JCTC Journal of Chemical Theory and Computation

Article

Cite This: J. Chem. Theory Comput. 2019, 15, 2847−2862

pubs.acs.org/JCTC

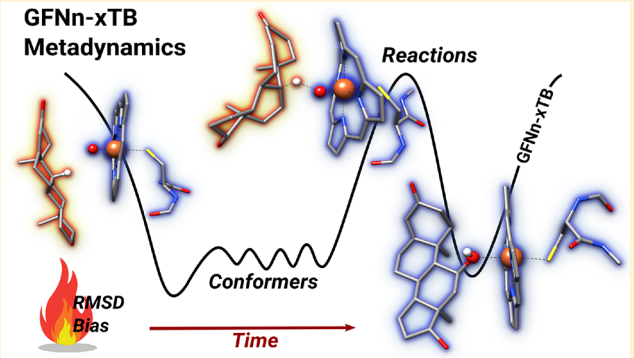
Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations

Stefan Grimme*

Mulliken Center for Theoretical Chemistry, Institute for Physical and Theoretical Chemistry, University of Bonn, Beringstrasse 4, 53115 Bonn, Germany

Supporting Information

ABSTRACT: The semiempirical tight-binding based quantum chemistry method GFN2-xTB is used in the framework of meta-dynamics (MTD) to globally explore chemical compound, conformer, and reaction space. The biasing potential given as a sum of Gaussian functions is expressed with the root-mean-square-deviation (RMSD) in Cartesian space as a metric for the collective variables. This choice makes the approach robust and generally applicable to three common problems (i.e., conformer search, chemical reaction space exploration in a virtual nanoreactor, and for guessing reaction paths). Because of the inherent locality of the atomic RMSD, functional group or fragment selective treatments are possible facilitating the investigation of catalytic processes where, for example, only the substrate is thermally activated. Due to the approximate character of the GFN2-xTB method, the resulting structure ensembles require further refinement with more sophisticated, for example, density functional or wave function theory methods. However, the approach is extremely efficient running routinely on common laptop computers in minutes to hours of computation time even for realistically sized molecules with a few hundred atoms. Furthermore, the underlying potential energy surface for molecules containing almost all elements (Z = 1−86) is globally consistent including the covalent dissociation process and electronically complicated situations in, for example, transition metal systems. As examples, thermal decomposition, ethyne oligomerization, the oxidation of hydrocarbons (by oxygen and a P450 enzyme model), a Miller-Urey model system, a thermally forbidden dimerization, and a multistep intramolecular cyclization reaction are shown. For typical conformational search problems of organic drug molecules, the new MTD(RMSD) algorithm yields lower energy structures and more complete conformer ensembles at reduced computational effort compared with its already well performing predecessor.



1. INTRODUCTION

A fast and reliable exploration and screening of the chemical compound and reaction space can complement many experimental studies and accelerate the computational prediction of novel reactions and functional compounds. General approaches for this purpose are preferably based on pure quantum chemistry (QC) methods which are opposed to most classical force-fields (FFs) able to break and form almost arbitrary types of bonds. Their inherent advantage over modern machine-learning (ML) models[1] is the ability to be applicable without a special training stage basically out-of-the-box to almost any chemical system. However, the major drawback of QC based reaction and compound space exploration (RCSE) tools is their high computational cost. In a pioneering work, Martinez and co-workers employed Hartree−Fock or DFT-based molecular dynamics (MD) simulations under high pressure and temperature conditions to build a so-called ab initio nanoreactor.[2] These inspiring simulations require thousands to millions of energy and force evaluations making the treatment of large systems (>50 atoms) on common computer hardware (e.g., laptops) routinely difficult. For further references on theoretical RCSE see the recent reviews from the Zimmermann[3] and Reiher[4] groups. Very early attempts for computer-assisted reaction planning and prediction were published by the group of Ugi already three decades ago.[5,6]

The basic idea presented here is to combine semiempirical tight-binding QC with root-mean-square-deviation (RMSD) based meta-dynamics for three important problems: (1) conformer search, (2) chemical compound (product) space exploration, and (3) finding reaction paths. As described in detail in the section 2, the same basic approach (biasing potential) is used in all three cases. We are aiming at automatic routine treatments running in at most hours of wall execution time on laptop computers for sizable molecules. The corresponding results will normally not be taken as the final answer to the respective problem but have exploratory and

Received: February 14, 2019
Published: April 3, 2019

screening character providing input for further (usually DFT) refinements. In this spirit, the proposed method complements existing methods for RCSE and modern reaction path finding algorithms (see, for example, refs 7 and 8). Moreover, the new conformer generation algorithm introduced here is more accurate and efficient than our recent multilevel MF-MD-GC procedure[9] and hence will replace it by default. It will be described and tested in greater detail in a separate publication.[10] For recent overviews on the practically very important conformational search problem which is still not generally solved for small peptides or macrocyclic drug compounds, see refs 11−13.

A highlight of the underlying QC approximation and hence of the entire approach presented here is the reliable treatment also of electronically complicated structures like, for example, reactive open-shell species or transition metal complexes. General force-fields for such systems are rare.[14,15] Two years ago we presented a semiempirical DFTB3[16,17] variant termed GFN-xTB, which mostly follows a global and element-specific parameters-only strategy and is consistently parametrized for all elements through radon.[18] The original purpose of the method and main target for the parameter optimization has been the computation of molecular geometries, vibrational frequencies, and noncovalent interaction energies. As such, it has been successfully used in structure optimizations of organometallic complexes[19,20] and structural sampling.[9,21−23] Apart from that, the method performed well in high-temperature molecular dynamics (MD) simulations of electron impact mass spectra.[24] Very recently, it is has been extended by including multipole electrostatic as well as one-center exchange-correlation terms leading to higher accuracy (at lower empiricism) specifically for noncovalent interactions and conformational energies.[25] This revised, so-called GFN2-xTB method is used throughout this work. In combination with the Fermi-smearing technique[26,27] at finite electronic temperature (typically a few thousand Kelvin), it enables the dissociation of covalent bonds as well as a qualitatively correct description of almost arbitrary open-shell electronic situations, which is a prerequisite for the construction of a general nanoreactor. The main drawback of both GFN methods is their limited accuracy for thermochemical properties, meaning that the potential energy surface (PES) is globally reasonable and consistent but can be partially distorted and too inaccurate to reveal fine details of reactions thus usually requiring additional higher-level treatments.

The use of the Cartesian RMSD as a collective variable in QC driven meta-dynamics has to the best of the author's knowledge not been investigated before. An RMSD based penalty potential and several variants of it are implemented in a common force-field MD driver[28] but seem to have escaped wider attention in the QC community. A somewhat related penalty function based on mass-weighted RMSD has been used by Levine et al.[29] for optimizing conical intersections. Coupled with FFs for finding protein conformations or describing drug-binding events on free energy surfaces, various RMSD type path collective variables have been discussed very recently by Hovan et al.[30] A method dubbed extended version of diffusion-map directed molecular dynamics (extended DM-d-MD) for decoupling slow and fast molecular motions with a weighted RMSD bias was proposed by the group of Clementi in the FF-based peptide folding context.[31,32] Already two decades ago, Schlitter et al.[33] introduced a FF-based method called targeted MD (TMD) for accelerating conformational transitions by applying a harmonic restrain on an RMSD variable with a moving restrain center (see also ref 34 for an overview).

Because of the enormous potential of the here-proposed method combination in many areas of chemistry, only selected, illustrative examples but no extensive benchmarking can be presented. The basic idea and theory is outlined in section 2 followed by examples for conformer searches of medium-sized and large organic molecules (with up to 176 atoms), decomposition as well as polymerization reactions, and a few typical reaction paths also involving transition metal complexes.

2. THEORY AND IMPLEMENTATION

The total energy E_{tot} of the system is the sum of the total (electronic) tight-binding QC energy E_{tot}^{el}, the biasing root-mean-square deviation (RMSD) potential E_{bias}^{RMSD}, and an optional reactor wall cavitation potential E_{bias}^{wall},

$$E_{tot} = E_{tot}^{el} + E_{bias}^{RMSD} + E_{bias}^{wall} \quad (1)$$

Further molecular constraining potentials, for example, to fix internal coordinates can be added of course and are for convenience contained in E_{tot}^{el}. The wall potential is normally used only for reaction space exploration in the nanoreactor for confinement and switched off for conformational or reaction path searches.

The same biasing potential E_{bias}^{RMSD} is used for all three purposes and is technically implemented for force and energy evaluations in the `xtb` code[35] (i.e., for optimization and MD runs in this work). In the dynamical context, the biasing potential is used in the spirit of so-called meta-dynamics[36] (MTD). The MTD algorithm is based on a description of the system by a set of collective variables (CVs) and the presence of a history-dependent potential that fills the minima of the PES over time to overcome large reaction barriers. More specifically the biasing potential used here is given by

$$E_{bias}^{RMSD} = \sum_{i=1}^{n} k_i \exp(-\alpha \Delta_i^2) \quad (2)$$

where n is the number of reference structures associated with the pushing ($k_i > 0$) or pulling ($k_i < 0$) strength k, Δ is the collective variable, and the parameter α determines the width (extension in space and time) of the biasing potential. In MTD runs, all k_i have the same positive values discouraging the system to come back to previous points. During the evolution of the simulation, more and more Gaussian potentials are summed up, preventing more and more the system to go back to previous real space coordinate regions. The frequency τ_{MTD} at which the potential is updated and the list of reference coordinates is enlarged by an actual MD snapshot is set by the user (typically 1−2 ps). Furthermore, a maximum number for n must be specified at which the oldest reference structure is removed from the list. In practice, duplicate reference structures do not appear in an MD and hence no structure comparison is conducted by default.

Related to MTD for exploring PES and in general for solving global optimization problems is the already mentioned TMD, and the so-called tabu search method which similarly employs lists of already visited structures (see, for example, ref 37). For predicting unimolecular reaction paths, a MTD related method termed "chemical flooding" that uses selected normal mode coordinates for the CVs was proposed by Müller et al.[38] This

DOI: 10.1021/acs.jctc.9b00143
J. Chem. Theory Comput. 2019, 15, 2847−2862

approach has been applied in FF-based conformational search for peptides by Ming et al.[39]

The choice of the CVs in MTDs is critical.[40] Here it is proposed to simply employ the standard root-mean-square deviation (RMSD) in Cartesian space as a metric. The RMSD Δ between the actual structure in a simulation and a reference structure in the above list is given by

$$\Delta_i = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(r_j - r_j^{\mathrm{ref},i})^2} \tag{3}$$

where r_j is a component of the Cartesian space vector of the actual molecule, $r_j^{\mathrm{ref},i}$ is the corresponding element in reference structure i (assuming the same atom numbering), and N is the number of atoms. For the optimal solid body transformation (rotation-translation) that minimizes the RMSD between the two sets of vectors requiring the alignment of the structures, the quaternion algorithm of Coutsias et al.[41] is employed, which additionally provides convenient access to the gradient of the RMSD as a function of the actual structure coordinates. The resulting $\frac{\partial \mathrm{RMSD}}{\partial r}$ terms are used to derive the atomic forces from the energy expression (eq 2). Note that the sum in eq 3 runs not necessarily over all atoms in the system thus allowing atom selective treatments. For example, the catalyst part in a reactive chemical system could be excluded from the RMSD computation which would push only the substrate away from the starting configuration to products. In section 4.2, the oxidation of hydrocarbons by a P450 enzyme model is modeled by this technique which, from the implementation point of view only requires setting up an atom list used for the RMSD computation.

Physically, the addition of each Gaussian potential increases the internal energy and, hence, in a MD run the temperature of the system. The RMSD (repulsive potential) increases (decreases) strongly for dissociation processes away from the reference structure, but usually, many strong covalent bonds prevent this thus avoiding trivial fragmentation to occur initially. Instead, lower-energy atomic movements like bending or torsion are typically first excited leading to a chemically realistic exploration of the PES by allowing otherwise impossible barrier crossings. If weakly bound complexes are treated, additional restraining potentials should be added to keep the fragments spatially close.

