Conformator: A Novel Method for the Generation of Conformer Ensembles

*Nils-Ole Friedrich,[1] Florian Flachsenberg,[1] Agnes Meyder,[1] Kai Sommer,[1] Johannes Kirchmair,[1,2,3] Matthias Rarey[1]**

[1] Universität Hamburg, Center for Bioinformatics, Bundesstr. 43, Hamburg, 20146, Germany

[2] Department of Chemistry, University of Bergen, N-5020 Bergen, Norway

[3] Computational Biology Unit (CBU), University of Bergen, N-5020 Bergen, Norway

*E-mail: rarey@zbh.uni-hamburg.de. Tel.: +49 40 42838-7351.

ABSTRACT

Computer-aided drug design methods, such as docking, pharmacophore searching, 3D database searching and the creation of 3D-QSAR models, need conformational ensembles to handle the flexibility of small molecules. Here we present Conformator, an accurate and effective knowledge-based algorithm for generating conformer ensembles. With 99.9% of all test molecules processed, Conformator stands out by its robustness with respect to input formats, molecular geometries and the handling of macrocycles. With an extended set of rules for sampling torsion angles, a novel algorithm for macrocycle conformer generation, and a new clustering algorithm for the assembly of conformer ensembles, Conformator reaches a median

minimum root-mean-square deviation (measured between protein-bound ligand conformations and ensembles of a maximum of 250 conformers) of 0.47 Å, with no significant difference to the highest-ranked commercial algorithm OMEGA and significantly higher accuracy than seven free algorithms, including the RDKit DG algorithm. Conformator is part of the NAOMI ChemBio Suite and is available as a standalone tool free for non-commercial use and academic research at https://software.zbh.uni-hamburg.de.

INTRODUCTION

Computational methods for 3D virtual screening, drug design and other applications depend on the ability of algorithms to represent the conformations that small molecules adopt upon binding to biomacromolecules. In particular, fast tools such as pharmacophore-based and shape-focused screening engines make use of pre-calculated, multi-conformational databases composed of compounds represented by (preferably small) conformer ensembles.[1–4]

The generation of representative conformer ensembles of small molecules poses significant challenges. Small molecules can have a substantial number of conformational degrees of freedom.[5] Upon binding, they may adopt conformations that are distinct from the low-energy conformations observed in the gas phase and in solution, such as strained conformations related to transition states.[6–9] On top of that, what constitutes the most appropriate algorithm for conformer ensemble generation depends on the specific purpose of use: fast algorithms may be preferred for sampling large molecular libraries for use with, for example, coarse virtual screening approaches such as pharmacophore models, whereas more time-consuming but more accurate algorithms are generally preferred for sampling small sets of molecules to be used e.g. for 3D QSAR. In consequence, a large number of conformer ensemble generators based on

various algorithmic approaches are available today. They are based, among others, on random and systematic search algorithms, molecular dynamics (MD) simulations, genetic algorithms (GA), distance geometry (DG) and knowledge-based approaches.[10] Two recent studies from our labs[11,12] directly compare the performance of seven free (the RDKit DG algorithm[13] and the Experimental-Torsion basic Knowledge Distance Geometry algorithm (ETKDG)[14], Confab,[15] Frog2,[16] Multiconf-DOCK[17] and the Balloon DG and GA algorithms[18]) and eight commercial (ConfGen,[19] ConfGenX,[20] cxcalc,[21] iCon,[22] MOE LowModeMD,[23] MOE Stochastic, MOE Conformation Import and OMEGA[24]) conformer ensemble generators. These studies were the first to employ comprehensive sets of high-quality structures of protein-bound ligands for benchmarking. In particular, a newly developed cheminformatics pipeline was utilized for the fully automated extraction and curation of a complete set of 10,936 high-quality structures of protein-bound ligands ("Sperrylite Dataset"[5]) from a total of over 350k ligand conformations (from structures deposited in the PDB). The support of the individual atoms of all ligands by the measured electron density was quantified by the electron density score for individual atoms (EDIA[25]). Based on the Sperrylite Dataset, a diverse subset of 2859 high-quality structures of unique ligands bound to their biomacromolecular targets ("Platinum Diverse Dataset"[12]) was compiled and provided to the scientific community for benchmarking. The outcomes of these studies show that commercial algorithms generally obtain higher accuracy and robustness than their free counterparts. OMEGA was confirmed as the leading commercial algorithm, with the distance geometry approach of RDKit and its knowledge-based counterpart, ETKDG, as the best-performing free alternatives.[11,12] Importantly, for all of the tested free algorithms severe geometrical errors related to wrong bond lengths and bond angles, as well as out-of-plane errors, were detected in the generated conformations. In contrast, for most of the tested commercial

algorithms only a few instances of anomalous geometries were observed. For OMEGA and iCon no geometric errors were identified.

