Tensor Train Optimization for Conformational Sampling of Organic Molecules

Chemistry

By Terra Quantum AG
terraquantum.swiss

2024

Tensor Train Optimization for Conformational Sampling of Organic Molecules

Christopher Zurek,[†] Ruslan A. Mallaev,[‡] Alexander Paul,[‡] Nils van Staalduinen,[†] Philipp Pracht,[¶] Roman Ellerbrock,[*,‡] and Christoph Bannwarth[*,†]

[†]*Institute of Physical Chemistry, RWTH Aachen University, Aachen, Germany*

[‡]*Terra Quantum AG, St. Gallen, Switzerland*

[¶]*Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA*

E-mail: romanellerbrock@gmail.com; bannwarth@pc.rwth-aachen.de

Abstract

Exploring the conformational space of molecules remains a challenge of fundamental importance to quantum chemistry: identification of relevant conformers at ambient conditions enables predictive simulations of almost arbitrary properties. Here, we propose a novel approach to enable conformational sampling of large organic molecules where the combinatorial explosion of possible conformers prevents the use of a brute-force systematic conformer search. We employ tensor trains as a highly efficient dimensionality reduction algorithm, effectively reducing the scaling from exponential to polynomial. In our approach, the conformational search is expressed as global energy minimization task in a high-dimensional grid of dihedral angles. Dimensionality reduction is achieved through a tensor train representation of the high-dimensional torsion space. The performance of the approach is assessed on a variety of drug-like molecules in

direct comparison to the state-of-the-art metadynamics based conformer rotamer ensemble sampling tool (CREST). The comparison shows significant acceleration of up to an order of magnitude, while maintaining comparable accuracy. More importantly, the presented approach allows treatment of larger molecules than typically accessible with metadynamics.

1 Introduction

While the structure activity relationship (SAR) is a well known concept to every chemist, drawing a causal link between a molecule's structure and its properties remains a major challenge of computational chemistry. In addition to the chemical composition and connectivity, the 3-dimensional arrangement of a molecule's atoms largely impacts its physical and chemical properties. Therefore, the search for all thermally accessible conformations is crucial for chemical simulations.[1] For this purpose, numerous approaches have been developed over the past decades.[2] Each of these approaches requires a structure generation procedure and subsequent energetic ranking of the obtained geometries.

While gradient optimization algorithms are suitable to identify local minima on the potential energy surface (PES), they provide no information on whether the overall energetically lowest structure is obtained.[3] To determine the global minimum on the PES, so-called global optimization algorithms are required.[4] Finding relevant minima in conformational space traditionally relies on initial guesses constructed from intuition, which makes it difficult to scale to systems with many degrees of freedom. To identify the global minimum automatically, multiple structures across the conformational space need to be examined. However, with increasing size and flexibility of a molecule, the number of possible conformations drastically increases necessitating efficient procedures for this global optimization problem. An early approach by Saunders is built upon Metropolis Monte Carlo (MC) simulations utilizing random displacements of atoms in Cartesian space. The acceptance rate of each individual step is based on molecular mechanics (MM) energies after subsequent gradient optimization.[5] Even for rather small molecules, the $3N$-6 internal degrees of freedom lead to a huge search space. A related global optimization approach is the basin hopping approach by Wales and Doye,[6] which was later adapted by Goedecker.[7] For conformational problems, neglecting rigid bond distances and valence angles allows for a significant simplification of the coordinate space as demonstrated by Chang *et al.*[8] As the individual modification of dihedral angles is not possible without damaging cyclic structures, the approach is primarily suited

for acyclic molecules. Lipton and Still extended the method to cyclic systems by applying ring closure constraints during the gradient optimization.[9]

In contrast to relying on entirely random structure modifications, molecular dynamics (MD) simulations provide a more chemically motivated strategy to conformer sampling. During an MD simulation, the internal motion of a molecule is simulated for a certain time at a fixed temperature. A conformer ensemble is generated from an MD trajectory by extracting snapshots at fixed time intervals and optimizing the structures using standard gradient methods. The application of MD to smaller oligopeptides was shown by Brooks *et al.* and even to an entire protein by Elber and Karplus using myoglobin as example.[10,11] While MD simulations, in principle, guarantee to converge to the global minimum, the required simulation times and the associated number of energy evaluations is usually unfeasible, thus limiting MD to a rather local analysis of the conformational space.[12] To accelerate structural changes upon simulation, two general approaches can be pursued: either increasing the internal energy of the system by raising the temperature or effectively flattening the PES to facilitate passing interconformational energy barriers. The former approach is applied in simulated annealing, which is applicable to MD as well as to Metropolis MC simulations.[13] Simulated annealing mimics controlled crystal growth by starting the simulation at high temperatures and then gradually lowering it, which was demonstrated as a viable method for conformer sampling by Wilson *et al.* in the context of MC and by Lelj *et al.* for MD simulations.[14,15] Alternatively, the PES can be flattened by either a constant boost potential as introduced by Hamelberg *et al.* or a dynamically updated bias potential as described by Laio and Parrinello significantly lowering the required time demand.[16,17] The latter is referred to as metadynamics (MTD) simulation, which was much simplified by Grimme using the Cartesian root mean square deviation (RMSD) of the atomic positions as collective variable for constructing the bias potential.[18] Based on this strategy, Pracht *et al.* developed CREST making primarily use of MTD but also MD simulations as well as genetic crossing (GC) to generate conformer and rotamer ensembles of almost any type of molecules.[19,20] Here, rotamers refer to degenerate

conformers that only vary in permutations of chemically equivalent atoms.

While the aforementioned approaches can eventually lead to the global minimum, they lack a directional force guiding the algorithm towards the global minimum. Different from that, buildup approaches are guided by knowledge of favorable conformations of fragments, which are used to construct the molecule's conformers. This fragmentation procedure was applied to the analysis of a pentapeptide constructed from amino acid conformers by Vásquez and Scheraga.[21] However, the definition of fragments is, to some extent, arbitrary and the quality of the final ensemble heavily relies on the availability of fragments in the respective data base. Therefore, a more general approach treating every possible bond rotation explicitly would be desirable. So far, this has been impossible due to the unfavorable exponential scaling of systematic conformer search approaches.

Tensor networks have been a powerful tool used in condensed matter physics and chemical physics to mitigate exponential scaling in multi-dimensional quantum systems.[22] A commonly known approach is the density matrix renormalization group (DMRG) originally introduced by White *et al.*, where a tensor train (TT) is used to describe 1D spin lattice systems.[23] Another well-known approach is the multi-configurational time-dependent Hartree (MCTDH) method developed by Meyer and coworkers to describe quantum dynamics of molecules.[24] Tensor networks factorize high-dimensional tensors into products of lower dimensional tensors where the nodes in the network represent the tensors and edges represent their indices. While tensor networks efficiently deal with high-dimensional systems, they are limited to applications in linear algebra. Analogously, high-dimensional functions can be separated into lower dimensional ones using the tensor network formalism by using tensor network grids. Consequently, a high-dimensional grid is decomposed into multiple lower-dimensional grids. The first tensor network grid was proposed by Manthe, called the correlation discrete variable representation (CDVR).[25] This idea was later extended from simple network types to general trees.[26] Recently, the scaling of the approach was drasti-

cally reduced, and schemes for systematic convergence were developed.[27,28] Independently, Oseledets and coworkers developed the cross approximation scheme for TTs which follows similar ideas but was first developed for the compression of functions.[29] The main difference compared to the CDVR approach is the way in which the different grids are combined and that lower dimensional grids are subsets of the high-dimensional grid in the cross approximation. Tensor network grids have a variety of uses including integration, compression, and optimization.[30] In the present work, tensor network grids provide the foundation for our approach of systematically finding conformers.

Each of the aforementioned approaches inevitably requires numerous energy evaluations, necessitating either MM or low-cost quantum mechanics (QM) methods to maintain a manageable computational demand. Most prominent examples in the field of organic chemistry are the General AMBER Force Field (GAFF) originally introduced by Wang *et al.*, particularly, since the implementation of a new charge model by He *et al.*. Other prominent force fields are the most recent versions of OPLS and OpenFF developed by Lu *et al.* and Boothroyd *et al.*, respectively.[31–35] A comprehensive benchmark by Lim *et al.* indicates the superior performance of the latter two methods.[36] Regarding large biomolecular systems predominantly Amber and CHARMM force fields are used, typically designed for the application to either proteins, nucleic acids, carbohydrates or lipids.[37–44]

While MM methods are usually preferred due to their low computational cost, their parametrization usually restricts them to a narrow range of molecules. Hence, they are not suitable for a universal conformer search approach.[45] Most promising for energy evaluation is the GFN family of methods, in particular the force field GFN-FF and the semiempirical quantum mechanics (SQM) method GFN2-xTB.[46,47] Since GFN2-xTB is efficient, reproduces molecular geometries particularly well, and is parametrized for the majority of elements ($Z \leq 86$), it is often the method of choice in quantum chemical investigations of medium-sized organic and inorganic molecules and is well suited for large conformer screenings.