Instantaneous addition of a biasing potential in eq 2 can cause instabilities in the MD with large time steps due to extensive local heating of the system. This is avoided by making the potential additionally time dependent. The prefactor of a new Gaussian is multiplied with a damping function f_{dmp} given by

$$f_{\mathrm{dmp}} = \frac{2}{1 + \exp(-\kappa k)} - 1 \tag{4}$$

where k is an MD step counter and $\kappa = 0.03$ is an empirical parameter adjusted such that about 90% of the full strength is smoothly switched on within the subsequent 50−100 MD steps. The counter k is initialized to zero for each new Gaussian. In all runs, the initial start structure is included in the reference list.

Compared to other more complicated algorithms, the present approach does not require any predefinition of special system coordinates. In a rather democratic way, the RMSD-based potential allows all atoms to collectively explore unexplored regions of the PES which is, for example, a very important property for finding new, low-energy molecular conformations. The only practical requirement for stable MTDs is to keep the total energy within reasonable limits. Thus, all simulations are run with a standard algorithm in the NVT ensemble using the Berendsen thermostat[42] at a heat transfer time constant of 0.5 ps. For typical values of k_i (few mE_h per atom), the average system temperature in a steady state is then usually higher than the temperature of the heat bath (usually 300−400 K). The corresponding average values will be given in the examples.

The principles of the approach are schematically outlined on an example system (discussed later in detail) for which in Figure 1 the total energy as a function of time in an MTD(RMSD) simulation is shown. The heating as well as reaction events are clearly visible from the plot.

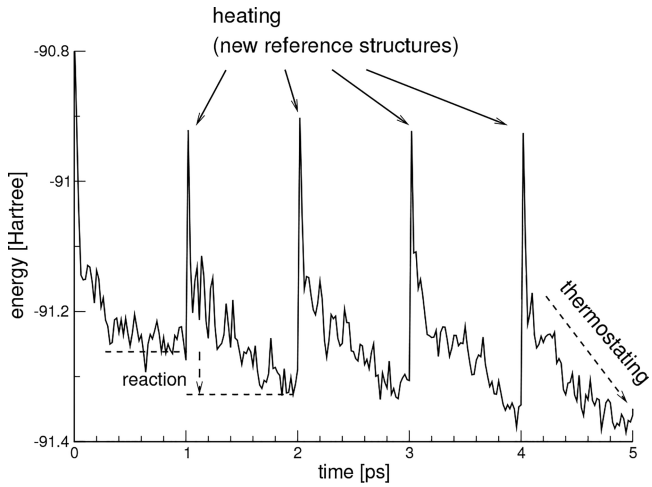


Figure 1. Energy as a function of time in a MTD(RMSD) simulation of the Miller-Urey system ($CO/H_2/H_2O/NH_3/CH_4$ mixture) as an example.

The extreme acceleration of conformational processes by effectively reducing the intramolecular barriers is illustrated in Figure 2, where for comparison snapshots along short 5 ps standard MD and MTD(RMSD) simulations for a long alkane chain are shown.

The efficient exploration of wide regions of the conformational PES by the MTD(RMSD) algorithm is clearly visible from this example. Note that the two simulations were conducted at approximately the same average temperature, indicating that the effect of the biasing potential is very different from a uniform heating. With very similar settings large scale molecular motions can also be induced in big systems. The protein Crambin with 642 atoms is used as an example to demonstrate the generality of the method (see Figure 3).

As can be seen from the plot of the structures, very different but still chemically realistic conformations and a complete helix-random coil transition can be obtained in relatively short simulation times, which are easily accessible with TB methods (the trajectory was completed in 7 h computation time on a quad-core laptop). This opens a route to a well-founded QM treatment of proteins and related flexible molecules and aggregates.

If the RMSD-biasing potential is applied in geometry optimization mode, estimates for low-energy reaction paths

DOI: 10.1021/acs.jctc.9b00143
J. Chem. Theory Comput. 2019, 15, 2847−2862

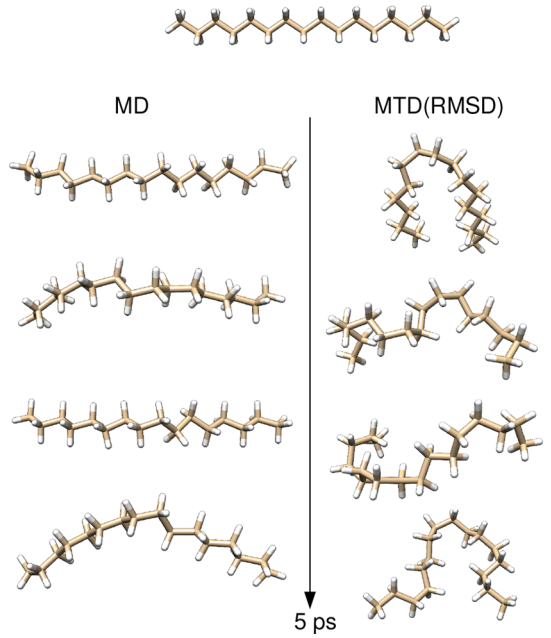


Figure 2. Structures of $C_{16}H_{34}$ conformational snapshots in two 5 ps MTD(RMSD)/GFN2-xTB simulations with $k_i/N = 0.002E_h$ and $\alpha = 0.8$ Bohr^{-1}. The average temperature in both runs was about 440 K.

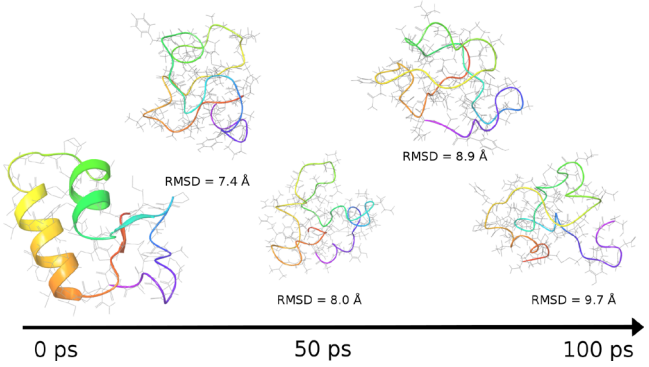


Figure 3. Structures of five Crambin snapshots in a 100 ps MTD(RMSD)/GNF0-xTB[43] simulation with $k_i/N = 0.002E_h$ and $\alpha = 0.3$ Bohr^{-1}. The average temperature was about 410 K. The RMSD values with respect to the optimized PDB start structure are also shown.

and transition states can be obtained conveniently. Here, we propose a two-point procedure where only the reactant (index one) and product (index two) reference structures are taken ($n = 2$). In a conventional geometry optimization starting from the reactant state, the system is pushed by the biasing potential from the starting point toward the product to which it is additionally pulled. Typically in such simulations, the pulling strength is absolutely larger than the pushing. The resulting crude path is refined by additional unconstrained and incomplete optimization steps at every point. The obtained estimates for transition states and intermediates should be further refined by other techniques. Note that this procedure is computationally only slightly more costly than a standard geometry optimization.

In nanoreactor simulations (or conformational search of noncovalently bound complexes) the molecules are restrained to move inside a spherical volume centered at the space-fixed origin by an appropriate time-independent wall potential[44]

$$E_{\text{bias}}^{\text{wall}} = k_{\text{wall}} \sum_{j=1}^{N} \log[1 + \exp(-\beta(R - r_j))] \tag{5}$$

where N is the number of atoms, R is the radius of the sphere, r_j are the atomic position vectors (initially transformed to the center-of-mass), and $\beta = 10$ Bohr^{-1} and $k_{\text{wall}} = 0.019E_h$ (corresponding to a temperature of 6000 K) are parameters determining the steepness and strength of the potential, respectively. This form has the advantage of a vanishing potential inside the cavity and yielding a constant force outside.

3. TECHNICAL DETAILS

All simulations were conducted with the in-house written `xtb` code.[35] The nanoreactor and paths simulations are initiated by the xtb code flags "-reactor" and "-path", respectively. The conformational search is organized by a separate program (see below) using xtb system calls with different MTD settings. For comparative DFT calculations, the TURBOMOLE suite of programs was used.[45,46] Problem specific technical details are given below. If computer timings are given, they mostly refer to a quad-core laptop with an Intel-i5−6300HQ CPU.

3.1. Conformer Search. General technical settings of the MTD(RMSD) and the previous MF-MD-GC procedure are the same and described in detail in ref 9. Both methods are implemented in a computer code named `crest` (conformer-rotamer ensemble sampling tool) available free for academic use and accessible through the xtb user list.[10] By default, the program enforces very tight optimization thresholds in the xtb code allowing one to distinguish between conformers and rotamers. In conformational search procedures, it is absolutely essential for any electronic structure or other atomistic method to provide numerically very accurate energy and gradients in order to be able to very tightly converge the structure optimizations. Otherwise, for larger systems, duplicate structures can not be identified, leading to wrong conformational thermostatistics and an overhead in subsequent DFT or WFT treatments. An important feature of `crest` is to find conformers as well as rotamers. A conformer belongs to a set of stereoisomers, each of which is characterized by a distinct energy minimum. Rotamers arise from restricted bond rotation (or other low-barrier motions like inversion), leading to an interchange of nuclei but to minima with identical energies. Enantiomers are special rotamer cases (mirror images). Rotamers contribute substantially to the molecular entropy, and the completeness of the overall conformation-rotation ensemble (CRE) can be assessed by a maximized entropy S_{CR} (or minimized free energy G_{CR}) according to the standard thermodynamic expressions

$$S_{\text{CR}} = R \sum_{i=1}^{\text{CRE}} p_i \log p_i \tag{6}$$

as a sum over all species found with population p_j

$$p_j = \frac{\exp(-\Delta E_j/\text{RT})}{\sum_{i=1}^{\text{CRE}} \exp(-\Delta E_i/\text{RT})} \tag{7}$$

and relative energy ΔE_j at absolute temperature T (R is the molar gas constant). The $G_{\text{CR}} = -TS_{\text{CR}}$ values (at $T = 298$ K) will be discussed in section 4.1. In the conformational MTD simulations, the SHAKE[47] algorithm for constraining all covalent bonds (based on standard covalent radii estimation)

DOI: 10.1021/acs.jctc.9b00143
J. Chem. Theory Comput. 2019, 15, 2847−2862

with an MD time step $d\tau$ of 5 fs is used to integrate the equations of motion. The atomic mass of hydrogen is set to 2 amu (deuterium). The MTD run time is dependent on an estimated flexibility of the molecule and typically is between $t = 0.5 \times N$ and N ps per run. The precise algorithm to estimate the required run time will be discussed in a forthcoming publication in which the conformational search procedure is described in more detail. In total, 12 independent MTDs with different settings for α in the range of 0.2−1.5 $Bohr^{-1}$ and k_{push} of 0.3−2.5mE_h are conducted from which equidistant snapshots are taken every 0.1 ps. The default biasing potential add-frequency τ_{MTD} is one per ps. Thus, for a typical drug molecule with 75 atoms, the total MTD time is ≈1 ns with about 1000 initially optimized snapshots. The technical parameters described above were obtained manually by trial and error in test calculations on molecules in the benchmark set and similar ones. The snapshots are geometry optimized in a multilevel, three-step-filtering procedure by first applying crude optimization threshold settings (option: `-opt crude` in the `xtb` code) followed by a tightly converged optimization run (option: `-opt tight`). The final ensemble is optimized very tightly (option: `-opt vtight`). In these three optimizations, all structures less than 18, 12, and 6 kcal/mol, respectively, above the lowest conformer (default energy windows) are taken into account. By this technique, the initial thousands of geometry optimizations are quickly completed, resulting typically only in hundreds of "costly" full optimizations to be conducted in the final step. The last value of 6 kcal/mol is conservatively chosen for a subsequent DFT treatment in which the lowest fully optimized conformers within this window would be further considered. Note that under standard experimental conditions only conformers in an energy window of 1.5−2 kcal/mol are relevant (i.e., have populations >3−5%). The computational bottleneck of the entire procedure is often the initial low-accuracy optimization of the MTD snapshots which is, however, running massively parallel. A standard thermostat bath temperature of 300 K is used in the MTD for cooling in order to avoid extensive rescaling of the atomic velocities. The average temperature in the conformational MTD is then typically 350−500 K. Elevated temperatures of about 400−500 K in the MD simulations were also found to be beneficial in the previous MF-MD-GC procedure. In order to improve ensemble completeness in particular for flexible systems with many alkyl chains, the previous genetic-crossing (GC) procedure[9] in automatically created Z-matrix coordinates is additionally executed. Here, the number of crossed and optimized structures is limited to $min(5000, t \times 50$ in ps). Additionally, 12 regular, unbiased MDs at 400 and 500 K running at half the time of the preceding MTD starting from the lowest six conformers obtained so far are included in order to increase the number of rotamers and low-barrier conformers which may have escaped the higher-energy MTD treatment. If in any of the above parts of the algorithm a new, lower lying conformer is found, the entire procedure is restarted with the first MTD steps. In this case, the hitherto generated ensembles are analyzed and merged. In this respect, the new algorithm is similarly iterative as the previous MF-MD-GC procedure. The advanatge of these restarts is mainly seen for large cases where the initial structure is often way off the global minimum. No solvation model was applied and in the comparison (see section 4.1), both algorithms were started from the same Cartesian coordinates taken mostly from the PubChem database.[48,49] If the 3D-structure was not available, the Open Babel code[50] was employed for 2D−3D conversion. The default GFN2-xTB electronic temperature of 300 K is applied in the Fermi-smearing procedure.