In this work we introduce Conformator as a new conformer ensemble generator that is free for non-commercial use and academic research, and which addresses several of the limitations shared by most of the existing free algorithms. Conformator is a knowledge-based conformer ensemble generator that builds on concepts of the previously introduced CONFECT algorithm.[26] Major conceptual advancements of Conformator over CONFECT include a novel approach to sampling the conformational space of macrocycles, a new efficient clustering algorithm, an extended set of rules for sampling torsion angles, and capabilities for handling SMILES and InChI input. Together with the revised and extended torsion angle library of Guba et al.[27] these advancements make Conformator a highly accurate and effective algorithm that stands out by its robustness with respect to input formats, molecular geometries and the handling of macrocycles.

METHODS

Conformer Generation Algorithm

Conformator is a conformer ensemble generator built on established concepts of incremental construction of conformers. At its core, Conformator consists of a torsion driver enhanced by an elaborate algorithm for the assignment of torsion angles to rotatable bonds, plus a new clustering component that compiles ensembles efficiently by taking advantage of the fact that the lists of generated conformers are partially presorted. The clustering algorithm minimizes the number of comparisons between pairs of conformers that are required in order to effectively derive individual RMSD thresholds for molecules and to compile the ensemble.

Conformator features two conformer ensemble generation modes, “Fast” and “Best”. As their names suggest, the emphasis of Fast is on computational efficiency whereas that of Best is on accuracy. Both modes include checks that ensure chemically correct bond lengths and bond angles, as well as the planarity of conjugated systems including rings.

Conformator reads molecular structures from SD and MOL2 files as well as from SMILES and InChI notations. By default, Conformator generates a new set of 3D atom coordinates as a starting point for conformation generation. Thus, Conformator does not rely on input coordinates and generates a canonicalized order of atoms and bonds (similar to canonical SMILES)[28]. This representation serves as a unique and independent starting point for conformer ensemble generation (Figure 1).

After parsing, the molecule is compartmentalized at any acyclic, non-terminal single bond that is not connected to a methyl, trifluoromethyl or nitrile group (following the concept of rigid rotor approximation). Each of these single bonds are assigned all torsion angle values of matching fragments recorded in the torsion angle library developed by Schärfer et al.[29] and revised by Guba et al.[27] As part of the construction of conformers, optimal bond angles based on the Valence Shell Electron Pair Repulsion (VSEPR) model are assigned.[30,31] Bond lengths of acyclic adjacent atoms used in the construction of conformers are calculated from the sum of covalent radii. They are adjusted for different atom types, taking into account the local molecular environment (e.g. delocalization). Details on the exact procedure and exceptions are reported in ref 26.