2 Theory

2.1 Conformer Search

We formalize the conformer search as the optimization problem $\underset{\mathbf{X}}{\operatorname{argmin}} E(\mathbf{X})$ where E is the (electronic ground state) energy of an N-atom molecule and $\mathbf{X}$ represents the $3N$ Cartesian coordinates. Since the energy is usually invariant to translation and rotation, we express the problem more naturally in terms of the $3N-6$ internal coordinates (denoted by $\mathbf{Q}$) and instead solve $\underset{\mathbf{Q}}{\operatorname{argmin}} E(\mathbf{Q})$. Suitable choice of coordinate representation drastically reduces correlation between the coordinates, which facilitates finding minima. We use bond, angle, and torsion (BAT)-coordinates, $\mathbf{Q}$ (described in detail later), and select only the relevant dihedral angles. We represent them on a grid which results in a tensor $E_{i_1\cdots i_d} = E(Q_{i_1\cdots i_d})$ storing the molecule's potential energy as a function of the internal coordinates Q. Here, $i_1 \cdots i_d$ denote the individual indices of the d-dimensional tensor. Searching all geometries would scale exponentially in the number of rotatable bonds. We avoid this by using TTs to quickly search for the optimal coordinates corresponding to the energetically lowest conformer with only polynomial effort. In our approach, we combine the search with a local gradient optimizer, so that the tensor network mainly serves as a tool to generate points in the vicinity of multiple different suitable local minima and to avoid clustering of points in one or a few minima.

2.2 Tensor Train Optimization

The TT representation of an order d tensor $\mathbf{A}$ with elements $A_{i_1\ldots i_d}$ reads

$$A_{i_1\ldots i_d} = \sum_{l_1\ldots l_d}^{r_1\ldots r_d} A^1_{i_1 l_1} \cdot A^2_{l_1 i_1 l_2} \cdot \ldots \cdot A^d_{l_d i_d}. \tag{1}$$

where the original tensor is factorized into a sequence of 2nd and 3rd order tensor cores $\mathbf{A^k}$ with elements $A^k_{l_{k-1} i_k l_k}$. These tensor cores carry new virtual indices l_{k-1} and l_k that connect

neighboring tensor cores, except for the first and last tensor core which only have a single virtual index l_1 and l_d, respectively. The upper bounds of the virtual indices, $r_1, ..., r_d$, are the called ranks or *bond dimensions*. In general, the ranks required to accurately represent tensors $\mathbf{A}$ can grow exponentially order d. In many practical applications, however, small r_k are sufficient to represent the tensor and we say that it has a low-rank TT representation.[48] In principle, the TT representation can be obtained by performing a sequence of singular value decompositions on the core tensor. This, however, would require *a priori* knowledge of the large core tensor.[48] If the full tensor itself is unknown, the TT representation of $\mathbf{A}$ can be generated by evaluating elements *on demand* from a function $A_{i_1 \ldots i_d} = A(x_{i_1}, ..., x_{i_d})$. Furthermore, it is desirable to evaluate this function very rarely, since evaluating the function itself is costly. The TT cross approximation[29] enables us to compress a tensor into TT format from accessing just a few elements. Additionally, we obtain an element close to the modulus largest value as an entry in one of the core tensors. Due to these properties, it is a highly efficient algorithm for finding function minima (or maxima).[49]

We first introduce the matrix cross approximation and extend the discussion to the more general TT cross approximation. Matrix cross approximation[50] aims at finding a rank-r representation of a matrix with elements A_{ij} by accessing only a few rows and columns. The target representation consists of a set of skeleton rows with indices I_r and columns I_c approximating the matrix as

$$A_{ij} \approx \sum_{I_c I_r} A_{i I_c} A^{-1}_{I_c I_r} A_{j I_r}, \tag{2}$$

where $A^{-1}_{I_c I_r}$ denotes elements of the inverse of the so-called cross matrix $A_{I_c I_r}$. If A is of rank r and described by *any* r linear independent rows and columns, the representation is accurate. If not, the choice of rows and columns matters. A convenient choice is the set of rows and columns that maximizes the volume of the cross matrix $|\det(A)| = \prod_k \sigma_k$ where σ_k are the singular values of the cross matrix.[51] The volume of a matrix is closely related to the Frobenius norm $||A||_F = \sum_k \sigma_k^2$ of a matrix, but there exist algorithms that allow performing updates on the volume with preferable scaling. The matrix cross approximation

proceeds as follows:

1. Choose a random set of linear independent columns I_c.
2. Find the $r \times r$ submatrix with maximal volume in the rectangular column submatrix matrix $A(:, I_c)$ using the `maxvol` algorithm[51] (here we use matlab notation).
3. Set I_r to the rows obtained in the `maxvol` algorithm.
4. Now search for the $r \times r$ submatrix in the row submatrix $A(I_r, :)$ using `maxvol` and set I_c to the obtained columns.
5. Continue these iterations, optimizing row and column indices, until convergence is reached.

The TT cross approximation performs a matrix cross approximation, but on the flattened tensor cores at every node. At a site k, the $r \times N \times r$ tensor is reshaped to a $(r \cdot N) \times r$ matrix and the maximal volume submatrix is evaluated. The resulting rows and columns determine which elements of the high-dimensional tensors are evaluated at the next site. A macro iteration or *sweep* entails performing the previously described micro-iterations for every tensor core. The sweeps are performed back and forth over all tensor cores until convergences is reached. We refer the interested reader for more details to a recent in-depth review of tensor cross approximation by N. Fernández et al.[30] This work also presents a partially rank-revealing LU-decomposition as a more stable alternative to the maxvol algorithm.

3 The Tensor Train Conformer Search (TTConf)

In this work, we present the tensor train conformer search (TTConf), a tool for the global optimization of molecular conformations. TT optimization is used to screen many local minima in internal coordinates of the molecule. We combine it with gradient-based local optimization to converge energies within local minima and thereby reduce the required TT ranks. For

the scope of this work, we will use BAT coordinates to describe molecular conformations and use a subset of dihedral angles, i.e., the ones characterizing single bond rotations, to span the conformationally relevant coordinate space. While BAT coordinates provide a natural description for a wide variety of molecules, other kinds of internal coordinates may be better suited for certain substructures like rings. The design of more elaborate internal coordinates, however, will be part of future work. For this work, we simply work in the space of torsion angles for the global optimization and combine it with a local geometry optimization to yield the entries of the high dimensional tensor in Eq. 1. The quality of conformers found by TTConf is expected to rely heavily on the chosen theory level that defines the PES. TTConf works with any kind of energy function that could be coming from force field, SQM, and *ab inito* approaches. In this work, the single-point evaluations and gradient optimization are carried out using the GFN family of methods which provide a good trade off between speed and accuracy. To perform geometry optimizations in parallel the `mdopt` subroutine of CREST (version 3.0.1)[52] is used in combination with the interface to the `tblite` package.[53] In the following, we will describe the details of the TTConf conformer sampling algorithm as outlined in Fig. 1.

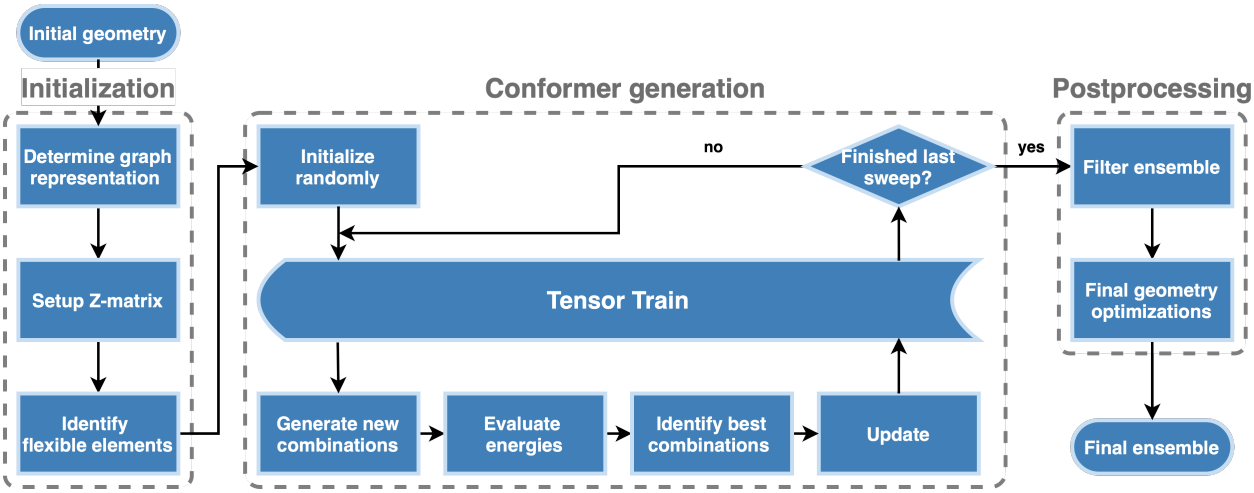

Figure 1: General workflow of the TTConf conformer sampling algorithm.