3.2. Reaction Path Estimates. By default three runs with increasing push/pulling strengths $[-k_2 = (2-4) \times k_1]$ at typical values of $\alpha = 0.5-1$ are conducted. Very tight optimization thresholds are applied in order to avoid trapping in spurious geometries. The subsequent incomplete optimizations without the biasing potential at every point of the path (typically 30−200) are limited to 2−4 geometry optimization steps such that a fall-back to the reactant/product is prevented. It is recommended that the user checks the resulting path for chemical reliability. In particular if the RMSD between reactant and product states is large, the reaction remains sometimes incomplete with the default settings and in such cases manually adjusted values for $k_1/k_2/\alpha$ should be applied. For thermally allowed reactions, the default GFN2-xTB electronic temperature of 300 K is used. The resulting paths are compared with those from an accurate GSM (growing string method[7]) treatment with 25−30 points on the path using exactly the same start-end structures as well as the same technical settings in GFN2-xTB.

3.3. Nanoreactor. The default settings for this run type are $n = 100$ with $k_i/N = 0.02$, $\alpha = 0.6$, an electronic temperature of 6000 K in GFN2-xTB, and a thermostat bath temperature of 298 K. The actual choice of the electronic temperature is not critical but should be guided by the expected biradical/multireference character in the system (i.e., higher for statically correlatedcases, e.g., covalent bond breakings). From the trajectory every tenth point is preoptimized and analyzed if a reaction occurred. These partially optimized structures are fully relaxed, further analyzed, and unique structures are written to a "product file" containing the final ensemble of individual product molecules. In order to allow any reaction, SHAKE is normally not used. Because locally high temperatures and large atomic velocities can occur, small MD time steps of $d\tau = 0.2-0.5$ fs are employed.

4. RESULTS AND DISCUSSION

4.1. Conformer Search. Before "real-life" conformational problems will be discussed, the advantage of the MTD(RMSD) algorithm is further illustrated on an interesting example continuing the already shown case of alkane folding in section 2. The alkane (alkyl chain) case is in some sense trivial because many weakly coupled, simple one-dimensional modes (torsions) only need to be permuted to generate most conformational minima (as done in typical chemo-informatics algorithms). However, the physics of a typical conformational process in large systems (particularly including rings) is often more involved through a complicated collective motion of many atoms occurring simultaneously. As an example, the well-known helix inversion (racemization) of so-called helicenes is considered. The barrier for the typical case of [6]helicene is about 35 kcal/mol[51] and can hardly be overcome in a conventional MD simulation. As shown in Figure 4, by applying standard values for the bias leads to the correct reaction event within only about 3 ps.

For procedures working in internal (predefined) coordinates this case may be already challenging while it seems rather trivial for the MTD(RMSD) approach because the collective atom torsion along the reaction coordinate strongly decreases the RMSD while other low-energy motions barely exist. More

DOI: 10.1021/acs.jctc.9b00143
J. Chem. Theory Comput. 2019, 15, 2847−2862

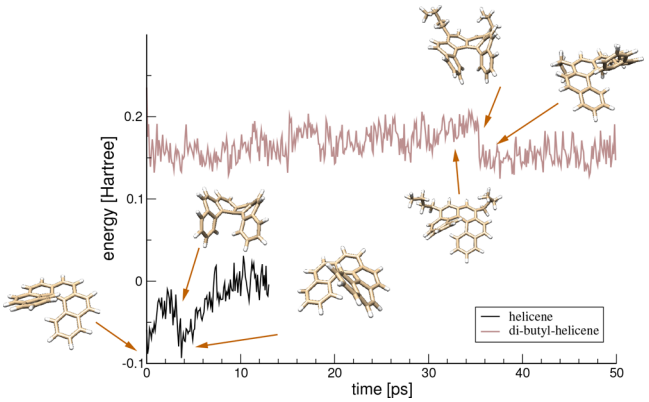


Figure 4. Energy as a function of time in a MTD(RMSD) simulation of [6]helicene and di-*n*-butyl[6]helicene with the same RMSD bias parameters (k_i/N = 0.002 E_h, α = 0.4 Bohr^{-1}). The energies are shifted to enable convenient visualization.

realistic are, however, cases in which such complicated movements occur simultaneously with other rather floppy and fast vibrational modes. In order to explore the limits of the present approach, we constructed a worst case model scenario by substituting the helicene in remote position by two *n*-butyl alkyl chains. The very positive result of this test is that also under these theoretically unfavorable conditions the racemization occurs (see the upper trace in Figure 4). In fact the reaction is merely slowed down by the presence of the alkyl chains meaning it does not occur before the biasing potential has filled up the many alkyl chain minima properly. That the algorithm works so well with a standard setup and without any definition of special coordinates for a very complicated but nevertheless realistic scenario is very encouraging. It can be hoped that even "exotic" global conformational minima in large systems can be found routinely by this algorithm.

Realistic example molecules for conformational searches are partially taken from our recent study[9] where the MF-MD-GC algorithm has been introduced. We consider the respective MF-MD-GC results employing the same underlying electronic structure method for the PES (GFN2-xTB) as a very reasonable and challenging comparison. The compiled diverse test set consists mainly of medium- to large-size organic drug molecules, but also two organometallic systems (a transition metal catalyst and a big adenosyl-cobalmin complex, entries 14 and 21) are included to show the generality of the method. Note that the largest system Vancomycin with almost 180 atoms is already at the upper size limit of what is of interest in medicinical chemistry and that some notoriously difficult (macro)cyclic cases are also included. The starting structures of the investigated molecules are shown in Figure 5.

In Table 1 computer timings and quality of the resulting conformer ensemble are compared for both algorithms based on the same GFN2-xTB gas phase PES. As a measure for completeness, the lowest total energy, the number of unique conformers in a 3 kcal/mol energy window, and the free conformer ensemble energies G_{CR}(298 K) are used. A good algorithm maximizes and minimizes the latter two quantities, respectively, and of course provides the correct global conformer minimum (which, however, is normally not known). Note that both algorithms have been tested extensively for correctness on simple textbook cases[52] regarding the global minimum where this information is available. Due to the partially stochastic nature of the algorithm (initialization of the MD velocities) together with parallel execution, the ensembles obtained from different runs with the same settings and starting structure are slightly different with typical deviations of about ±5−10% regarding the number of conformers found and the value of G_{CR}. It is recommended in large (>100 atoms) and critical cases to run the entire procedure at least two times to ensure that the results are conclusive. For further technical details, see the Supporting Information of ref 9.

Inspection of the results in Table 1 reveals the overall advantage of the new algorithm. In almost all cases the found ensemble is larger (very significantly in cases 9, 11, and 15) and the resulting conformer free energy is lower as indicated by negative values in column nine of Table 1. Mostly, the same lowest lying conformer is found with both methods giving confidence in the correctness of the global minimum structure. Note that a free energy comparison is only meaningful if both methods yield the same lowest-lying conformer. However, for eight molecules a lower-energy conformer is found with the new algorithm. In four cases (16, 20−22) is the new conformer found significantly lower in energy by about 3 kcal/mol or more. This means that it is unlikely that the correct one (at a higher DFT or WFT level) was included in the ensemble obtained from the old procedure. For Vancomycin (the largest system studied) and Adcobalmin$^+$, the huge energy gain of >10 kcal/mol indicates a complete failure of the old algorithm. Noteworthy is also that the difficult protein folding problem in the alanine decamer (15) is treated efficiently by MTD-(RMSD)-GC. Such inherently flexible molecules with potentially strong interactions like intramolecular hydrogen bonding represent a worst-case scenario with an entropically disfavored narrow funnel leading to the global minimum. Here, the lower (better) G_{CR} value wih the old method is rooted in a larger number of methyl group rotamers found while the more important number of true conformers is much better with the MTD(RMSD)-GC treatment.

The computation times with the new algorithm for the smaller cases are reduced by a factor of about 3−5, but the efficiency improvement of MTD(RMSD)-GC compared to MF-MD-GC at least partially holds with increasing system size. Speed-ups for some of the larger systems with MTD(RMSD)-GC can still be significant (e.g., for 13−14) but mostly the two algorithms are on par. This is impressive considering the much better results and that the old MF-MD-GC algorithm already outperformed standard methods.[9] For more than about 150 atoms, MTD(RMSD)-GC is slower by a factor of 2−3 but, however, thereby providing a significantly improved result. The starting and best structures found in the Supporting Information are reported also suggesting a challenge for other competitive search algorithms (which may call the `xtb` code for an on-the-fly computation of the GFN2-xTB PES).

One reason for inefficiency of the older procedure is that it is based on a normal-mode analysis of the lowest energy conformer. If the latter changes during the search, the calculation has to be restarted very often leading to a substantial computational overhead in cases where the starting structure is way off the global minimum. In the MTD approach, the number of restarts is reduced because the PES is globally explored and regions far away from the starting point are reached more quickly. Importantly, this is carried out in parallel, and the default number of 12 runs in each iteration with the subsequent optimizations are completely independent leading to an almost perfect speed-up with the number of