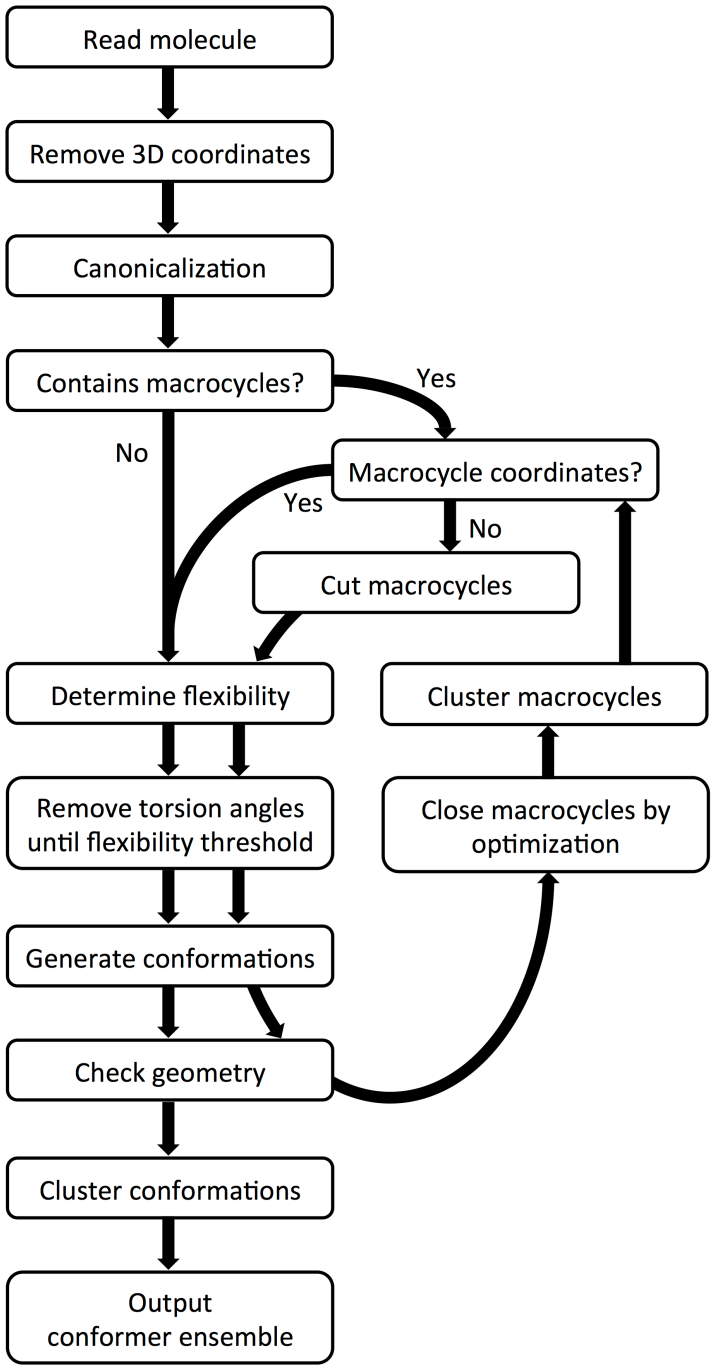

Figure 1. Schematic depiction of the conformer ensemble generation approach followed by Conformator. The boxes show the major algorithmic steps including the loop for macrocycle conformer generation.

Once all possible torsion angles have been assigned based on this SMARTS pattern matching procedure,[32] individual torsion angle values are removed during an iterative process until the maximum number of possible conformers (based on the combination of all assigned torsion angles, neglecting potential clashes) no longer exceeds the maximum number of generated candidate conformers for clustering. The number of torsion angles assigned to a rotatable bond depends on the bond's centricity in the molecule, the overall flexibility of the molecule, and the sampling parameters defined by the user (such as the maximum ensemble size). The centricity is estimated from the topological distance of the rotatable bond to the farthest atoms calculated on the molecular graph with the Floyd-Warshall algorithm.[33] Rotatable bonds located at the center of a molecule are assigned more alternative torsion angle values compared to rotatable bonds of terminal fragments. This is because fragments close to the center of a molecule are more likely to have a determinant effect on the overall conformation. More specifically, fragments located at the center of a molecule keep many if not all torsion angles recorded for a specific SMARTS pattern in the torsion angle library whereas fragments located away from the center of the molecule are assigned only a few of the most frequently observed torsion angles. The overall aim of this procedure is the reduction of the number of conformers to be generated and analyzed during the clustering process (typically hundreds of thousands or even millions of conformations) by two to three orders of magnitude. The flexibility of a molecule is estimated based on the maximum number of possible conformations resulting from the enumeration of all torsion angle values stored in the library (without the consideration of potential clashes). The maximum number of generated candidate conformers for clustering is the product of the maximum allowed ensemble size (user-adaptable parameter; in this study 50 or 250) and a factor of 10 (Fast) or 20 (Best).

Once all torsion angles for conformer enumeration have been selected, the conformer generation process is initiated, starting from the most central fragment and following a standard incremental construction approach.[34] Initially, a depth-first search of the most likely torsion angles is carried out in order to ensure that the most relevant torsion angles are represented in the conformer ensemble and that the conformer generation produces the conformers which are likely most relevant. Provided that the number of conformers resulting from this depth-first search does not exceed the maximum number of candidate conformers for clustering, breadth-first search (starting again from the most central fragment) is carried out iteratively to explore all selected torsion angles and, hence, generate additional candidate conformers.

During conformer generation, topological symmetry classes of each heavy atom of the molecule are calculated in a canonical way using a variant of the CANON algorithm.[35] Based on these, local symmetries are detected and considered during torsion angle enumeration in order to avoid the generation of duplicate conformers. Since local symmetry detection depends on the used central fragment, not all symmetries can be detected and a final symmetry clustering via complete automorphism enumeration is performed to remove similar conformers due to global symmetries.