3.1 Determining the Molecular Graph Representation

Based on a user-provided input structure (in Cartesian coordinates), an initial gradient optimization is performed at the requested level of theory to ensure a chemically reasonable starting geometry. From the resulting geometry, covalent bonds are identified from which a molecular graph is generated. The graph setup and analysis are realized with the Python package `networkx`, which facilitates the determination of the pairwise shortest-path distances and identification of ring structures.[54]

3.2 Setup of Internal Coordinates

Z-matrix-type coordinates are often a preferable choice as they allow to describe the natural motion of the molecule by mostly independent variation of the internal coordinates. The lower the correlation between the coordinates, the lower is the required rank in the TT optimization. In TTConf, the Z-matrix is constructed in such a way that every bond rotation is described by exactly one proper dihedral angle facilitating conformer generation, while every remaining atom is referenced via improper dihedrals. The order of the atoms is based on their respective shortest-path distance from the origin, which is chosen to be one of the two most distant atoms (futher details in the Supporting Information Sec. 1).

3.3 Identification of Relevant Dihedrals

Within the set of internal coordinates, we need to identify dihedral angles, that correspond to rotatable bonds. For this purpose, the Wiberg bond orders (WBOs) at the GFN2-xTB level of theory are evaluated. All bonds with WBO$<$1.1 are characterized as single bonds and considered to be rotatable, while bonds with higher WBOs are assumed to not contribute to the conformational degrees of freedom. The only exceptions for this criterion are hydroxyl groups, which are rotated regardless of the respective bond order, to allow for modification of carboxyl- and phosphate groups. As for example, the rotation of methyl groups primarily

results in different rotamers, those bond rotations can be excluded from the conformer search. For this purpose, tetrahedrally coordinated atoms are identified and their direct neighboring atoms are checked for equivalence using atomic priorities as implemented in the `molbar` package.[55] If three neighbors have the same priority, the respective bond rotation can only lead to rotamers. Hence, this torsion is discarded for the conformational sampling. This provides a more general criterion that covers not only methyl, but also *tert.*-butyl and CF_3 groups as well. In addition to bond rotations, asymmetrically substituted nitrogen centers are identified that can be inverted by modification of the improper dihedral angle of one of its substituents (nitrogen inversion). Finally, flexible ring structures are identified which require separate treatment to avoid ring opening when changing dihedral angles in the ring. The description of rings is detailed in Sec. 3.6. The order of relevant dihedrals within the TT representation is equal to their order within the Z-matrix. Since the Z-matrix follows the longest chain of atoms, neighboring bonds are usually described via adjacent nodes in the TT. However, at branching points in the molecule, this cannot be realized.

3.4 Conformer Generation

The TT cross approximation facilitates the conformer search task by decomposing it into sequential updates of dihedrals of individual parts of the molecule. If there is no information on favorable conformations available in the beginning, the TT is initialized randomly and updated after each step. More specifically, the tail configurations contain only random numbers in the first sweep and subsequentially get replaced by the optimum values when passing through the TT. The evaluation of a core tensor involves generation of new geometries based on the respective combinations of dihedral angles, calculation of the corresponding energies (after a local geometry optimization) according to Sec. 3.5, and selection of the r best head conformations to update the TT (Fig. 2). Processing every tensor core once is referred to as one sweep. After a sweep is completed, the search direction is reversed. TTConf terminates after a specified number of sweeps has been completed.

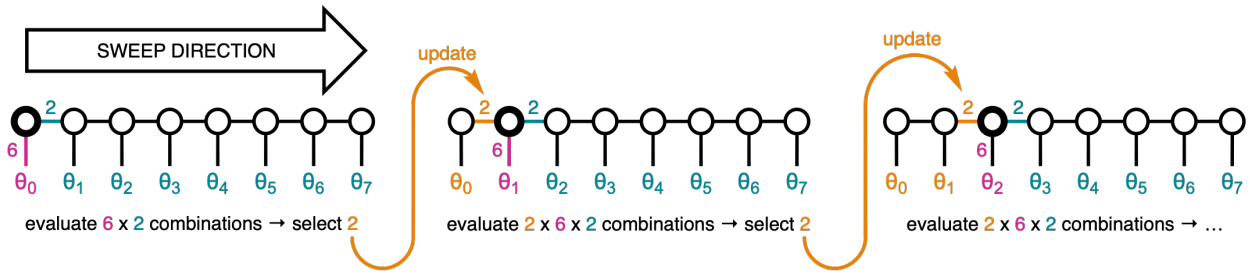

Figure 2: Representation of the overall global minimization procedure.

The conformer generation for a single core tensor is illustrated by the example of Raltitrexed in Fig. 3 A. In this example the head conformations are combinations of θ_0 and θ_1, the tail conformations include θ_3 to θ_7, while all allowed values of θ_2 are sampled. The respective combinations result in the third order tensor depicted in Fig. 3 B.

The subsequent selection of the new head conformations is shown in Fig. 4. For this purpose, the respective energies are calculated and the corresponding third order tensor is flattened along the dimension of tail conformations. Within the resulting matrix the new head conformations are selected based on the two columns leading to the maximum volume submatrix as described in Section 2.2.

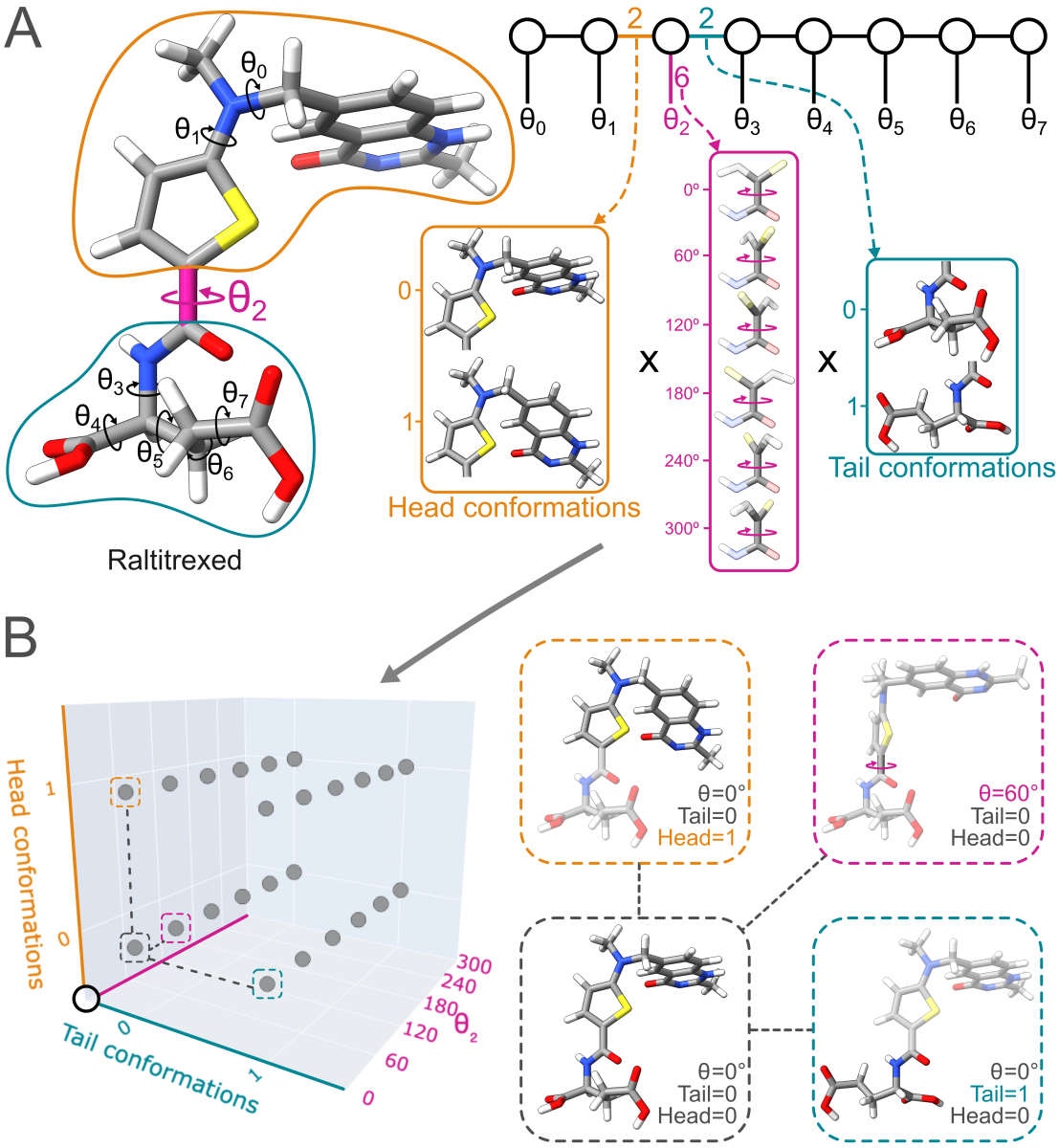

Figure 3: Conformer space of Raltitrexed represented in terms of a TT. As an example, we focus on the third rotatable bond — highlighted in pink (θ_2) — and chose 6 values for its dihedral angle and a rank of 2 for the other edges. At the third edge during the sweep through the TT, we already found two ($rank = 2$) head conformations, i.e. two sets of values for θ_0 and θ_1. These two head conformations are combined with 6 dihedral values for the current bond and two tail conformations, i.e., two sets of values for θ_3 through θ_7. The energy can be represented as three dimensional tensor with 24 elements, shown as cube in panel B. In panel B we additionally show selected structures indicating the changes along the 3 axes of the cube, where the parts of the molecule remaining constant are transparent.