DOI: 10.1021/acs.jctc.9b00143
J. Chem. Theory Comput. 2019, 15, 2847−2862

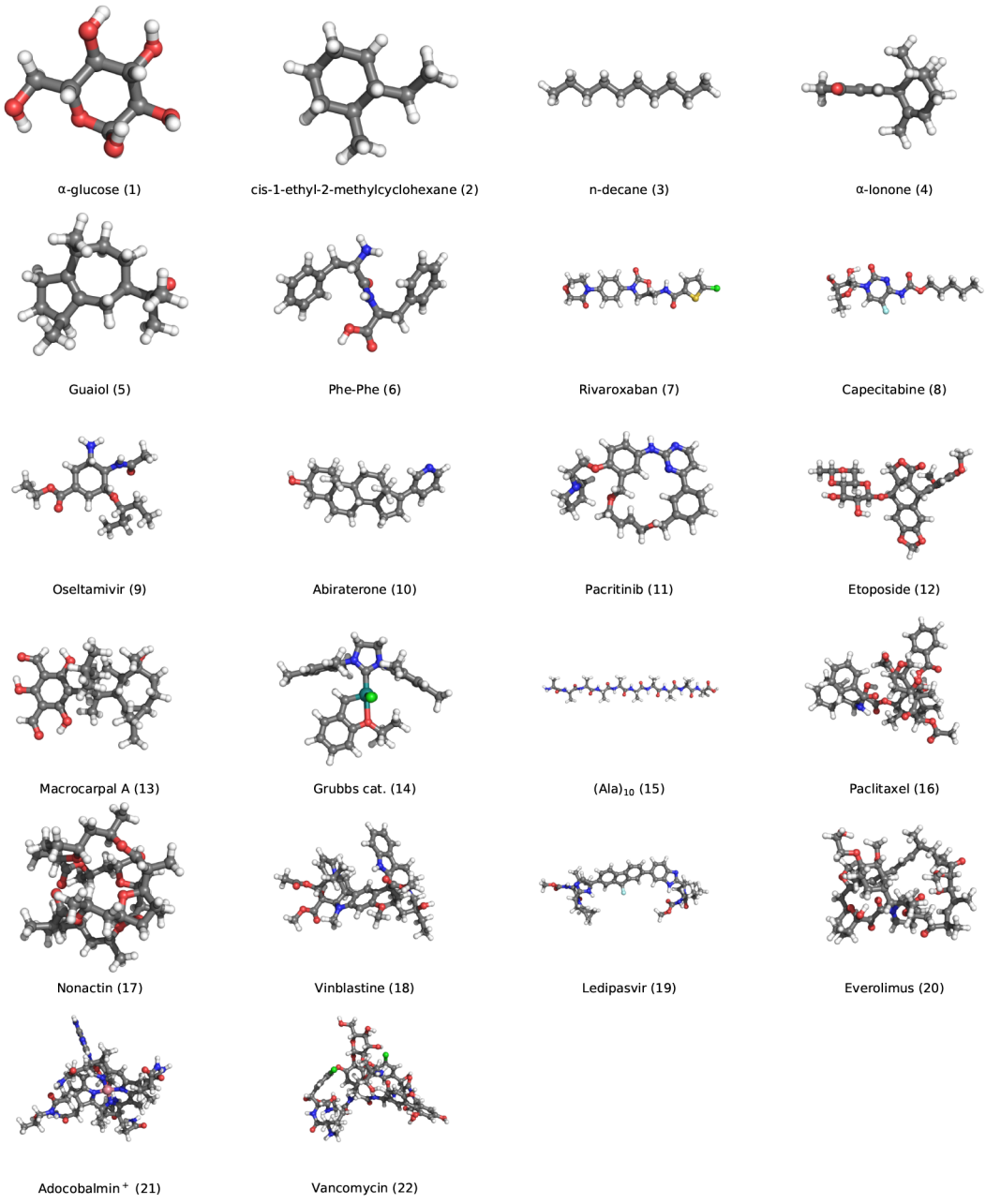


Figure 5. Molecular starting structures in the conformer search benchmark.

available processors. The search depth is user controlled by a single parameter (the length of the MD), and in this way the results can be systematically improved. Further details and more examples for the new MTD(RMSD)-GC algorithm will be described elsewhere. In particular, the advantage of the locality of the biasing potential will be further investigated there for cases like conformational search of transition states and for molecules on surfaces or in large cavities which require a chemically meaningful restriction of the dimensionality of the PES.

4.2. Nanoreactor. *4.2.1. Thermal Decomposition of Benzene and Ferrocene.* Thermal decomposition reactions are relatively easy to simulate because the RMSD bias pushes the reactants intrinsically to a more fragmented state since the RMSD strongly increases in such directions. If the radius of the spherical nanoreactor is small (high mass density ρ), however, fragmentation can mostly be avoided thus favoring more isomerizations. This is first demonstrated for benzene isomers which is a widely studied and well-known system.[54] Figure 6 shows optimized structures of the isomerization products resulting from a 100 ps long MTD(RMSD) simulation at a mass density of $\rho = 8$ g/cm^3 corresponding to a cavity radius of about 5 Bohr with $k_i/N = 0.04E_h$, $\alpha = 0.7$ Bohr^{-1}, and $\tau_{MTD} = 2$ ps.

The average temperature in the simulation was about 1700 K and yields various commonly known products like the Dewar (38), prismane (39), benzvalene (41), and bicyclopropenyl (7) valence isomers and other well-known species like fulvene (36) or hexa-1,3-diene-5-yne (8). Under these conditions, only 15% of the products are fragmented into

DOI: 10.1021/acs.jctc.9b00143
J. Chem. Theory Comput. 2019, 15, 2847−2862

Table 1. Conformational Search Results for Typical Organic Drugs and Two Transition Metal Systems[a]

			found[b]		lowest energy		G_{CR}		time (min:s)[c]	
	molecule	atoms	new	old	new[d]	δ(old)[e]	new[f]	δ(old)[g]	new	old
1	α-glucose	24	6	6	−43.36901	0.0	−0.62	−0.14	0:54	3:02
2	et-me-CH[h]	27	5	5	−28.48112	0.0	−2.04	−0.13	0:37	2:38
3	*n*-decane	32	154	112	−32.65145	0.0	−3.79	−0.51	3:08	3:31
4	α-Ionone	34	7	5	−42.08019	0.0	−3.06	0.13	1:30	6:05
5	Guaiol	42	33	18	−49.47445	0.0	−2.91	0.17	1:55	7:00
6	$(Phe)_2$	43	67	123	−66.79618	0.0	−2.75	−0.14	7:25	28:47
7	Rivaroxaban	47	25	18	−86.85147	0.0	−1.80	−0.22	7:28	20:55
8	Capecitabine	47	128	104	−81.33389	0.0	−3.00	0.18	8:01	18:36
9	Oseltamivir	50	283	72	−70.68494	0.0	−3.99	−0.80	13:45	29:58
10	Abiraterone	57	6	7	−74.21800	0.0	−2.11	−0.01	3:08	15:05
11	Pacritinb	67	180	87	−100.17834	0.0	−2.99	−0.67	28:25	58:02
12	Etoposide	74	67	56	−131.37186	−0.2	−3.23	−0.55	16:23	37:43
13	Macrocarpal	74	17	11	−104.81044	0.0	−3.08	0.57	14:04	45:20
14	Grubbs cat.	75	3	4	−108.13610	0.0	−4.18	−0.02	14:45	60:58
15	$(Ala)_{10}$	103	54	5	−164.98715	0.0	−3.51	0.59	246:47	218:31
16	Paclitaxel[I]	113	19	13	−186.57693	−3.1	−3.54	−0.48	100:30	156:52
17	Nonactin[j]	116	18	19	−167.31756	−0.5	−4.59	−3.10	193:13	104:24
18	Vinblastine	117	43	20	−176.54447	−0.6	−4.05	−0.35	110:23	161:41
19	Ledipasvir	119	82	38	−189.98552	−0.6	−3.69	−0.51	68:12	127:44
20	Everolimus	151	52	34	−215.73973	−6.6	−3.46	−1.50	701:10	181:29
21	Adcobalmin^{+}[k]	170	17	25	−262.29129	−14.6	−4.23	−2.05	767:28	230:24
22	Vancomycin	176	113	41	−312.85325	−19.5	−3.66	−1.89	1708:33	441:06

[a]A comparison of the previous MF-MD-GC search algorithm (old) with the newly proposed MTD(RMSD)-GC procedure (new) is given. [b]Number of distinct conformers found (excluding rotamers) in a 3 kcal/mol relative energy window. [c]Total wall computation time on a 80-core machine (Intel-Xeon-Gold-614 CPU). [d]Total energy of lowest conformer in E_h. [e]Energy difference in kcal/mol to the lowest conformer found with the new algorithm. Negative values indicate better performance of the new algorithm. [f]Free energy of conformer/rotamer ensemble at 298 K in kcal/mol. [g]Difference of G_{CR} value between the new and old algorithm in kcal/mol. Negative values indicate better performance of the new algorithm. [h]cis-1-Ethyl-2-methylcyclohexane. [I]Starting structure taken from ref 15. [j]Lowest conformer found in ref 9 used as starting structure. [k]Starting structure taken from ref 53.

two or more molecules or atoms. This behavior can readily be inverted by choosing a smaller mass density as simulation parameter. Some of the resulting isomers have very interesting structures featuring [e.g., inverted tetracoordinate carbon (23 and 27)]. It seems very encouraging that a single simulation provides about a quarter of all possible 218 benzene isomers without applying any prior knowledge and special input. The total computation time for this simulation including optimization of the products was 26 min on a quad-core laptop computer.

Guessing decomposition or isomerization products for benzene may be guided by good chemical intuition, but this likely fails for larger or more "exotic" compounds, and hence, a nanoreactor could be a great advantage in such cases. As a more challenging example including a transition metal, we investigate ferrocene which can result in electronically very complicated situations. It is shown, that even in such cases GFN2-xTB can be used without any modifications. A similar system setup as for benzene is used with slightly smaller bias due to the higher reactivity of ferrocene compared to benzene (ρ = 10 g/cm^3, $k_i/N = 0.01E_h$, α = 0.7 Bohr^{-1}, and τ_{MTD} = 2 ps) (see Figure 7).

The average temperature in the simulation was about 700 K and reveals various interesting isomers. Again, only a small amount of fragmentation products of about 10% is observed. As expected, no alternatives for lower coordinated Cp-rings are found. Instead, some highly coordinated Cp-ring isomeric structures of polycyclic and cage type are produced. Some rather exotic molecules (which are still within an energy window of only about 120 kcal/mol) are obtained showing the utility of this approach to widen the horizon of traditional chemical thinking. The total computation time for this simulation including optimization of the products was 39 min on a quad-core laptop computer.

4.2.2. Ethyne Polymerization. In ref 2, the authors investigated 39 ethyne molecules in a nanoreactor and observed spontaneous oligomerization leading to complicated hydrocarbons involving aromatic rings. The same system is studied here with the standard setup ($k_i/N = 0.02E_h$, α = 0.6 Bohr^{-1}), a mass density of approximately ρ = 2 g/cm^3, and an overall run time of 40 ps. A plot with the energy distribution of the reaction products (full stoichiometry) formed over time is shown in Figure 8.

The average temperature in the simulation was about 1800 K. The reaction started after 2 ps with a dimerization to an open C_4H_4 species. Subsequently, longer chains, three-membered rings and even aromatic compounds (benzene) are obtained. As intermediates, polyenes formed by hydrogen abstraction from ethyne are observed. The largest molecule formed after about 20 ps with sum formula $C_{66}H_{66}$ is qualitatively similar to that reported in ref 2, and overall, the product distribution (129 distinct species) seems to be chemically realistic. The total computation time for this simulation including complete optimization of all products was 12 h on a quad-core laptop (8 h for the MTD at a time step of 0.4 fs).