Conformations for rings formed by up to nine heavy atoms are calculated using conformations from a ring template library embedded in NAOMI[36] as described by Schärfer et al.[26] Ring systems are incrementally constructed from individual ring conformers. Following the concept of unique ring families (URFs) reported by Kolodzik et al.[37] (a recent reimplementation by Flachsenberg et al.[38] was used for Conformator), at most one relevant cycle (RC) per URF is selected for ring system conformation generation. Starting from the RC with the highest connectivity, the remaining cycles are attached while considering atom geometries according to

VSEPR and taking into account the available stereo information. Within a tailored optimizer, simplified force field terms for bond distortion, angle bending and torsion energy are used for evaluating the deviations of molecular geometries from the ideal values and for assessing steric clashes. The tailored optimizer subsequently relaxes the assembled ring system conformation.

This optimizer is also used to generate additional low-energy conformations based on initial template conformations to generate an ensemble of ring system conformations. Rings formed by more than nine atoms are handled by a new algorithm for sampling the conformations of macrocycles (see Conformer Generation for Macrocycles).

Conformations causing clashes are rejected as early as possible during the incremental construction process. Intramolecular clashes are defined as overlaps of more than 30% of the van der Waals radii of 1-4-connected (or more distant) heavy atom pairs that are not part of the same ring system. Alternatively, users can choose for Conformator to include hydrogen atoms in the clash calculation.

The configuration of any defined stereogenic centers is preserved by the algorithm, whereas the configuration of any undefined R/S-stereogenic centers is arbitrarily chosen once per molecule. Undefined E/Z-stereogenic centers are enumerated (limited only by steric hindrances and the maximum ensemble size). In the case of undefined stereogenic centers, the macrocycle conformation generation (see section "Conformer Generation for Macrocycles") may produce a mix of stereoisomers (R/S and E/Z). Arbitrarily selecting one stereoisomer could prevent the algorithm from finding any reasonable result, especially in the case of E/Z isomers.

Clustering of Conformers

A new algorithm based on sphere exclusion clustering[39,40] was developed as part of Conformator for the efficient assembly of conformer ensembles (Algorithm S1, Figure S1). The clustering algorithm is the final step of the conformer ensemble generation. It aims to reduce the number of computationally expensive geometric comparisons of pairs of conformers required for the assembly of ensembles of a defined maximum size by exploiting the fact that sequentially generated conformers are likely to be highly similar to each other. To an outside observer the list of conformers generated by Conformator will appear to be the result of a systematic search which explores valid torsion angles for one rotatable bond after the other. Geometric deviations between pairs of sequentially generated conformers are likely small because they often differ only by one torsion angle. Large deviations are less common and are often related to clashes which, when occurring during early stages of the search, can result in the rejection of whole branches of the search tree. The number of comparisons (RMSD calculations) between conformers is heavily reduced by traversing the list of conformers forward and the list of cluster centers backwards. This increases the probability of similar conformers being compared early. When a similar enough conformer (defined by a RMSD threshold) is identified, the conformer is removed from the list of candidates and not compared to any further conformers.

During clustering, Conformator adjusts the minimum RMSD distance between conformers and determines an appropriate RMSD threshold for each individual molecule in order to generate ensembles that do not exceed the maximum ensemble size. This RMSD threshold depends on the maximum ensemble size and quality level, as well as the size and flexibility of the molecule. The algorithm is heuristic but deterministic, i.e., it produces the same result given the same list of conformations (note that, unless the user requests that input coordinates be used as a starting

point for conformer generation, the list of conformations generated during each run is identical for a given molecule).

Conformator does not rank conformers explicitly (although the first conformers generated by the algorithm are more likely based on the most commonly observed torsion angles). The conformers of an ensemble of small size (e.g. five conformers) will not necessarily be part of an ensemble of larger size (e.g. 50 conformers) because for small ensembles Conformator may prioritize conformers of high diversity over conformers with more commonly observed torsion angles. It is also unlikely that the first few conformers of an ensemble of larger size are those that would be included in an ensemble of small size. For this reason, in order to obtain ensembles of desired size, users are advised to not extract individual conformers but to define an adequate maximum ensemble size prior to ensemble generation.

The clustering algorithm (illustrated in Figure S1 and reported as pseudo code in Algorithm S1) involves the following key steps (with *radius* and *increase* having the values 0.1 Å and 0.05 Å for Best, and 0.5 Å and 0.5 Å for Fast):

1. An empty list of cluster centers is created.
2. The first conformation becomes the first cluster center.
3. Each conformer in the list of conformers is compared to the reversed list of cluster centers.
4. If the conformer is
    - a) similar to an existing cluster center (RMSD smaller than *radius*), then the conformer is immediately discarded.