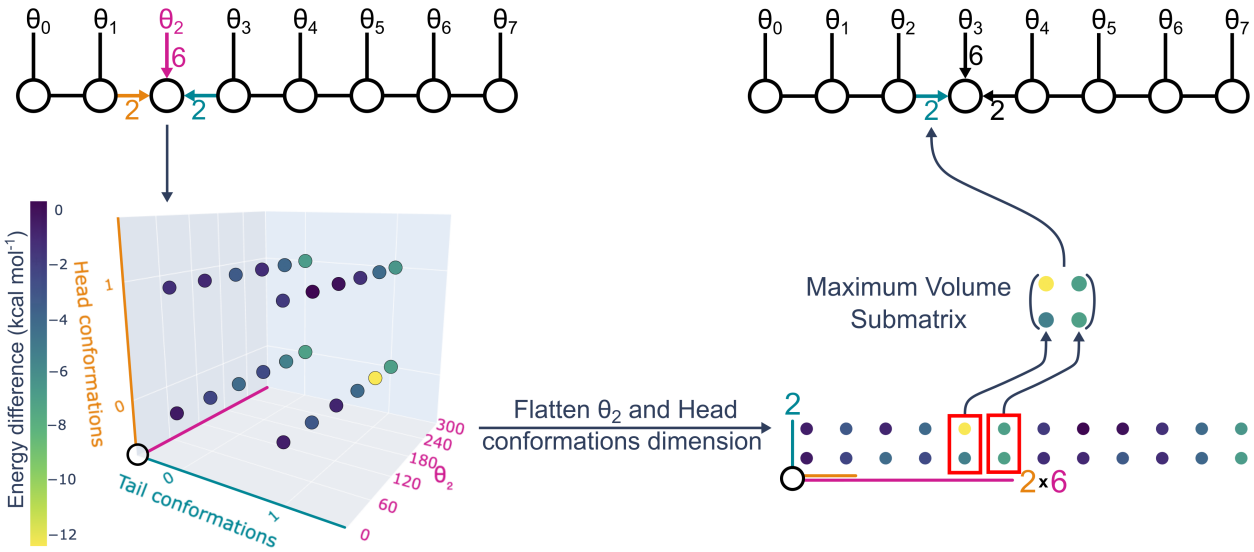

Figure 4: Update of the TT determining the new head conformations going into the fourth node. The cube at the third edge is flattened to a (2×12) matrix, where the colors indicate the relative energy of the conformer relative to the initial structure. The new head conformations are found as two columns of the matrix for which the volume of the (2×2) matrix is maximal.

3.5 Potential Energy Evaluation

To calculate the energy corresponding to a certain set of torsion angles, the initial (input) structure is modified in the respective internal coordinates and subsequently converted back to Cartesian coordinates. To keep the number of MM or QM calculations minimal, the adjacency matrix is calculated for the newly generated conformer (according to Sec. 3.1) and compared to the initial atomic connectivity. Any differences to the initial graph indicate clashing parts of the molecule, which most likely lead to unreasonable geometries after gradient optimization. Therefore, the respective geometries are not further considered and are assigned an arbitrary energy value of $+\infty$. Instead of computing only single-point energies at each generated conformer geometry, every considered geometry is optimized at the requested level of theory to its nearest local minimum. After re-evaluating the respective atomic connectivity, all optimized conformers with matching topology are added to a conformer ensemble. The ensemble is filtered for duplicates based on the pairwise energy difference, RMSD, and relative difference in rotational constants after every tensor core eval-

uation. This is done to reduce memory usage by storing only conformers within the requested energy window (default 6 kcal mol^{-1} relative to the lowest conformer found).

3.6 Generation of Ring Conformers

Ring structures need to be treated differently from acyclic bond rotations and nitrogen inversions, as an independent variation of dihedrals inside ring systems leads to bond breaking. Therefore, rings are identified in advance and the entire ring conformation is used as a single variable in the overall conformer search instead of individual torsion angles. For this purpose, a separate modified TT optimization is performed for every ring by variation of individual dihedral angles of the flexible bonds inside the ring. This obviously leads to ring-opened structures that are invalid candidates as conformers. To ensure intact ring structures after such an individual bond torsion modification, ring "repair" is achieved through a structure optimization with a harmonic force field as used in the structure idealization of the MolBar identifier.[55] This force field energy expression is based on harmonic potentials for every covalently bonded atom pair as defined by the initial adjacency matrix. Additionally, weak Coulomb-type repulsion is included for every non-bonded atom pair. Harmonic potentials are also included for every bond angle between three neighboring atoms and for improper (i.e. non-torsion) dihedral angles. Double and triple bonds are constrained via harmonic potentials for respective proper dihedral angles. The reference values for the harmonic potentials are calculated based on the initial molecular geometry (see Ref. 55 for details).

From the resulting ring conformer ensembles, up to 11 ring conformations are determined based on their corresponding conformer energy. Bond distances, angles, and torsions might all differ between those ring conformations, and are hence saved to characterize each ring conformer. Hence, the entire molecular conformational space used in the final TT optimization is now expressed as combination of individual bond rotations, nitrogen inversions, and different ring conformers. In our implementation, spiro- and bicyclic ring structures are treated as single fragment.

4 Computational details

All calculations were performed on an AMD EPYC 7763 using 8 threads and 16 GB of memory. Unless stated otherwise `xtb` 6.7.0,[56] CREST 3.0.1[52,53] and MolBar 1.1.1[57] were used. The TTConf calculations were performed with an in-house program written in Python.[58] All CREST MTD-based conformer searches were run with the default settings. The calculations were performed without solvation corrections.

5 Results and discussion

5.1 Benchmarking for Drug-like Molecules (CD25)

For a first assessment, the CD25 benchmark set introduced by Pracht and Grimme was chosen, as it provides a diverse set of organic molecules covering a wide range of molecular sizes and compositions (Fig. S3).[59] Those molecules were analysed with TTConf at three different settings of rank r and number of sweeps s using CREST as reference. We used the energy difference of the lowest TTConf and CREST conformers (ΔE) and the relative runtime to compare the performance of both conformational sampling approaches (Fig. 5).

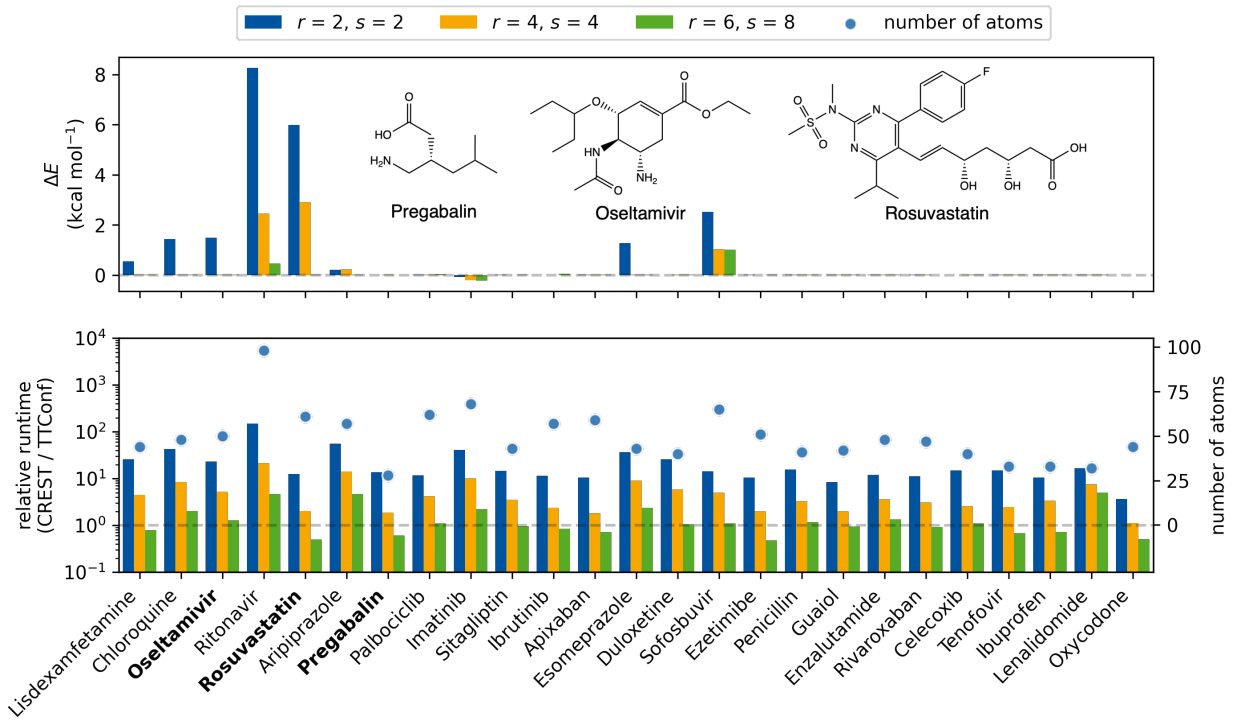

Figure 5: Energy deviation $\Delta E = E_{\text{method}} - E_{\text{min}}$ of the lowest conformer obtained with each method relative to the overall lowest one found (top). Relative runtimes and number of atoms of the molecules in the CD25 dataset. Three different TTConf settings are shown. r defines the considered rank in the TT optimization and s is the number of sweeps.