4.2.3. Oxidation of Cyclohexane. As an example for a reaction with multiple open-shell molecules, the oxidation of

DOI: 10.1021/acs.jctc.9b00143
J. Chem. Theory Comput. 2019, 15, 2847−2862

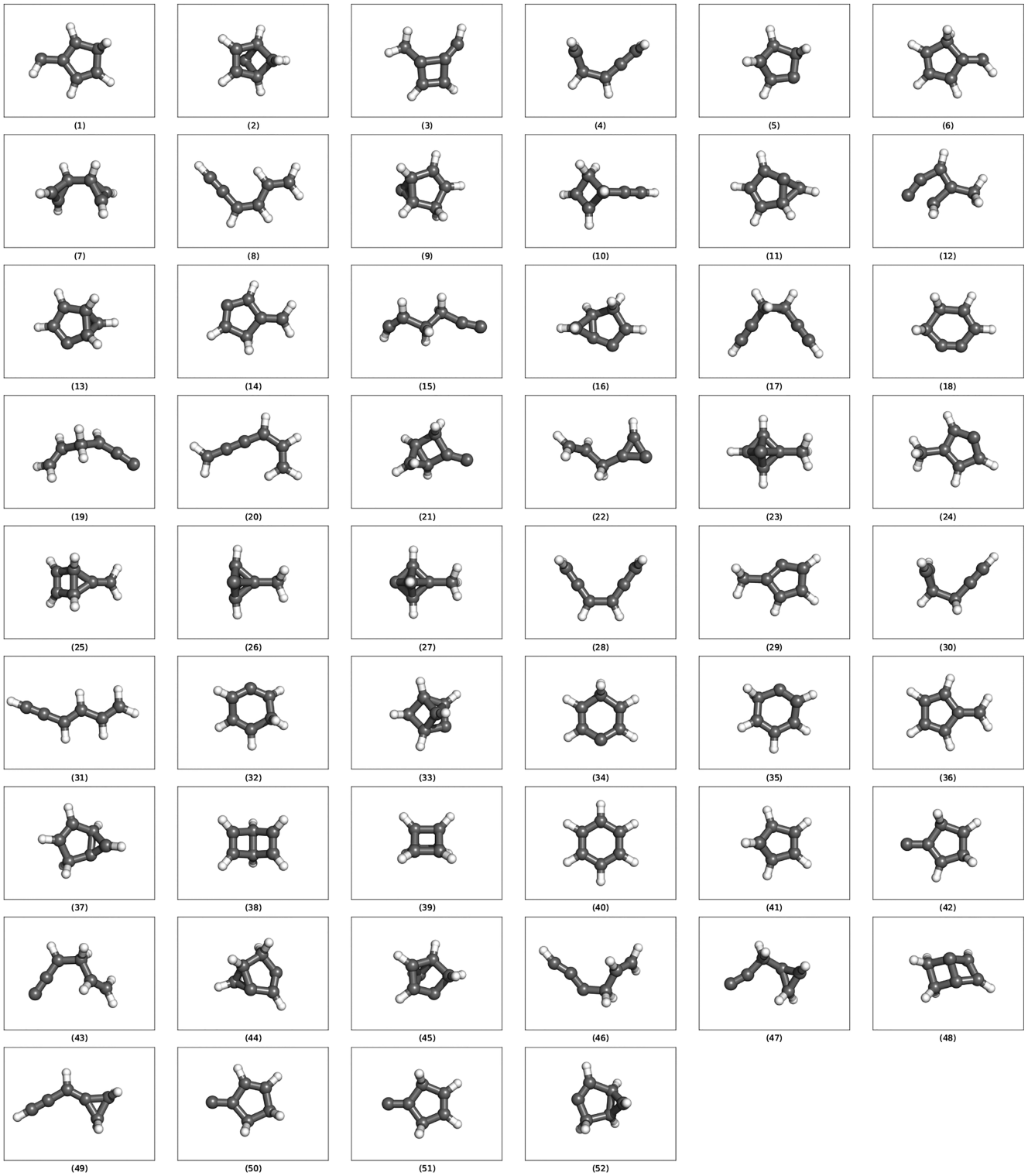


Figure 6. Automatically generated list of the 52 isomers found in the thermal decomposition of benzene (40).

cyclohexane by five dioxygen molecules is considered. The GFN2-xTB method (as GFN-xTB) does not properly distinguish between $\langle S^2 \rangle$ spin eigenstates and only conserves a total spin, which is $S = 5$ over the entire reaction. This means that spin-crossing is not allowed, and hence, some of the products are not in their electronic ground state, thus limiting the significance of the results to some extent. Nevertheless, it is shown that the simulation runs robustly also for such a complicated case which is difficult to treat "out-of-the-box" with other atomistic methods (for an ReaxFF treatment of hydrocarbon oxidation see ref 55). A plot with the energy distribution of the reaction products (entire system) as obtained from a 20 ps long MTD(RMSD) with parameters $\rho = 4$ g/cm^3, $k_i/N = 0.02$ E_h, $\alpha = 0.6$ Bohr^{-1} is shown in Figure 9.

DOI: 10.1021/acs.jctc.9b00143
J. Chem. Theory Comput. 2019, 15, 2847−2862

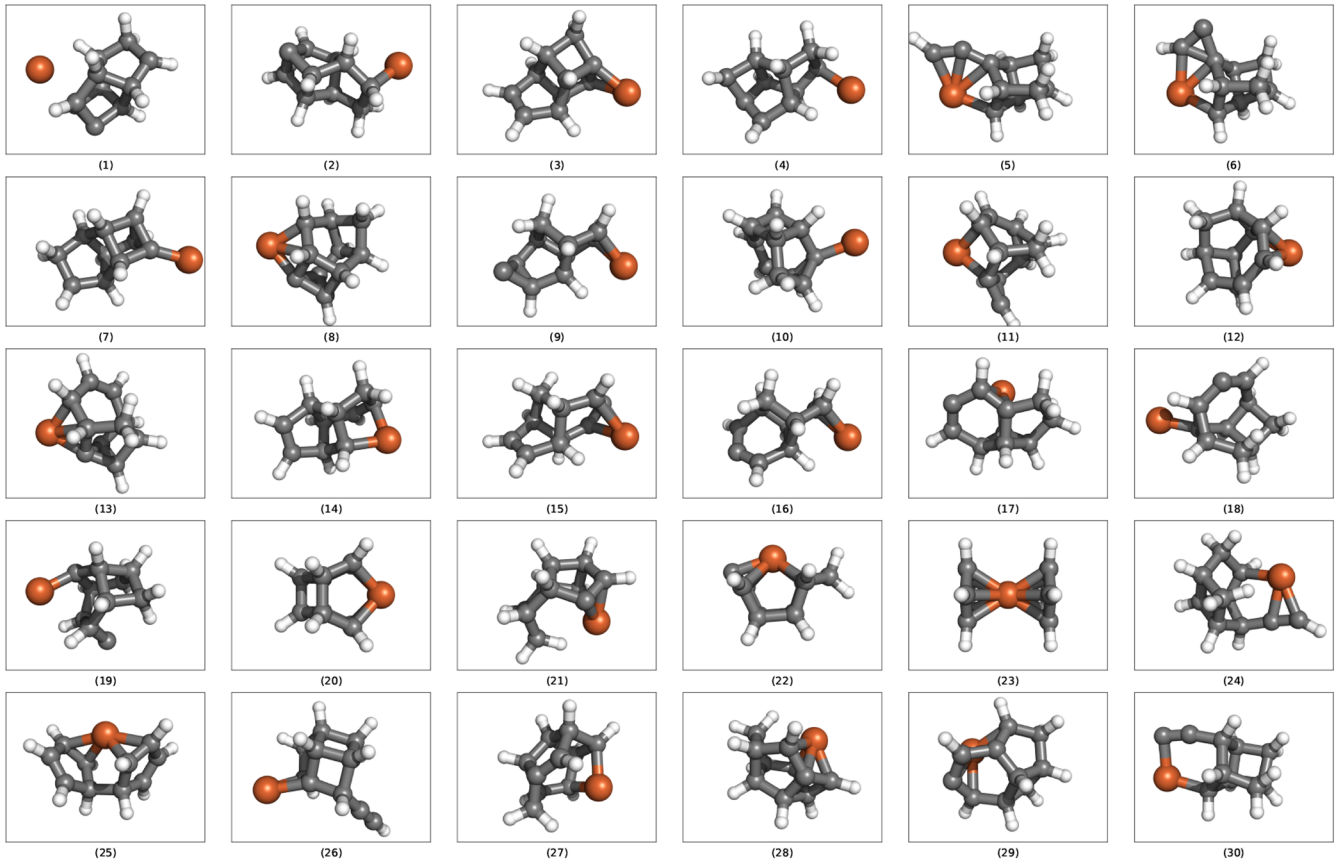


Figure 7. Automatically generated list of the 30 isomerization products found in the thermal decomposition of ferrocene (23).

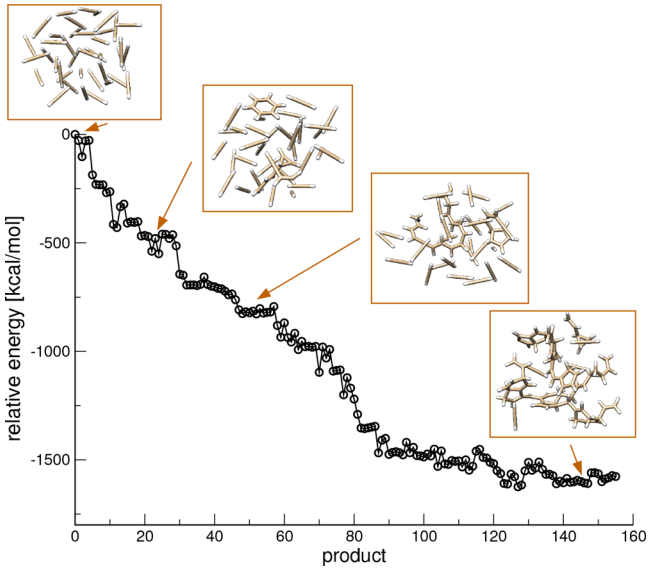


Figure 8. Energy distribution of reaction products obtained from snapshot optimization along the trajectory in the oligomerization of 39 ethyne molecules. The insets show some structures as examples.

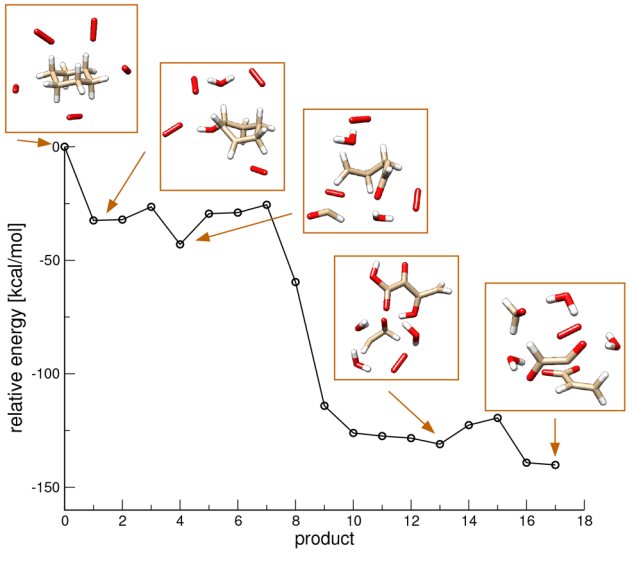


Figure 9. Energy distribution of reaction products obtained from snapshot optimization along the trajectory for the oxidation of cyclohexane with five $^3O_2$ molecules. The insets show some example structures.

The average temperature in the simulation was about 1700 K. The reaction started immediately by hydrogen abstraction and addition of OH to the respective carbon radical center. Further quickly appearing intermediates are water as well as hydrogen peroxide. After around 2 ps, the ring is opened leading to many radical species which are to be expected under these conditions. At later stages, common products with varying degree of oxidation like formaldehyde, the acrylic acid radical, or carbon dioxide are observed. Overall the simulation provides a chemically sensible degradation mechanism and meaningful reaction products. The total computation time for this simulation including complete optimization of all products was 12 min on a quad-core laptop computer.