    - b) dissimilar to any of the existing cluster centers, then the conformer is added to the list of cluster centers.
5. If the number of cluster centers reaches the maximum ensemble size, *radius* is increased as specified by the *increase* parameter and the clustering process is restarted with an empty list of cluster centers and the list of remaining conformers.
6. When all conformers are assigned to a cluster center and the ensemble size is equal to or below the maximum ensemble size, the list of cluster centers is reported as the conformer ensemble.

Conformer Generation for Macrocycles

Conformers for macrocyclic ring systems are generated using a novel algorithm. First, all macrocycles are sliced by cutting bonds until no macrocycles are left. Next, conformations are generated for these structures without macrocycles, which serve as starting points for the rebuilding of the macrocycles by a local optimization algorithm. The following sections describe these processes in detail. Schematics of the conformer generation algorithm for macrocycles are provided in Figure S2.

Preprocessing of Macrocyclic Structures for Conformer Generation

In the following, all rings formed by more than nine atoms are termed *macrocycle*; all others are termed *small rings*. This distinction is necessary because conformations for small rings are covered by the ring template library (see Conformer Generation Algorithm). The concept of unique ring families (URFs)[37,38] is used to consider one ring family at a time instead of processing individual rings. URFs are a unique, chemically meaningful and polynomial description of the rings in a molecule.

First, all URFs of the molecule are identified.[37,38] An URF is called macrocyclic if it contains at least one ring with more than nine atoms. All ring systems are processed independently. All *macrocyclic* URFs in a ring system are iteratively cut at one single bond outside of *small rings* until the resulting ring system no longer contains any *macrocycles*. In case a molecule contains exactly one *macrocycle* this process results in the cutting of one bond. By choosing exactly one bond to be cut during each iteration, the molecule remains connected. The single bond to be cut is chosen by prioritizing carbon-carbon and then carbon-incident bonds. If no such bond exists, the same priority rule is applied to bonds in conjugated systems. Bonds that are not adjacent to small rings are favored in the selection process. Double bonds, triple bonds and bonds that are part of small rings are not cut. Macrocycles consisting entirely of small rings are incrementally constructed from individual ring conformers. Following the cutting of a bond, new single bonds equal in length to the original bond are introduced by attaching two dummy atoms.

Generation of Conformers for Preprocessed Macrocyclic Structures

Diverse conformations of the preprocessed macrocyclic structures are generated with Conformator's standard algorithm following the exact same procedure as described above (see Conformer Generation Algorithm; Figure 1).

Rebuilding the Macrocycles by Numerical Optimization

The conformations generated during the previous process are used as starting points for a gradient-based numerical optimization procedure that aims to reconstitute macrocycles by superimposing the dummy atoms with the atoms they replaced during the cutting step. Note that the initial conformations already have valid geometries at this point, obviously with the exception of the part where the macrocyclic bond is to be reintroduced. The optimization is

performed employing internal coordinates, namely the torsion angles and bond angles in the macrocycles. By this strategy the number of parameters is reduced down to at most one bond angle per atom and one torsion angle per bond.

Local optimization is performed using a reimplementation of the BFGS-B algorithm,[41,42] which was modified to not allow any atoms to move by more than 0.5 Å per iteration. This modification, inspired by recent work on the refinement of the positions of water molecules in protein crystal structures,[43] was made to increase the locality of the optimization method and avoid unreasonably large changes in geometry. The local optimization is performed only on the atoms of the macrocycle (all other atoms of the molecule are not considered) and no part of the macrocycle is fixed (except for individual atoms in small rings, which are moved as a unit).

The here introduced macrocyclic optimization score (MCOS, see Eq. (1)) is used to reconstruct the macrocycle. It includes several well-known components from common force fields and some components specific to the optimization of macrocycles. The formulae of the terms in Eq. (1) are provided in the Figures S3 to S9 in the SI, the weights were determined empirically and are provided in Table S1. Please note that the MCOS and the individual score contributions are dimensionless and are not genuine energy terms.

$$MCOS = w_{overlay}\, S_{overlay} + w_{bond}\, S_{bond} + w_{angle}\, S_{angle} + w_{limit}\, S_{limit} + w_{torsion}\, S_{torsion} + w_{torsion,conjugated}\, S_{torsion,conjugated} + w_{clash}\, S_{clash} \quad (1)$$

The overlay score given in Eq. (2) is the central part of the scoring function.

$$S_{overlay} = \sum_{\{i,j\} \in cutbonds} \frac{1}{2}(distance(i, dummy(i))^2 + distance(j, dummy(j))^2), \quad (2)$$

where $\{i,j\}$ is a cut bond and *dummy(j)* is the dummy atom replacing atom *j* as a terminal atom adjacent to atom *i*.