Considering the energy deviations we observe that the most basic setting ($r = s = 2$) already provides comparable accuracy to CREST for most of the benchmark molecules. Pregabalin is shown as one example in Fig. 5. On average, a speedup of an order of magnitude can be achieved with these settings compared to an MTD-based conformer search with the CREST. However, with increasing complexity of the molecule, the conformer search becomes more challenging, requiring an increase of r and s to maintain a high accuracy as can be seen, e.g., for Oseltamivir. Only a few cases require settings beyond $r = s = 4$ like Rosuvastatin. This is, to some extent, attributed to intramolecular non-covalent interactions as exemplified by the lowest energy conformers of Rosuvastatin (Fig. 6 a).

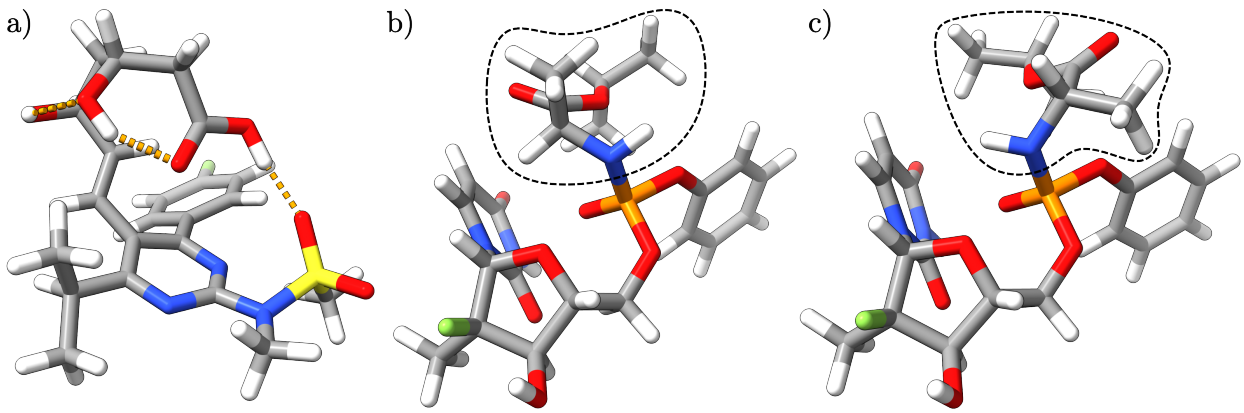

Figure 6: Panel a): lowest energy conformer of Rosuvastatin obtained with CREST and TTConf($r = 6, s = 8$) showing intramolecular hydrogen bonds. Finding this conformer with TTConf required a comparably large number of sweeps. Panel b): lowest energy conformer of Sofosbuvir obtained with CREST. Panel c): TTConf($r = 6, s = 8$) conformer of Sofosbuvir most similar to the CREST conformer. The main difference to panel b) is mostly due to the nitrogen inversion that had not been recognized in TTConf.

In this example, the structure dominating element is a hydrogen bond between two opposite ends of the molecule which requires the algorithm to recognize the favorable combination of several bond orientations eventually enabling the hydrogen bond. As this interaction of distant functional groups is due to a specific combination of multiple bond torsions and not well represented by two neighboring bond torsions alone, it is difficult to capture with low rank TT approximations.

Among the CD25 benchmark set, Sofosbuvir is the only example, where the energy deviation runs into a limit of approximately $1\,\mathrm{kcal\,mol^{-1}}$, while the computational demand and the associated runtime increases significantly. Comparing the generated CREST and TTConf ensembles reveals, that the TTConf ensemble includes a structure fairly similar to the lowest energy CREST conformer primarily differing in the configuration of one nitrogen atom (Fig. 6 b and c). In this case, the mentioned nitrogen atom was not recognized as invertible by TTConf, hence restricting direct access to the global minimum despite random inversion upon gradient optimizations. Therefore, TTConf is not able to invert the respective configuration systematically and reach the global minimum, which requires a more robust

identification algorithm of those nitrogen atoms in future developments.

5.2 Ring Sampling

While TTConf is primarily designed to sample acyclic structures, it also includes a generally applicable ring variation routine as outlined in Sec. 3.6. For a quick overview, the most common ring structures in natural products according to Chen *et al.* were investigated (molecules **1-14**) as well as some more complex ring structures (molecules **15-21**).[60] The respective structural formulas are shown in Fig. S2. The respective energy deviations and relative runtimes compared to the MTD-based CREST for the 21 tested molecules are shown in Fig. 7.

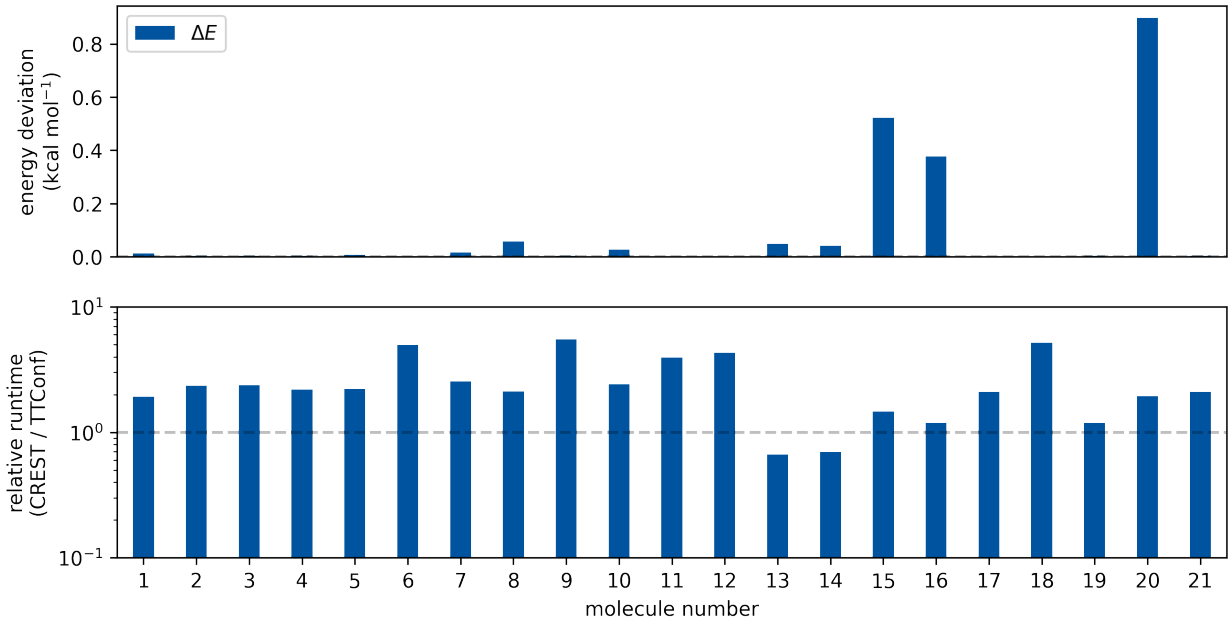

Figure 7: Energy deviations $\Delta E = E_{\mathrm{TTConf}} - E_{\mathrm{CREST}}$ and relative runtimes compared to CREST MTD-based conformer sampling of typical small ring structures ($r = 3, s = 8$). For the structural formulas, see Fig. S2

For most of the test cases, CREST and TTConf identified the same conformer as global minimum. The largest deviations occur particularly for nitrogen containing compounds **15** and **20** with an energy deviation of 0.5 and 0.9 kcal mol^{-1}, respectively. Comparing the respective conformer ensembles as shown in Fig. 8 for ring structure **20** reveals that the

major deviation originates from different configurations of nitrogen atoms as it was already observed for Sofosbuvir (Fig. 6 b and c). Since inversions are realized by modification of improper dihedral angles, endocyclic nitrogen atoms are prone to errors. Depending on how the Z-matrix is constructed the respective nitrogen atom can be part of the ring closing bond, making it impossible to define an improper dihedral angle and thus preventing inversion.

