. Petrova, T. V. Harris, X. Li, Z. Ke, F. Liu, H.-B. Li, L. Ding, and K. Morokuma, "The ONIOM method and its applications," Chem. Rev. **115**, 5678–5796 (2015).
[175]L. W. Chung, H. Hirao, X. Li, and K. Morokuma, "The ONIOM method: Its foundation and applications to metalloenzymes and photobiology," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **2**, 327–350 (2012).
[176]H. M. Senn and W. Thiel, "QM/MM methods for biomolecular systems," Angew. Chem., Int. Ed. **48**, 1198–1229 (2009).
[177]S. Ahmadi, L. Barrios Herrera, M. Chehelamirani, J. Hostaš, S. Jalife, and D. R. Salahub, "Multiscale modeling of enzymes: QM-cluster, QM/MM, and QM/MM/MD: A tutorial review," Int. J. Quantum Chem. **118**, e25558 (2018).
[178]K.-S. Csizi and M. Reiher, "Universal QM/MM approaches for general nanoscale applications," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **13**, e1656 (2023).
[179]P. Seeber, S. Seidenath, J. Steinmetzer, and S. Gräfe, "Growing spicy ONIOMs: Extending and generalizing concepts of ONIOM and many body expansions," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **13**, e1644 (2023).
[180]M. A. Majewski, W. Stawski, J. M. Van Raden, M. Clarke, J. Hart, J. N. O'Shea, A. Saywell, and H. L. Anderson, "Covalent template-directed synthesis of a spoked 18-porphyrin nanoring," Angew. Chem., Int. Ed. **62**, e202302114 (2023).
[181]C. Plett, A. Katbashev, S. Ehlert, S. Grimme, and M. Bursch, "ONIOM meets *xtb*: Efficient, accurate, and robust multi-layer simulations across the periodic table," Phys. Chem. Chem. Phys. **25**, 17860–17868 (2023).
[182]P. Pracht, J. W. R. Morgan, and D. J. Wales, "Exploring energy landscapes for solid-state systems with variable cells at the extended tight-binding level," J. Chem. Phys. **159**, 064801 (2023).
[183]S. Grimme, M. Müller, and A. Hansen, "A non-self-consistent tight-binding electronic structure potential in a polarized double-ζ basis set for all *spd*-block elements up to Z = 86," J. Chem. Phys. **158**, 124111 (2023).
[184]M. Müller, A. Hansen, and S. Grimme, "An atom-in-molecule adaptive polarized valence single-ζ atomic orbital basis for electronic structure calculations," J. Chem. Phys. **159**, 164108 (2023).