$S_{overlay}$ scores the distance between the dummy atoms and the atoms in the original macrocycle they replaced. Ideally, this distance should be close to 0 (see Figure S3). The overlay score ensures that the bond angle and bond length across the cut bond will be restored during local optimization. It also supports the preservation of local stereochemistry.

The bond angle term S_{angle} uses a harmonic potential (calculated on the angle cosine, see Figure S4) to account for deviations from the ideal values (see Conformer Generation Algorithm and ref 26). It is calculated only for bond angles directly altered during optimization (i.e. angles involving bonds along the macrocycle that are optimization parameters) and the angles involving the cut bonds. During local optimization, bond angles are box-constrained such that no bond angle may be set to values greater than 179 degrees (if the atom does not have linear VSEPR geometry) and smaller than 0 degrees. This is to prevent unreasonable bond angle changes or even inversions of the local stereochemistry as bond angles usually stay rather close to the respective ideal values. The bond angle constraints are further supported by the penalty S_{limit} in the scoring function for bond angles in macrocycles, which leads to a preference of bond angles between 30 and 150 degrees (see Figure S5). Both terms S_{angle} and S_{limit} are multiplied by a function (see Figure S7) that reduces the scores to 0 in cases where any bond length adjacent to the angle approaches 0 Å. This is necessary because bond angles are not defined in cases where two defining atoms are placed on top of each other.

In addition, the bond length term S_{bond} uses a harmonic potential (see Figure S6) to account for deviations from ideal values (see Conformer Generation Algorithm and ref 26). Only the bond lengths of the cut bonds are scored.

The torsion angle score for bonds within ($S_{torsion,conjugated}$) and outside ($S_{torsion}$) of conjugated systems is calculated using the same torsion angle potential but different weights. The (continuous) torsion angle potential is based solely on torsion angle peaks recorded in a freely available torsion angle library derived from the CSD.[27] It uses the von Mises function as the kernel for curve approximation[44] with a tailored equation for kappa. We estimate the curve width through connecting the second peak tolerance and the peak score from the torsion library with the measure of concentration of the von Mises function (kappa). Due to the numerical optimization steps in continuous torsion space, torsional angles may differ from the angles stored in the torsion library (note that the angles start from those stored in the torsion angle library).

The torsion angle potential is multiplied by a function (see Figure S8) that reduces the torsion angle score to 0 in cases where any bond angle along that torsion bond is either close to 0 or 180 degrees (such bond angle values may be observed for cut bonds where the bond angle is not directly modified and therefore not subject to the box constraints). This is necessary because the torsion angle, as a function of the four atom coordinates, has a discontinuity when three consecutive atoms are collinear. The torsion angle potential is furthermore multiplied by the same function described above for S_{angle} and S_{limit} that reduces the score to 0 in cases where bond lengths are close to 0 Å (Figure S7).

To prevent intramolecular clashes, the clash term S_{clash} was added to the MCOS. S_{clash} is a quadratic function that penalizes van der Waals overlaps between 1-4-connected (or further away) heavy atoms that exceed the threshold level of 30% (see Figure S9).

Postprocessing and Filtering of Macrocyclic Structures for Conformer Generation

Following the optimization procedure, the cut bonds are reintroduced to close the macrocycle conformations again, and the dummy atoms are removed. In the rare event that the resulting macrocycle has assigned a configuration that does not correspond to the conformation of the input structure, the conformer is rejected. The geometry of all atoms forming macrocycles is then checked and, if required, optimized to resemble VSEPR geometries by adjusting the position of the macrocycle substituents.

All *macrocycle* conformations are then checked for bond lengths and angles that deviate strongly from the known optimal value.[26] The optimal values for bond length and bond angles were the same as used for the optimization; for allowed deviations see ref 45. Furthermore, the planarity of conjugated macrocycles (e.g. protoporphyrin IX, *PP9*) is tested by checking their bonds for torsion angles deviating from 0 or 180 degrees. Since macrocycles can adopt highly strained conformations a maximum deviation of 20 degrees of torsion angles in conjugated macrocycles is allowed. Only in cases where no (approximately) planar conjugated system can be generated are non-planar alternative conformations considered.

Before utilizing the macrocycle conformations for ensemble generation, the conformations are sorted by their final MCOS and subjected to one iteration of clustering utilizing the identical clustering algorithm (see Clustering of Conformers) with an RMSD threshold of 0.1 Å. The

sorting step prior to the clustering step ensures that for each cluster the best-scored conformation is selected.