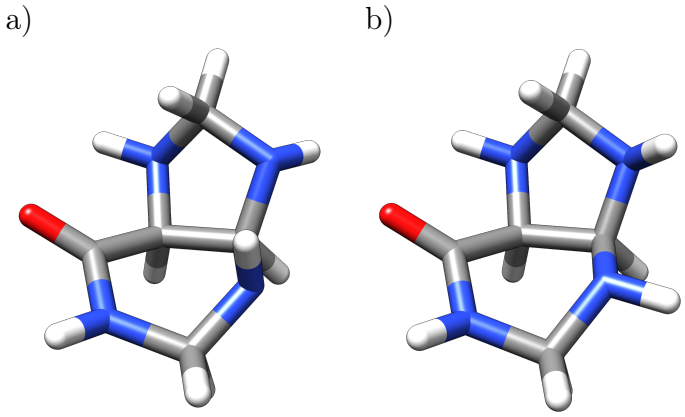

Figure 8: Structure comparison of a) the lowest energy CREST conformer and b) the most similar TTConf conformer of ring structure **20**, which showed the largest deviation in Fig. 7. The difficulty arises primarily from the handling of the nitrogen inversion in TTConf.

Considering the relative runtimes, rather small benefits for larger ring structures are found with TTConf. Especially bicyclic compounds (**13** and **14**) or spirocyclic compounds (**16** and **19**) take as long as or even slightly longer than CREST, while providing comparable accuracy.

5.3 Statistics on Large Conformational Dataset: BACE

To achieve a more comprehensive view of TTConf's performance, the BACE dataset was chosen comprising 1511 organic molecules ranging from 17 to 184 atoms. Different from CD25, this set also covers charged species.[61] As the dataset provides only conformational energies that were generated with CREST (version 2.9) but no information on the respective runtimes, the present comparison focuses solely on the accuracy of TTConf in finding the lowest energy conformer. For the TTConf runs, $r = 3, s = 8$ was chosen as a reasonable

compromise of runtime and accuracy (Sec. 5.4). To further interpret the origin for observed energy deviations, the dataset was divided based on the molecules' flexibility and degree of branching. As measure for the flexibility, the number of flexible torsions is used. The degree of branching is determined as shown in Fig. 9 taking into account only the molecule's flexible bonds. Out of the 1511 molecules, 15 molecules showed varying MolBar identifiers[55] in the reference CREST ensembles. Since this indicates changes in the topology and topography that are outside the pure conformational space, those molecules were excluded from the investigation.

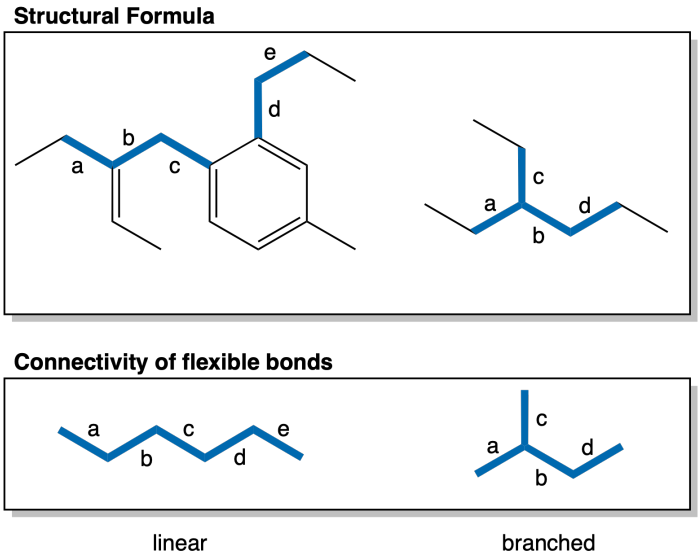

Figure 9: Definition of linear and branched structures according to the respective flexible bonds (highlighted in blue).

In Fig. 10, the distribution of the respective energy deviations of the lowest energy CREST and TTConf conformers are shown along with their average value. Regarding the entire dataset, the distribution is centered close to 0 kcal mol^{-1} primarily covering the range from -2 to 2 kcal mol^{-1} leading to an average deviation of only 0.58 kcal mol^{-1} in total. However, in several cases, deviations of >5 kcal mol^{-1} can be observed, where TTConf struggles to identify the global minimum. Vice versa, though less frequently, TTConf is also able to outperform CREST in some cases.

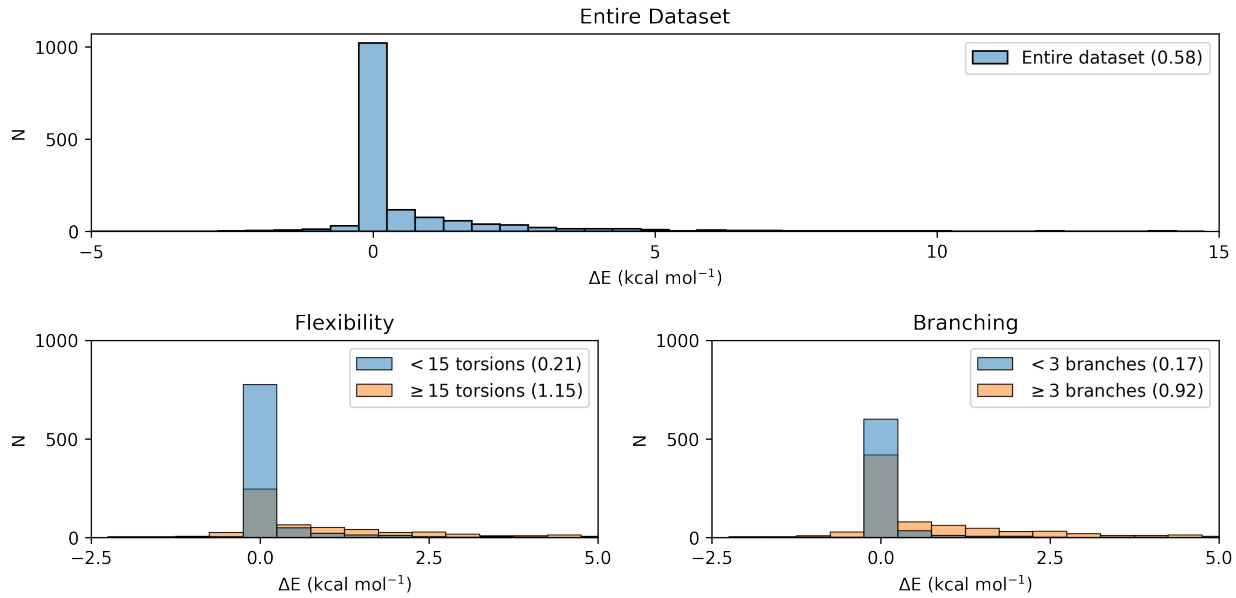

Figure 10: Distribution of the energy deviations of CREST and TTConf for the BACE dataset comprising 1511 organic molecules ranging from 17 to 184 atoms. Apart from the results for the entire dataset (top), the results for with subdivision for flexibility and branching is shown as well (bottom).

Taking a look at the influence of the system's flexibility and the number of branching points on the energy deviation, it becomes clear that TTConf performs better for smaller, predominantly linear molecules. As already indicated by the CD25 dataset, a higher rank approximation is beneficial, whenever correlations between distant bonds occur, e.g., when intramolecular non-covalent interactions play a role. This observation is confirmed, as higher deviations are found for molecules with more than 15 flexible bonds, where correlations between distant bonds are more likely. However, the more critical point seems be the molecular connectivity. While predominantly linear structures can be described with high accuracy at an average energy deviation of $0.17\,\mathrm{kcal\,mol^{-1}}$, strongly branched molecules lead to a much broader distribution including several outliers and therefore a higher average deviation of $0.92\,\mathrm{kcal\,mol^{-1}}$. This indicates a limitation of the inherently linear structure of the TT, since some neighboring bonds (at the branching point) are not found in consecutive order in the TT.

5.4 Optimal Settings

As the accuracy and time demand of the TTConf approach are strongly dependent on the assumed rank (r) and the number of sweeps (s), proper settings have to be determined to ensure reliable results at reasonable computational effort. Increasing r or s results in sampling of a larger portion of the conformational space, leading to a higher accuracy in finding the lowest energy conformer. While the choice of r influences the number of evaluated conformations quadratically, s contributes only linearly. To achieve the target accuracy, it is thus desirable to use an increased number of sweeps while working with a low-rank approximation. However, for an optimal cost-accuracy ratio r and s should be chosen as small as possible, while maintaining the desired accuracy. For this purpose, the CD25 benchmark set was sampled with a total of 20 different combinations of r and s using the mean ΔE and relative runtime (with respect to CREST's MTD conformer sampling) as performance criteria (Fig. 11).