DOI: 10.1021/acs.jctc.9b00143
J. Chem. Theory Comput. 2019, 15, 2847−2862

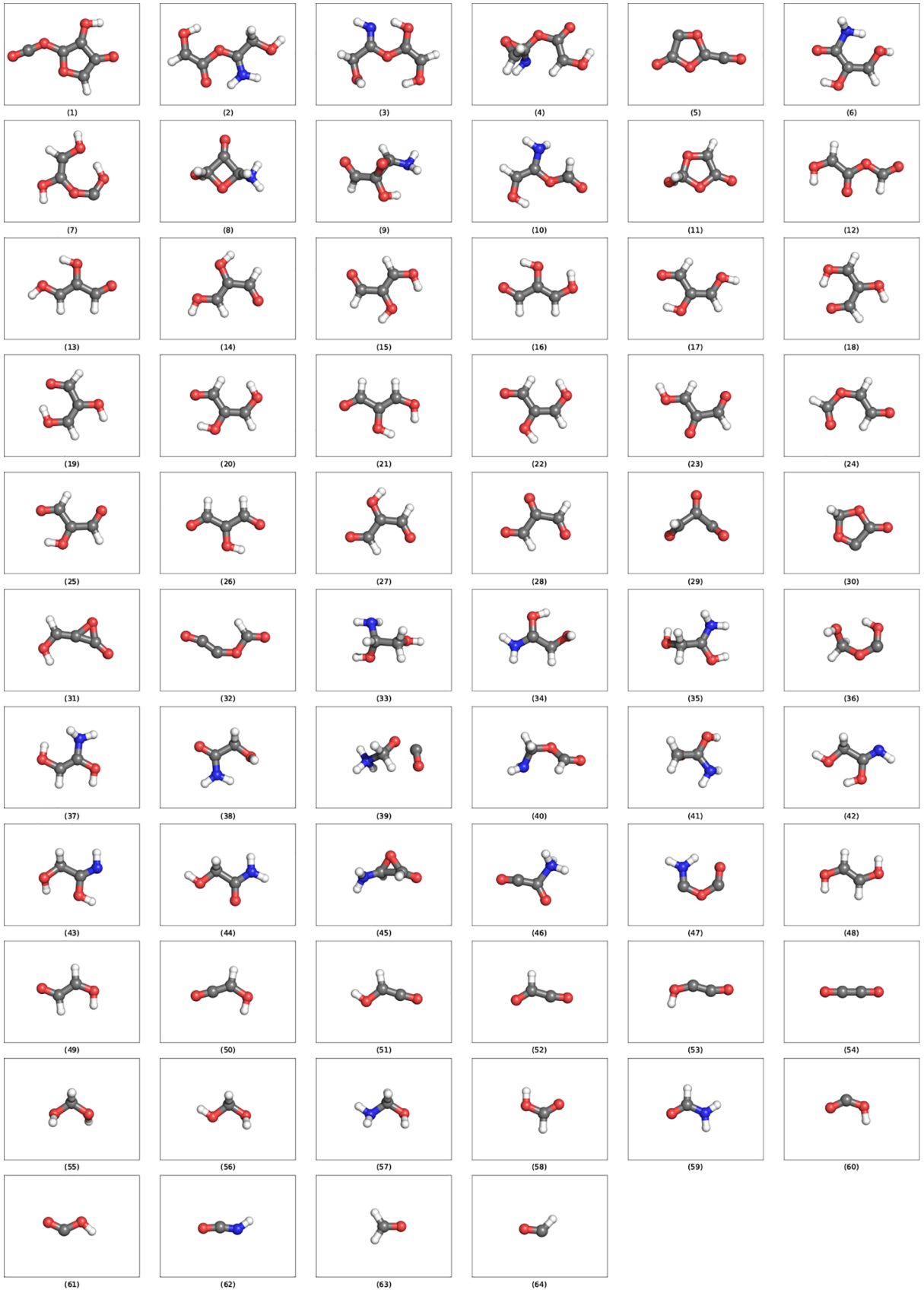


Figure 10. Automatically generated list of the 64 reaction products for the Miller-Urey mixture ordered according to molecular mass.

4.2.4. Miller-Urey System. In ref 2, the authors investigated a mixture of hydrogen, methane, water, ammonia, and carbon monoxide in order to simulate the experiment of Miller and Urey for the formation of biomolecular precursor molecules in

DOI: 10.1021/acs.jctc.9b00143
J. Chem. Theory Comput. 2019, 15, 2847−2862

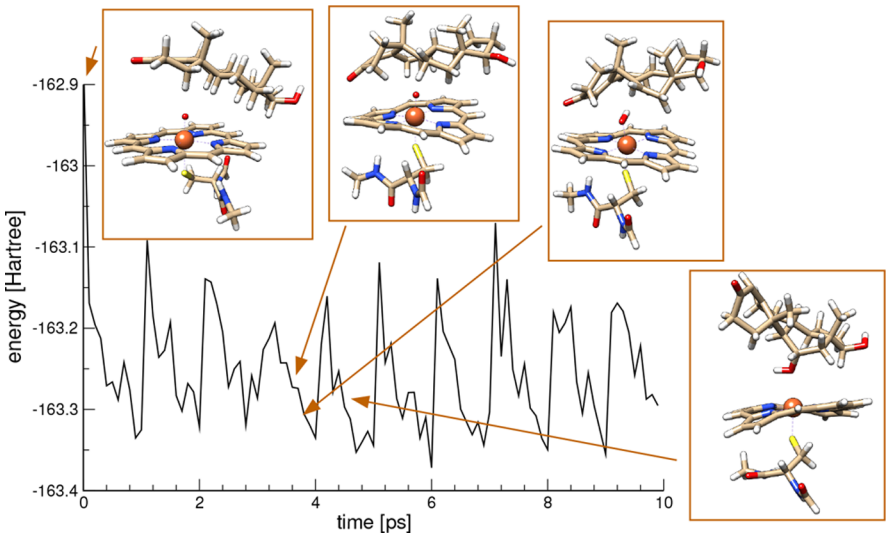


Figure 11. Energy as a function of time in a 10 ps MTD(RMSD) simulation of a P450 model catalyst reacting with testosterone.

the primordial atmosphere.[56] Here, a similar system was simulated consisting of 5CO, $5H_2$, $5H_2O$, $5NH_3$, and $2CH_4$ at a mass density of approximately $\rho = 5$ g/cm^3, $k_i/N = 0.02E_h$, $\alpha = 0.7$ Bohr^{-1}, and an overall run time of 50 ps. The structures of the molecular reaction products (i.e., individually optimized cut-outs from the reaction mixture) are listed in Figure 10.

The average temperature in the simulation was about 1600 K. The reaction started at around 1 ps by dimerization of CO to linear C_2O_2 (54). This molecule can add another CO, and the resulting three-membered ring is subsequently opened by nucleophilic attack of water or ammonia. Direct hydrogenation of CO leads to formaldehyde (63) and, if further water or ammonia is involved, to formic acid (58) or formamide (59). Also the larger molecules formed as the dihydroxyaldehydes (13−17) or well-known ring structures [e.g., (11) or (30)] are chemically meaningful. Under the chosen conditions, methane was practically unreactive, and it is not clear if this is "correct" or an artifact of the chosen QC method. Nevertheless, some products resemble very closely the ones found originally or later in the corresponding experiments.[57] The total computation time including complete optimization of all products was about 6 h on a quad-core laptop (at a time step of 0.2 fs; small MD steps are required due to the many relatively fast moving hydrogen atoms in this system).

4.3. P450 Model Oxidation of Testosterone. The current scientific knowledge of the mechanism and reactivity of the P450 enzyme oxidation system has been reviewed recently.[58] Here, a model complex without the protein is taken from previous studies of Shaik and co-workers.[59] As an example, the reaction with testosterone is considered and some important intermediates for the observed hydroxylation, which are known to occur in this system,[58] are shown in Figure 11.

The average temperature in the simulation was about 700 K. The reaction started around 3.5 ps with the hydrogen abstraction at position 11. Very shortly after this primary event, the OH bound to the iron center is directly transferred back to the carbon yielding the hydroxylation product complexed to P450 as an intermediate (rightmost structure in Figure 11). This so-called "rebound" scenario is in agreement with current knowledge about the mechanism for this and similar substrates.[58] The GFN2-xTB method seems to provide a very good description of the PES for this electronically complicated open-shell transition metal complex. Here, it was essential to restrict the RMSD evaluation (and thus the entire biasing potential) to the substrate, which keeps the catalyst part "cold" (i.e., at the thermostat temperature of about 300 K) and structurally intact over long simulation times. This option for selective heating is a key advantage of the present method. Moreover, except for the standard wall potential, no other bias or constrains have been applied. With the used setup ($\rho = 2$ g/cm^3, $k_i/N = 0.015E_h$, $\alpha = 0.7$ Bohr^{-1}), about 50% of independently started trajectories yield oxidation products within 10 ps. The total computation time for a 10 ps long MTD simulation is about 30 min on a quad-core laptop computer. The described procedure may be helpful for the detection of so-called degradation "soft-spots" (functional groups in drugs preferentially reacting with water or oxygen), which are usually predicted with empirical chemo-informatics methods.[60] Work in this direction is in progress in our lab and will be reported elsewhere.

4.4. Reaction Paths. Before four typical path finding examples will be shown in this section, the purpose of the approach is clarified. Compared to the minimum energy path finding method used here as a reference (GSM[7]), the proposal to optimize from a starting structure to a given product just by applying two pushing/pulling RMSD bias potentials (RMSD-PP) is extremely simple. Actually, it seems intriguing that this works rather generally. The approximate, so-defined RMSD-PP method is normally 1 to 2 orders of magnitude faster than the "exact" GSM path optimizer, which should be kept in mind when the results are compared. Hence, RMSD-PP is suggested as an extremely fast path or transition state estimator that provides useful input for other methods (e.g., direct DFT-based transition optimization). Its main disadvantage is that the optimum biasing potential parameters k_{push}/k_{pull} and α vary with the system by up to 50% such that usually a few different settings have to be tested for each new reaction.

4.4.1. Racemization of Tris(pentafluorophenyl)borane. Delocalized, collective torsion motions are difficult to describe in general as discussed already in section 4.1 for the case of [6]helicene. Here, another practically relevant example for this type of problem is discussed. The propeller-type molecule tris(pentafluoro)phenylborane, which is an important building block in chemically very interesting so-called frustrated-Lewis-

DOI: 10.1021/acs.jctc.9b00143
J. Chem. Theory Comput. 2019, 15, 2847−2862

pairs (FLPs), appears in mirror-image forms which readily interconvert and lead to diastereomers in real complexes.[61] The basic racemization process is investigated, and the comparison of the resulting RMSD-PP path (using k_{push}/N = 0.005, k_{pull}/N = −0.015, α = 1.1) with that obtained from the GSM reference treatment is shown in Figure 12.

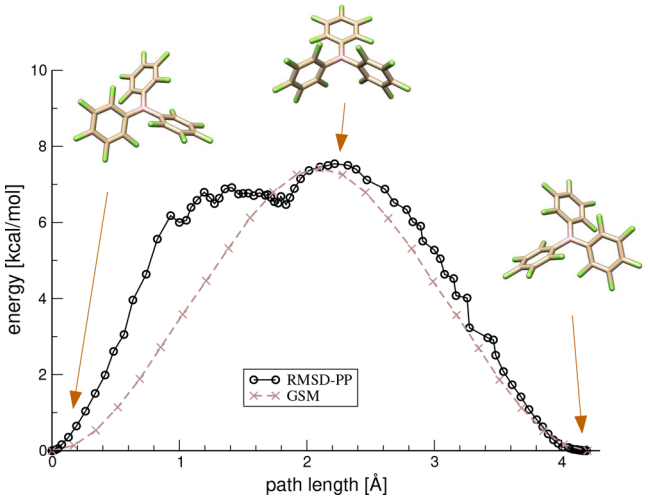


Figure 12. Minimum energy path for the racemization of tris(pentafluorophenyl)borane (GFN2-xTB) from the GSM reference treatment in comparison to the approximate RMSD-PP path.

The general racemization mechanism, the barrier height of 8 kcal/mol, and transition state structure is practically identical with both methods which is very encouraging keeping the complex movement of many atoms in mind. The computed barrier is in reasonable agreement with measurements based on NMR line-shape analysis.[61] Only in the first part of the reaction the approximate RMSD-PP path shows larger deviations from the reference of about 2 kcal/mol and some additional noise. The computation times are 10 s for RMSD-PP and 90 s for GSM on a quad-core laptop computer.

4.4.2. Multistep Cyclization to Lanosterine. A conceptually interesting and important enzymatic transformation in living organisms is the conversion of squalene (an open polyene chain) to polycyclic products, one of which is cholesterol. For a recent review including computational studies, see ref 62. Here, the multistep cyclization to the Lanosterine cation is treated in which sequentially four new C−C bonds are formed leading eventually to the well-known steroid framework. The comparison of the resulting RMSD-PP path (using k_{push}/N = 0.01, k_{pull}/N = −0.0015, α = 0.6) with that obtained from the GSM reference treatment is shown in Figure 13.

The challenge here for start-end-point path finding schemes and in particular for RMSD-PP is that the distance in the multidimensional reaction coordinate is large (RMSD of about 2.5 Å) and that the carbon chain is relatively flexible. Nevertheless, the two approaches yield similar results with the first two steps being practically barrierless and the third cyclization step is identified as rate determining. The computation times are 30 s for RMSD-PP and 74 min for GSM on a quad-core laptop computer.