Output Summary

In addition to any warnings and errors, Conformator prints out a single-line summary for each processed molecule. The summary includes information on the name of the molecule, the number of generated conformers, and stereochemistry. The user may request additional output, such as the minimum pairwise RMSD between a generated conformer and the input conformer, and the minimum pairwise RMSD between any generated conformers. Note that these options may lead to substantially longer runtimes.

Benchmarking Conformer Ensemble Generators

Preparation of the Benchmark Dataset for Computation

The Platinum Diverse Dataset used for benchmarking conformer ensemble generators is a representative subset of the Platinum Dataset.[46] Both datasets were compiled according to the method described in ref 11, with the improvements described in ref 12 and downloaded from ref 47.

Conformer Ensemble Generation

In our previous benchmark studies, standard 3D structures (SDF format) generated from SMILES with NAOMI served as input for conformer ensemble generation for the RDKit DG algorithm and OMEGA. The same structures were used as input for CONFECT[26] in the present work. Conformator was benchmarked with both SMILES and 3D structures as input. Conformer

ensembles were calculated with the parameters described in the Results section and summarized in Table 1.

Table 1. Parameter Sets Applied to Conformer Ensemble Generation.

Algorithm	Mode[a]	Clustering[b]	Force field
Conformator	Best (default)	RMSD	n/MCOS[c]
Conformator	Fast	RMSD	n/MCOS[c]
CONFECT	3[d]	TFD[e]	TrAmber[f]
RDKit DG[g]	n/a	RMSD	UFF[48]
OMEGA[g]	default	RMSD	mmff94s_NoEstat[h]

[a] Parameter sets and search modes supplied by the developers of the respective algorithms.

[b] Distance measure for clustering conformers to form ensembles. Default values were applied.

[c] Macrocycle Optimization Score (MCOS). Only used for macrocycle optimization.

[d] Setting recommended by the developers.[49]

[e] Torsion fingerprint distance.[50]

[f] TrAmber is a hybrid force field partly based on TAFF[51] and used for resolving clashes by small rotations of torsion angles.

[g] Best-performing parameter set in our previous study.[12]

[h] MMFF94 variant that includes all MMFF94s terms except those for Coulomb interactions.

RMSD Calculations, Geometry Checks and Runtime Measurements

The RMSD between pairs of conformers was calculated with NAOMI.[36] NAOMI determines the RMSD based on the best superposition of a pair of conformers, taking into account molecular symmetry via complete automorphism enumeration.

NAOMI was also utilized to determine the deviation of atom angles and bond lengths from known optimal values as well as the divergence of aromatic rings and ring systems (up to 6 bonds per relevant cycle) from planarity.[45] Runtimes of conformer ensemble generation were measured for SD files containing single molecules.

Statistical Analysis

The Mann−Whitney U test was used to test for statistical significance at $\alpha = 0.05$ and $\alpha = 0.01$, with the Holm−Bonferroni method[52] applied to control the familywise error rate. The p-values are reported for pairwise comparisons of the conformer ensemble generators at maximum ensemble sizes 250 and 50 in the Supporting Information (Table S2 and S3).

Hardware Setup

All calculations were performed single-threaded on Linux workstations running openSUSE 42.2 and equipped with Intel Xeon processors (2.2–2.7 GHz) and 126 GB of main memory (Conformator typically uses less than 1 GB of memory).

RESULTS

Benchmarking Conformator

The accuracy and efficiency of Conformator in representing protein-bound ligand conformations was assessed using the same dataset[46] and following the same testing procedure[12] previously applied to the benchmarking of the commercial algorithms ConfGen,[19] ConfGenX,[20] cxcalc,[21] iCon,[22] MOE[23] and OMEGA.[24] In a second, earlier published study[11] we compared the performance of the free conformer ensemble generators Balloon (two different algorithms),[18] the RDKIT DG[13] and ETKDG[14] algorithms, Confab,[15] Frog2[16] and Multiconf-DOCK.[17] This study also followed the identical testing protocol but utilized an earlier version of the Platinum Diverse Dataset.[53] We have previously shown[12] that the marginal differences in the composition of both versions of the Platinum Dataset have no significant impact on any study outcomes. This means that all results presented in the current work can be directly compared to the results reported in either of our previous studies.