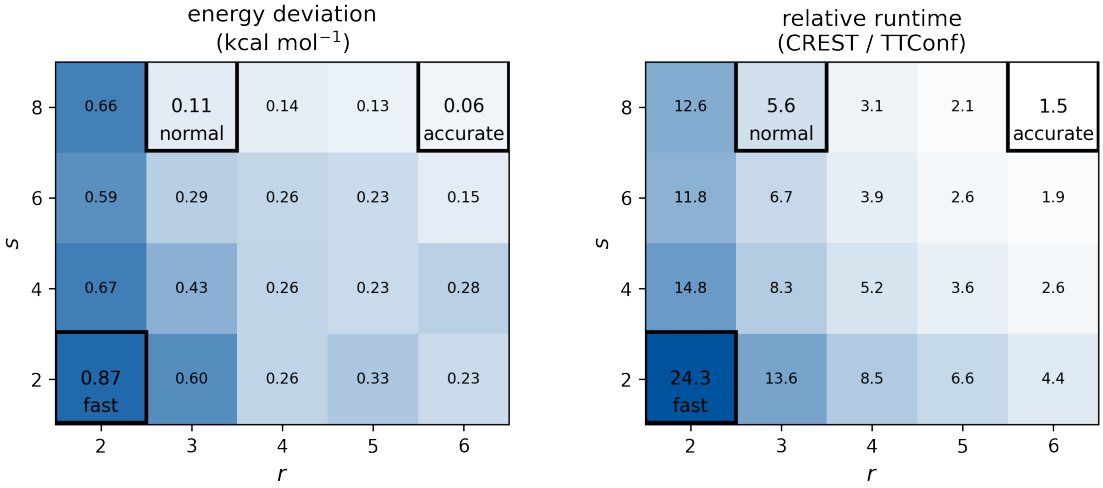

Figure 11: Assessment of the optimal settings for rank and number of sweeps based on the CD25 dataset. Comparison of the mean energy deviation of the respective lowest energy TTConf and CREST conformers (left) and the associated mean relative runtimes (right).

As expected, ΔE decreases when choosing a higher rank and number of sweeps, but significantly increases the runtime of the sampling procedure. While $r = 2$, $s = 2$ provides an average 24-fold speed up, the average ΔE of 0.87 kcal mol^{-1} indicates an insufficient

accuracy on average. As seen in Fig. 5, this deviation originates primarily from large and more branched molecules like Ritonavir (max. deviation of 8 and 6 $\mathrm{kcal\,mol^{-1}}$), while the majority of less branched molecules is handled with sufficient accuracy already at this level. The computationally most demanding setting of $r = 6, s = 8$ on the other hand shows excellent accuracy, but at a significantly reduced speedup of just 1.5 on average. The setting $r = 3, s = 8$ provides a good balance between cost and accuracy with an average deviation ΔE of 0.11 $\mathrm{kcal\,mol^{-1}}$ and an average speed up of 5.6 compared to CREST's MTD.

Based on these results, we suggest three default settings labeled *fast*, *normal* and *accurate* as shown in Fig. 11. Besides using fixed default settings, Fig. 5 indicates that a molecule-specific adjustment of the required rank and number of sweeps based on a molecule's flexibility and number of branches can be beneficial. For smaller, less branched molecules and smaller molecules like Pregabalin, the "fast" setting ($r = 2$, $s = 2$) is sufficient, while branched systems and longer chains (with potential for intramolecular noncovalent interactions) require more effort. In that case, the "normal" ($r = 3, s = 8$) or even "accurate" settings ($r = 6, s = 8$) should be used. In the supporting information, we provide additional results for selected molecules (Fig. S7).

6 Conclusions

We present a novel systematic conformer search approach for the investigation of drug-like molecules called tensor train conformer search (TTConf). Our approach addresses the inherent combinatorial explosion of grid search algorithms by using a low-rank tensor train (TT) representation of the high dimensional tensor of dihedral angle combinations. Through TT cross approximation this low-rank representation can be achieved with limited number of energy evaluations, which are typically the most time demanding task. As the algorithm allows the scaling to be reduced from exponential to polynomial, exploration of the low energy conformational space of drug-like molecules can be achieved at a significantly reduced time

demand compared to the current state-of-the-art metadynamics-based conformer search as implemented in CREST. The current applicability is primarily acyclic molecules, while it was shown that also different ring structures can be treated. This makes TTConf very suitable for sampling conformers of a broad variety of organic molecules. Future developments will focus on an improved ring sampling algorithm, tailoring the tensor network to the respective molecule for a better description of heavily branched structures. Furthermore, the approach is to be extended to metal-organic compounds.

Acknowledgement

The authors thank Michael Perelshtein for fruitful discussions and for pointing out errors in the manuscript. C.B. gratefully acknowledges support by the Ministry of Culture and Science of the State of North Rhine-Westphalia (MKW) through the NRW Rückkehrprogramm. P.P. acknowledges support by the German Academic Exchange Service (DAAD) for a postdoctoral research fellowship.

Supporting Information Available

More details on the setup of the molecular graph, additional figures with structural formulas of the benchmark compounds and additional benchmark results can be found in the supporting information.

References

(1) Bursch, M.; Mewes, J.-M.; Hansen, A.; Grimme, S. Best-Practice DFT Protocols for Basic Molecular Computational Chemistry. *Angew. Chem. Int. Ed.* **2022**, *61*, e202205735.

(2) Hawkins, P. C. D. Conformation Generation: The State of the Art. *J. Chem. Inf. Model.* **2017**, *57*, 1747–1756.

(3) Schlegel, H. B. Geometry optimization. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 790–809.

(4) Wales, D. J. Exploring Energy Landscapes. *Annual Review of Physical Chemistry* **2018**, *69*, 401–425.

(5) Saunders, M. Stochastic exploration of molecular mechanics energy surfaces. Hunting for the global minimum. *J. Am. Chem. Soc.* **1987**, *109*, 3150–3152.

(6) Wales, D. J.; Doye, J. P. K. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *Phys. Chem. A* **1997**, *101*, 5111–5116.

(7) Goedecker, S. Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems. *J. Chem. Phys. 120*, 9911–9917.

(8) Chang, G.; Guida, W. C.; Still, W. C. An internal-coordinate Monte Carlo method for searching conformational space. *J. Am. Chem. Soc.* **1989**, *111*, 4379–4386.

(9) Lipton, M.; Still, W. C. The multiple minimum problem in molecular modeling. Tree searching internal coordinate conformational space. *J. Comput. Chem.* **1988**, *9*, 343–355.

(10) Brooks, B. R.; Pastor, R. W.; Carson, F. W. Theoretically determined three-dimensional structure for the repeating tetrapeptide unit of the circumsporozoite coat protein of the malaria parasite Plasmodium falciparum. *Proc. Natl. Acad. Sci.* **1987**, *84*, 4470–4474.

(11) Elber, R.; Karplus, M. Multiple Conformational States of Proteins: A Molecular Dynamics Analysis of Myoglobin. *Science* **1987**, *235*, 318–321.

(12) van Gunsteren, W. F.; Berendsen, H. J. C. Computer Simulation of Molecular Dynamics: Methodology, Applications, and Perspectives in Chemistry. *Angew. Chem. Int. Ed.* **1990**, *29*, 992–1023.

(13) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671–680.

(14) Wilson, S. R.; Cui, W.; Moskowitz, J. W.; Schmidt, K. E. Conformational Analysis of Flexible Molecules: Location of the Global Minimum Energy Conformation by the Simulated Annealing Method. *Tetrahedron Lett.* **1988**, *29*, 4373–4376.

(15) Lelj, F.; Grimaldi, P.; Cristinziano, P. L. Conformational energy minimization by simulated annealing using molecular dynamics: Some improvements to the monitoring procedure. *Biopolymers* **1991**, *31*, 663–670.

(16) Hamelberg, D.; Mongan, J.; McCammon, J. A. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *J. Chem. Phys.* **2004**, *120*, 11919–11929.

(17) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci.* **2002**, *99*, 12562–12566.

(18) Grimme, S. Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations. *J. Chem. Theory Comput.* **2019**, *15*, 2847–2862.

(19) Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

(20) Pracht, P.; Grimme, S.; Bannwarth, C.; Bohle, F.; Ehlert, S.; Feldmann, G.; Gorges, J.; Müller, M.; Neudecker, T.; Plett, C.; Spicher, S.; Steinbach, P.; Wesołowski, P. A.;

Zeller, F. CREST—A program for the exploration of low-energy molecular chemical space. *J. Chem. Phys.* **2024**, *160*, 114110.

(21) Vásquez, M.; Scheraga, H. A. Use of buildup and energy-minimization procedures to compute low-energy structures of the backbone of enkephalin. *Biopolymers* **1985**, *24*, 1437–1447.