4.4.3. Hydrogen Transfer in a Cobalt Catalyst. As an example for a typical reaction occurring in a transition metal catalyzed reaction cycle we show the hydrogen transfer in a cobalt-catalyst that was recently investigated theoretically in

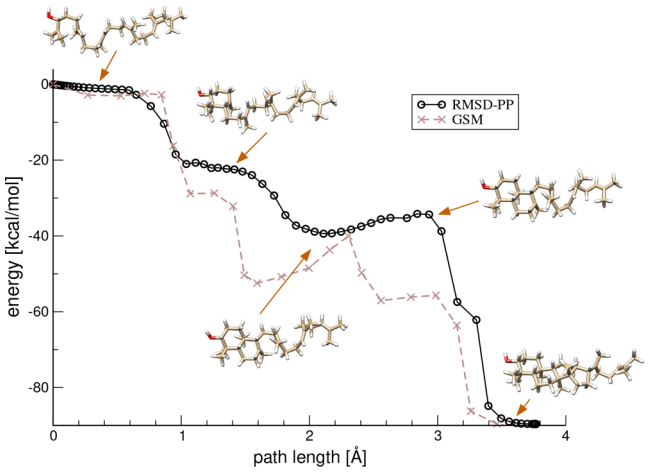


Figure 13. Minimum energy path for the polycyclization to the Lanosterine cation (GFN2-xTB) from the GSM reference treatment in comparison to the approximate RMSD-PP path.

our group.[63] The reaction consists of the complicated movement of a nitrogen atom to the metal, a hydrogen atom transfer from a CH bond to the metal-attached CO_3^-, and the formation of a new carbon−metal bond. A comparison of the two paths using k_{push}/N = 0.02, k_{pull}/N = −0.04, α = 0.6 for RMSD-PP is shown in Figure 14.

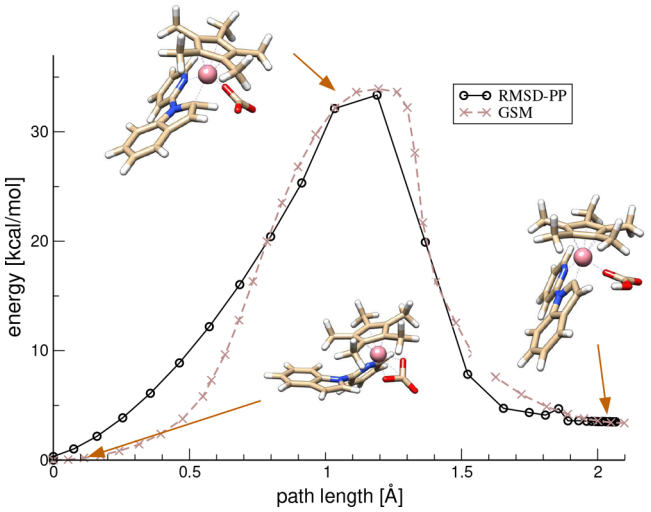


Figure 14. Minimum energy path for the hydrogen transfer in a cobalt catalyst (GFN2-xTB) from the GSM reference treatment in comparison to the approximate RMSD-PP path.

Here, the agreement between the results from the two methods is striking. The computed GFN2-xTB forward barrier of about 24 kcal/mol is in reasonable agreement with that from a single-point treatment at the B97-3c level[64] giving 32.2 kcal/mol. The computation times are 11 s for RMSD-PP and 7 min for GSM on a quad-core laptop computer.

4.4.4. Dimerization of Anthracene. Thermally "forbidden" chemical reactions often showing biradical character can be investigated by GFN2-xTB at high electronic temperatures which account implicitly for multiconfigurational effects.[27,65] In the following example of the [4 + 4]π thermal dissociation of the anthracene dimer, this property is exploited using an electronic temperature of 6000 K. The comparison of the two

DOI: 10.1021/acs.jctc.9b00143
J. Chem. Theory Comput. 2019, 15, 2847−2862

paths using $k_{push}/N = 0.01$, $k_{pull}/N = -0.02$, $\alpha = 0.6$ for RMSD-PP is shown in Figure 15.

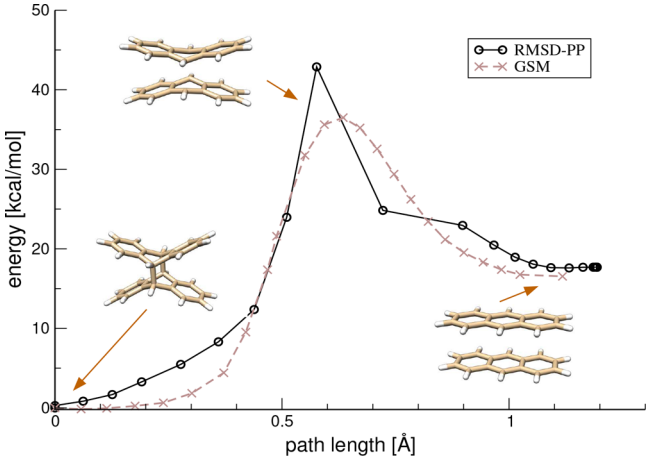


Figure 15. Minimum energy path for the dimerization of anthracene (GFN2-xTB) from the GSM reference treatment in comparison to the approximate RMSD-PP path.

Again, the agreement between the two paths is rather good although not as perfect as in the previous example. The computed forward (dissociation) barrier of 37 kcal/mol (43 with RMSD-PP) is in qualitative agreement with the observed thermal stability of dianthrancene[66] and previous theoretical results (45 kcal/mol[67]). The computation times are 11 s for RMSD-PP and 78 min for GSM on a quad-core laptop computer.

5. CONCLUSIONS

The GFN2-xTB tight-binding method is ideally suited to investigate chemical reaction space automatically using metadynamics (MTD) simulations. This approximate quantum chemical method is robust also for complicated electronic structures or at high temperatures and applicable for almost all elements of the periodic table. It describes in combination with the Fermi-smearing technique fundamental processes like covalent dissociation or conformational changes reasonably accurate. It runs at a very high computational speed such that even for large systems with a few hundred atoms, thousands to millions of energy and gradient evaluations are feasible on laptop or common workstation computers. The resulting potential energy surfaces are mostly qualitatively or even semiquantitatively correct but normally require further refinement by more sophisticated DFT or WFT methods. Hence, what is proposed here is of exploratory character and suggested as a starting point for typical chemical projects.

The generality of the underlying electronic structure method and the consistency of the PES is ideally combined with a rather general and robust biasing potential which is newly introduced to low-cost quantum chemistry here. It is based on a Gaussian function of the root-mean-square-deviation (RMSD) in Cartesian space for the collective variables. In a MTD simulation, it allows atom-selective heating and thereby efficient crossing of small and large chemical barriers depending on the strength of the bias (i.e., the problem under consideration). In a worst case model system (alkyl-substituted helicene), it could be shown that the simultaneous presence of very soft and stiff vibrational motions does not deteriorate the performance of the algorithm. The versatility of the MTD(RMSD)/GFN2-xTB approach is documented exemplarily for three important chemical problems: conformer search, chemical compound space exploration in a nanoreactor, and estimating reaction paths.

The progress made for the conformational problem is probably most striking. Compared to the previously developed, already rather good procedure, the new MTD(RMSD)-GC algorithm yields more complete conformer ensembles at significantly reduced computational effort. In some cases, lower-energy conformers could be found and even more important, large (e.g., Vancomycin) or otherwise challenging cases as the alkyl-substituted helicene can be treated, for which the old algorithm completely fails. Also the nanoreactor simulations were successful. The calculations for reasonably sized systems of hundreds of atoms complete robustly in little computation time (hours) and provide chemically sensible chemical reaction mechanisms and products. The results for the P450 model oxidation reaction of testosterone are impressive. Without any system-specific modification, the MTD(RMSD) simulations provide a qualitatively correct reaction mechanism for such an electronically complicated open-shell system. The key advantage here is the selectivity of the RMSD based heating in which the P450 catalyst part is excluded (i.e., kept cold).

Some disadvantage of the RMSD biasing potential is that weakly bound complexes are preferentially dissociated before other internal degrees of freedom are excited. Thus, if the PES of noncovalently bound systems are to be investigated, additional contrains like the reactor wall or fragment-center-of-mass distance-based restraining potentials should be applied. In the reactions of benzene and ferrocene, this has been exploited in order to favor isomerization over fragmentation reactions. In principle, the proposed MTD-(RMSD) tool could be coupled with any atomistic structure method. However, in the authors opinion, current force-fields are either too inaccurate and not sufficiently general (e.g., lacking the ability to generally break and form new bonds) or miss global consistency for the PES and hence would waste the strengths of the concept. Application of more sophisticated DFT treatments would of course increase the accuracy of the entire scheme, however, at the expense of impractically long computation times even with "low-cost" variants. The nanoreactor and conformational search procedures require typically about 10^4-10^6 energy and force evaluations per simulation for 30−100 atoms, which would severely limit the size of the routinely investigated molecules with DFT methods.

The proposed MTD(RMSD)-GC/GFN2-xTB conformer search algorithm is set to our new default procedure and more results will be presented in the near future for other conformational benchmarks, protomers, tautomers, and transition states as well as for the first step in the automatic computation of high-resolution NMR spectra. In its design, more emphasis was put to the completeness of the generated ensemble and in particular for finding the correct global minimum, than to extreme speed. In actual applications, one is usually willing to invest some more computation time to get with higher certainty the right result. The extremely efficient RMSD-PP guess for reaction paths and barriers could be used to semiautomatically assess whether the found conformers can really be interconverted and, hence, are relevant under the experimental conditions. It should be stressed at this point that the ultimate test for the quality of the produced GFN2-xTB

DOI: 10.1021/acs.jctc.9b00143
J. Chem. Theory Comput. 2019, 15, 2847−2862

conformer ensembles is, beside direct comparison to the experiment, to provide many energetically low (including the lowest) structures in subsequent high-level DFT or WFT treatments including solvation effects. For the real-life systems considered here this represents another challenging project carried out in our group currently. Eventually this may pave the way to a sparsely empirical theory of the three-dimensional structures of large molecules.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.9b00143.

Cartesian coordinates of the starting structures and best structures found for the conformer benchmark (TXT)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: grimme@thch.uni-bonn.de.

ORCID

Stefan Grimme: 0000-0002-5844-4371

Funding

This work was supported by the DFG in the framework of the "Gottfried Wilhelm Leibniz-Preis".

Notes

The author declares no competing financial interest.

■ ACKNOWLEDGMENTS

The author further thanks Dr. A. Hansen, Dr. C. Mück-Lichtenfeld, F. Bohle, M. Bursch, S. Dohm, S. Ehlert, J. Seibert, and P. Pracht for fruitful discussions, suggestions, implementations, and technical support.

■ REFERENCES

(1) Behler, J. Representing potential energy surfaces by high-dimensional neural network potentials. *J. Phys.: Condens. Matter* **2014**, *26*, 183001.

(2) Wang, W.; Titov, A.; McGibbon, R.; Liu, F.; Pande, V. S.; Martinez, T. J. Discovering chemistry with an ab initio nanoreactor. *Nat. Chem.* **2014**, *6*, 1044−1048.

(3) Dewyer, A. L.; Arguelles, A. J.; Zimmerman, P. M. Methods for exploring reaction space in molecular systems. *WIREs Comput. Mol. Sci.* **2018**, *8*, e1354.

(4) Simm, G. N.; Vaucher, A. C.; Reiher, M. Exploration of Reaction Pathways and Chemical Transformation Networks. *J. Phys. Chem. A* **2019**, *123*, 385−399.

(5) Jochum, C.; Gasteiger, J.; Ugi, I. The Principle of Minimum Chemical Distance (PMCD). *Angew. Chem., Int. Ed. Engl.* **1980**, *19*, 495−505.

(6) Ugi, I.; Bauer, J.; Bley, K.; Dengler, A.; Dietz, A.; Fontain, E.; Gruber, B.; Herges, R.; Knauer, M.; Reitsam, K.; Stein, N. Computer-Assisted Solution of Chemical Problems-The Historical Development and the Present State of the Art of a New Discipline of Chemistry. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 201−227.

(7) Zimmerman, P. M. Growing string method with interpolation and optimization in internal coordinates: Method and examples. *J. Chem. Phys.* **2013**, *138*, 184102.