(22) Orús, R. Tensor networks for complex quantum systems. *Nat. Rev. Phys.* **2019**, *1*, 538–550.

(23) White, S. R. Density matrix formulation for quantum renormalization groups. *Phys. Rev. Lett.* **1992**, *69*, 2863–2866.

(24) Beck, M.; Jäckle, A.; Worth, G.; Meyer, H.-D. The multiconfiguration time-dependent Hartree (MCTDH) method: a highly efficient algorithm for propagating wavepackets. *Phys. Rep.* **2000**, *324*, 1–105.

(25) Manthe, U. A time-dependent discrete variable representation for (multiconfiguration) Hartree methods. *J. Chem. Phys.* **1996**, *105*, 6989–6994.

(26) Manthe, U. A multilayer multiconfigurational time-dependent Hartree approach for quantum dynamics on general potential energy surfaces. *J. Chem. Phys.* **2008**, *128*, 164116.

(27) Ellerbrock, R.; Hoppe, H.; Manthe, U. A numerically exact correlation discrete variable representation for multi-configurational time-dependent Hartree calculations. *J. Chem. Phys.* **2023**, *158*, 244103.

(28) Ellerbrock, R.; Hoppe, H.; Manthe, U. A non-hierarchical multi-layer multi-configurational time-dependent Hartree approach for quantum dynamics on general potential energy surfaces. *J. Chem. Phys.* **2024**, *160*, 224108.

(29) Oseledets, I.; Tyrtyshnikov, E. TT-cross approximation for multidimensional arrays. *Linear Algebra Appl.* **2010**, *432*, 70–88.

(30) Núñez Fernández, Y.; Ritter, M. K.; Jeannin, M.; Li, J.-W.; Kloss, T.; Louvet, T.; Terasaki, S.; Parcollet, O.; von Delft, J.; Shinaoka, H.; Waintal, X. Learning tensor networks with tensor cross interpolation: new algorithms and libraries. *arXiv e-prints* **2024**, arXiv:2407.02454.

(31) He, X.; Walker, B.; Man, V. H.; Ren, P.; Wang, J. Recent progress in general force fields of small molecules. *Curr. Opin. Struct. Biol.* **2022**, *72*, 187–193.

(32) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(33) He, X.; Man, V. H.; Yang, W.; Lee, T.-S.; Wang, J. A fast and high-quality charge model for the next generation general AMBER force field. *J. Chem. Phys.* **2020**, *153*, 114502.

(34) Lu, C.; Wu, C.; Ghoreishi, D.; Chen, W.; Wang, L.; Damm, W.; Ross, G. A.; Dahlgren, M. K.; Russell, E.; Von Bargen, C. D.; Abel, R.; Friesner, R. A.; Harder, E. D. OPLS4: Improving Force Field Accuracy on Challenging Regimes of Chemical Space. *J. Chem. Theory Comput.* **2021**, *17*, 4291–4300.

(35) Boothroyd, S.; Behara, P. K.; Madin, O. C.; Hahn, D. F.; Jang, H.; Gapsys, V.; Wagner, J. R.; Horton, J. T.; Dotson, D. L.; Thompson, M. W.; Maat, J.; Gokey, T.; Wang, L.-P.; Cole, D. J.; Gilson, M. K.; Chodera, J. D.; Bayly, C. I.; Shirts, M. R.; Mobley, D. L. Development and Benchmarking of Open Force Field 2.0.0: The Sage Small Molecule Force Field. *J. Chem. Theory Comput.* **2023**, *19*, 3251–3275.

(36) Lim, V. T.; Hahn, D. F.; Tresadern, G.; Bayly, C. I.; Mobley, D. L. Benchmark assessment of molecular geometries and energies from small molecule force fields. *F1000Res.* **2020**, *9*, 1390.

(37) Tian, C.; Kasavajhala, K.; Belfon, K. A. A.; Raguette, L.; Huang, H.; Migues, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q.; Simmerling, C. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput.* **2020**, *16*, 528–552.

(38) Zgarbová, M.; Šponer, J.; Jurečka, P. Z-DNA as a Touchstone for Additive Empirical Force Fields and a Refinement of the Alpha/Gamma DNA Torsions for AMBER. *J. Chem. Theory Comput.* **2021**, *17*, 6292–6301.

(39) Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. GLYCAM06: A generalizable biomolecular force field. Carbohydrates. *J. Comput. Chem.* **2008**, *29*, 622–655.

(40) Dickson, C. J.; Walker, R. C.; Gould, I. R. Lipid21: Complex Lipid Membrane Simulations with AMBER. *J. Chem. Theory Comput.* **2022**, *18*, 1726–1736.

(41) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell Jr., A. D. CHARMM36: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Biophys. J.* **2017**, *112*, 175a–176a.

(42) Hart, K.; Foloppe, N.; Baker, C. M.; Denning, E. J.; Nilsson, L.; MacKerell, A. D. J. Optimization of the CHARMM Additive Force Field for DNA: Improved Treatment of the BI/BII Conformational Equilibrium. *J. Chem. Theory Comput.* **2012**, *8*, 348–362.

(43) Guvench, O.; Mallajosyula, S. S.; Raman, E. P.; Hatcher, E.; Vanommeslaeghe, K.; Foster, T. J.; Jamison, F. W. I.; MacKerell, A. D. J. CHARMM Additive All-Atom Force Field for Carbohydrate Derivatives and Its Utility in Polysaccharide and Carbohydrate–Protein Modeling. *J. Chem. Theory Comput.* **2011**, *7*, 3162–3180.

(44) Yu, Y.; Krämer, A.; Venable, R. M.; Brooks, B. R.; Klauda, J. B.; Pastor, R. W. CHARMM36 Lipid Force Field with Explicit Treatment of Long-Range Dispersion:

Parametrization and Validation for Phosphatidylethanolamine, Phosphatidylglycerol, and Ether Lipids. *J. Chem. Theory Comput.* **2021**, *17*, 1581–1595.

(45) Folmsbee, D.; Hutchison, G. Assessing conformer energies using electronic structure and machine learning methods. *Int. J. Quantum Chem.* **2021**, *121*, e26381.

(46) Spicher, S.; Grimme, S. Robust Atomistic Modeling of Materials, Organometallic, and Biochemical Systems. *Angew. Chem. Int. Ed.* **2020**, *59*, 15665–15673.

(47) 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.

(48) Oseledets, I. V. Tensor-Train Decomposition. *SIAM J. Sci. Comput.* **2011**, *33*, 2295–2317.

(49) Sozykin, K.; Chertkov, A.; Schutski, R.; Phan, A.-H.; Cichocki, A. S.; Oseledets, I. TTOpt: A Maximum Volume Quantized Tensor Train-based Optimization and its Application to Reinforcement Learning. Advances in Neural Information Processing Systems. 2022; pp 26052–26065.

(50) Tyrtyshnikov, E. Incomplete Cross Approximation in the Mosaic-Skeleton Method. *Computing* **2000**, *64*, 367–380.

(51) Civril, A.; Magdon-Ismail, M. On selecting a maximum volume sub-matrix of a matrix and related problems. *Theor. Comput. Sci.* **2009**, *410*, 4801–4811.

(52) Conformer-Rotamer Ensemble Sampling Tool. `https://github.com/crest-lab/crest`, last accessed on September 23nd 2024.

(53) Light-weight tight-binding framework. `https://github.com/tblite/tblite`, last accessed on September 23nd 2024.

(54) Hagberg, A. A.; Schult, D. A.; Swart, P. J. Exploring Network Structure, Dynamics, and Function using NetworkX. Proceedings of the 7th Python in Science Conference. Pasadena, CA USA, 2008; pp 11 – 15.

(55) Van Staalduinen, N.; Bannwarth, C. MolBar: A Molecular Identifier for Inorganic and Organic Molecules with Full Support of Stereoisomerism. `https://chemrxiv.org/engage/chemrxiv/article-details/65e3cd80e9ebbb4db9c71da0`.

(56) Semiempirical Extended Tight-Binding Program Package. `https://github.com/grimme-lab/xtb`, last accessed on September 23nd 2024.

(57) MolBar. `https://git.rwth-aachen.de/bannwarthlab/molbar`, last accessed on September 23nd 2024.

(58) The program can be requested from the authors.

(59) Pracht, P.; Grimme, S. Calculation of absolute molecular entropies and heat capacities made simple. *Chem. Sci.* **2021**, *12*, 6551–6568.

(60) Chen, Y.; Rosenkranz, C.; Hirte, S.; Kirchmair, J. Ring systems in natural products: structural diversity, physicochemical properties, and coverage by synthetic compounds. *Nat. Prod. Rep.* **2022**, *39*, 1544–1556.

(61) Subramanian, G.; Ramsundar, B.; Pande, V.; Denny, R. A. Computational Modeling of β-Secretase 1 (BACE-1) Inhibitors Using Ligand Based Approaches. *J. Chem. Inf. Model.* **2016**, *56*, 1936–1949.