(8) Plessow, P. Reaction Path Optimization without NEB Springs or Interpolation Algorithms. *J. Chem. Theory Comput.* **2013**, *9*, 1305−1310.

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

(10) Pracht, P., Grimme, S., to be submitted. Please contact xtb@thch.uni-bonn.de for the CREST program.

(11) van Gunsteren, W. F.; Bürgi, R.; Peter, C.; Daura, X. The Key to Solving the Protein-Folding Problem Lies in an Accurate Description of the Denatured State. *Angew. Chem., Int. Ed.* **2001**, *40*, 351−355.

(12) Sindhikara, D.; Spronk, S. A.; Day, T.; Borrelli, K.; Cheney, D. L.; Posy, S. L. Improving Accuracy, Diversity, and Speed with Prime Macrocycle Conformational Sampling. *J. Chem. Inf. Model.* **2017**, *57*, 1881−1894.

(13) Cavasin, A. T.; Hillisch, A.; Uellendahl, F.; Schneckener, S.; Göller, A. H. Reliable and Performant Identification of Low-Energy Conformers in the Gas Phase and Water. *J. Chem. Inf. Model.* **2018**, *58*, 1005−1020.

(14) van Duin, A. C. T.; Dasgupta, S.; Lorant, F.; Goddard, W. A. ReaxFF: A Reactive Force Field for Hydrocarbons. *J. Phys. Chem. A* **2001**, *105*, 9396−9409.

(15) Grimme, S. A General Quantum Mechanically Derived Force Field (QMDFF) for Molecules and Condensed Phase Simulations. *J. Chem. Theory Comput.* **2014**, *10*, 4497−4514.

(16) Yang, Y.; Yu, H.; York, D.; Cui, Q.; Elstner, M. Extension of the self-consistent-charge density-functional tight-binding method: third-order expansion of the density functional theory total energy and introduction of a modified effective coulomb interaction. *J. Phys. Chem. A* **2007**, *111*, 10861−73.

(17) Christensen, A. S.; Kubař, T.; Cui, Q.; Elstner, M. Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications. *Chem. Rev.* **2016**, *116*, 5301−5337.

(18) Grimme, S.; Bannwarth, C.; Shushkov, P. 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.* **2017**, *13*, 1989−2009.

(19) Bursch, M.; Hansen, A.; Grimme, S. Fast and Reasonable Geometry Optimization of Lanthanoid Complexes with an Extended Tight Binding Quantum Chemical Method. *Inorg. Chem.* **2017**, *56*, 12485−12491.

(20) Struch, N.; Bannwarth, C.; Ronson, T. K.; Lorenz, Y.; Mienert, B.; Wagner, N.; Engeser, M.; Bill, E.; Puttreddy, R.; Rissanen, K.; Beck, J.; Grimme, S.; Nitschke, J. R.; Lützen, A. An Octanuclear Metallosupramolecular Cage Designed To Exhibit Spin-Crossover Behavior. *Angew. Chem., Int. Ed.* **2017**, *56*, 4930−4935.

(21) Seibert, J.; Bannwarth, C.; Grimme, S. Biomolecular Structure Information from High-Speed Quantum Mechanical Electronic Spectra Calculation. *J. Am. Chem. Soc.* **2017**, *139*, 11682−11685.

(22) Pracht, P.; Bauer, C. A.; Grimme, S. Automated and efficient quantum chemical determination and energetic ranking of molecular protonation sites. *J. Comput. Chem.* **2017**, *38*, 2618−2631.

(23) Pracht, P.; Wilcken, R.; Udvarhelyi, A.; Rodde, S.; Grimme, S. High accuracy quantum-chemistry-based calculation and blind prediction of macroscopic pKa values in the context of the SAMPL6 challenge. *J. Comput.-Aided Mol. Des.* **2018**, *32*, 1139−1149.

(24) Asgeirsson, V.; Bauer, C. A.; Grimme, S. Quantum chemical calculation of electron ionization mass spectra for general organic and inorganic molecules. *Chem. Sci.* **2017**, *8*, 4879−4895.

(25) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB - an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652−1671.

(26) Mermin, N. D. Thermal Properties of the Inhomogeneous Electron Gas. *Phys. Rev.* **1965**, *137*, A1441−A1443.

(27) Grimme, S.; Hansen, A. A practicable real-space measure and visualization of static electron-correlation effects. *Angew. Chem., Int. Ed.* **2015**, *54*, 12308−12313.

(28) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **2014**, *185*, 604−613.
(29) Levine, B. G.; Coe, J. D.; Martinez, T. J. Optimizing Conical Intersections without Derivative Coupling Vectors: Application to Multistate Multireference Second-Order Perturbation Theory (MS-CASPT2). *J. Phys. Chem. B* **2008**, *112*, 405−413.
(30) Hovan, L.; Comitani, F.; Gervasio, F. L. Defining an Optimal Metric for the Path Collective Variables. *J. Chem. Theory Comput.* **2019**, *15*, 25−32.
(31) Preto, J.; Clementi, C. Fast recovery of free energy landscapes via diffusion-map-directed molecular dynamics. *Phys. Chem. Chem. Phys.* **2014**, *16*, 19181.
(32) Zheng, W.; Rohrdanz, M. A.; Clementi, C. Rapid Exploration of Configuration Space with Diffusion-Map-Directed Molecular Dynamics. *J. Phys. Chem. B* **2013**, *117*, 12769−12776.
(33) Schlitter, J.; Engels, M.; Krüger, P. Targeted molecular dynamics: A new approach for searching pathways of conformational transitions. *J. Mol. Graphics* **1994**, *12*, 84−89.
(34) Fiorin, G.; Klein, M. L.; Henin, J. Using collective variables to drive molecular dynamics simulations. *Mol. Phys.* **2013**, *111*, 3345−3362.
(35) *xtb*, version 5.9; University Bonn, 2019. Please contact xtb@thch.uni-bonn.de for access to the program.
(36) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562−12566.
(37) Stepanenko, S.; Engels, B. Gradient tabu search. *J. Comput. Chem.* **2007**, *28*, 601−611.
(38) Müller, E. M.; de Meijere, A.; Grubmüller, H. Predicting unimolecular chemical reactions: Chemical flooding. *J. Chem. Phys.* **2002**, *116*, 897−905.
(39) Chen, M.; Cuendet, M. A.; Tuckerman, M. E. Heating and flooding: A unified approach for rapid generation of free energy surfaces. *J. Chem. Phys.* **2012**, *137*, 024102.
(40) Barducci, A.; Bonomi, M.; Parrinello, M. Metadynamics. *WIREs Comput. Mol. Sci.* **2011**, *1*, 826−843.
(41) Coutsias, E. A.; Seok, C.; Dill, K. A. Using Quaternions to Calculate RMSD. *J. Comput. Chem.* **2004**, *25*, 1849−1857.
(42) Berendsen, H. C. J. *Simulating the Physical World: Hierachical Modeling from Quantum Mechanics to fluid Dynamics*; Cambridge University Press: Cambridge, 2007.
(43) GFN0-xTB is a first-order-only variant in the GFN family of TB methods using classical electrostatics which is right now developed in our group. It is about 3−10 times fast than GFN2-xTB and well suited for biochemical systems. It will be described in detail soon in a separate publication.
(44) Shiga, M.; Masia, M. Boundary based on exchange symmetry theory for multilevel simulations. I. Basic theory. *J. Chem. Phys.* **2013**, *139*, 044120.
(45) *TURBOMOLE*, version 7.2; Universitä Karlsruhe and Forschungszentrum Karlsruhe GmbH: Karlsruhe, 2017.
(46) Furche, F.; Ahlrichs, R.; Hättig, C.; Klopper, W.; Sierka, M.; Weigend, F. Turbomole. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2014**, *4*, 91−100.
(47) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327−341.
(48) Kim, S.; Thiessen, P.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound databases. *Nucleic Acids Res.* **2016**, *44*, D1202−D1213.
(49) PubChem. https://www.ncbi.nlm.nih.gov/ (accessed Mar 19,2019).
(50) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.
(51) Grimme, S.; Peyerimhoff, S. D. Theoretical study of the structures and racemization barriers of [n]helicenes (n = 3−6,8). *Chem. Phys.* **1996**, *204*, 411−417.
(52) Eliel, E. L.; Wilen, S. H. *Stereochemistry of Organic Compounds*; Wiley: New York, 1994.
(53) Qu, Z.-w.; Hansen, A.; Grimme, S. Co-C Bond Dissociation Energies in Cobalamin Derivatives and Dispersion Effects: Anomaly or Just Challenging? *J. Chem. Theory Comput.* **2015**, *11*, 1037−1045.
(54) Dinadayalane, T. C.; Priyakumar, U. D.; Sastry, G. N. Exploration of C6H6 Potential Energy Surface: A Computational Effort to Unravel the Relative Stabilities and Synthetic Feasibility of New Benzene Isomers. *J. Phys. Chem. A* **2004**, *108*, 11433−11448.
(55) Chenoweth, K.; van Duin, A. C. T.; Goddard, W. A. ReaxFF Reactive Force Field for Molecular Dynamics Simulations of Hydrocarbon Oxidation. *J. Phys. Chem. A* **2008**, *112*, 1040−1053.
(56) Miller, S. L.; Urey, H. C. Organic Compound Synthes on the Primitive Eart. *Science* **1959**, *130*, 245−251.
(57) https://en.wikipedia.org/wiki/Miller (accessed Mar 19, 2019).
(58) Dubey, K. D.; Shaik, S. Cytochrome P450-The Wonderful Nanomachine Revealed through Dynamic Simulations of the Catalytic Cycle. *Acc. Chem. Res.* **2019**, *52*, 389.
(59) Ogliaro, F.; Harris, N.; Cohen, S.; Filatov, M.; de Visser, S. P.; Shaik, S. A Model "Rebound" Mechanism of Hydroxylation by Cytochrome P450: Stepwise and Effectively Concerted Pathways, and Their Reactivity Patterns. *J. Am. Chem. Soc.* **2000**, *122*, 8977−8989.
(60) Gloriam, D. E.; Olsen, L.; Rydberg, P. The SMARTCyp cytochrome P450 metabolism prediction server. *Bioinformatics* **2010**, *26*, 2988−2989.
(61) Stephan, D. W.; Erker, G. Frustrated Lewis pairs: metal-free hydrogen activation and more. *Angew. Chem., Int. Ed.* **2010**, *49*, 46−76.
(62) Hess, B. A. Computational studies on the cyclization of squalene to the steroids and hopenes. *Org. Biomol. Chem.* **2017**, *15*, 2133−2145.
(63) Zell, D.; Müller, V.; Dhawa, U.; Bursch, M.; Presa, R. R.; Grimme, S.; Ackermann, L. Mild Cobalt(III)-Catalyzed Allylative C-F/C-H Functionalizations at Room Temperature. *Chem. - Eur. J.* **2017**, *23*, 12145−12148.
(64) Brandenburg, J. G.; Bannwarth, C.; Hansen, A.; Grimme, S. B97−3c: A revised low-cost variant of the B97-D density functional method. *J. Chem. Phys.* **2018**, *148*, 064104.
(65) Bauer, C. A.; Hansen, A.; Grimme, S. The Fractional Occupation Number Weighted Density as a Versatile Analysis Tool for Molecules with a Complicated Electronic Structure. *Chem. - Eur. J.* **2017**, *23*, 6150−6164.
(66) Grimme, S.; Peyerimhoff, S. D.; Bouas-Laurent, H.; Desvergne, J.-P.; Becker, H.-D.; Sarge, S. M.; Dreeskamp, H. Calorimetric and quantum chemical studies of some photodimers of anthracenes. *Phys. Chem. Chem. Phys.* **1999**, *1*, 2457−2462.
(67) Tapilin, V. M.; Bulgakov, N. N.; Chupakhin, A. P.; Politov, A. A. On the mechanism of mechanochemical dimerization of anthracene. Quantum-chemical calculation of the electronic structure of anthracene and its dimer. *J. Struct. Chem.* **2008**, *49*, 581−586.

DOI: 10.1021/acs.jctc.9b00143
J. Chem. Theory Comput. 2019, 15, 2847−2862