The following sections report on key performance figures computed for Conformator and CONFECT, some of which are summarized in Figure 2 and Table 2. In support of the discussions, results obtained as part of our previous study with the best-performing parameter sets (Table 1) for the RDKit DG algorithm (the best-performing free algorithm) and OMEGA (the best-performing commercial algorithm) are recited in the figures and tables of the current work. Results of the Mann−Whitney U test for statistical significance for maximum ensemble sizes of 250 and 50 are provided in the Supporting Information (Table S2 and S3). In the

following sections, four-letter codes refer to PDB entries and three-letter codes in italics refer to PDB ligand identifiers.

Table 2. Comparison of the Performance of Conformer Ensemble Generators on the Platinum Diverse Dataset[a]

Algorithm	Maximum ensemble size 50		Maximum ensemble size 250	
	mean	median	mean	median
	RMSD [Å]			
Conformator Best	**0.68**	**0.58**	**0.57**	**0.47**
Conformator Fast	0.75	0.66	0.64	0.53
CONFECT	0.92	0.74	0.78	0.67
RDKit DG	0.82	0.64	0.64	0.52
OMEGA	**0.67**	**0.51**	**0.57**	**0.46**
	Ensemble size			
Conformator Best	38	42	166	187
Conformator Fast	20	19	70	54
CONFECT	**18**	**15**	**50**	**38**

RDKit DG	42	49	180	229
OMEGA	34	50	118	74
		Runtime [s]		
Conformator Best	**2**	**1**	7	3
Conformator Fast	**2**	**1**	**3**	**1**
CONFECT	**2**	**1**	4	**1**
RDKit DG	4	3	18	14
OMEGA	**2**	2	**3**	2

[a] The best values obtained for RMSD (considering statistical significance), ensemble size and runtime by any of the tested algorithms are marked in bold.

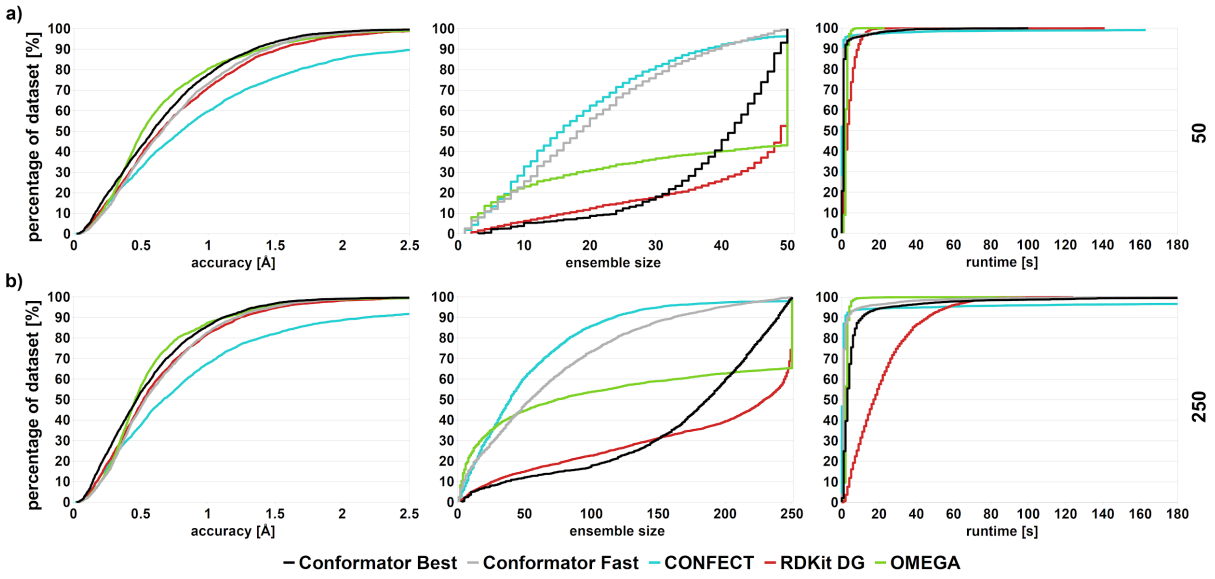

Figure 2. Percentage of protein-bound ligand conformations of the Platinum Diverse Dataset reproduced by the different algorithms within a certain accuracy (left), ensemble size (middle), and runtime per molecule (right) at maximum ensemble sizes (a) 50 and (b) 250 conformers. Steeper curves indicate better performance with respect to all three criteria.

Accuracy and Ensemble Size

This study, like most benchmark studies (including ours[11,12]), defines the accuracy of conformer ensemble generators by the minimum RMSD in Å measured between the experimentally determined protein-bound conformation and any conformer of the computed ensemble. Accuracy is, to some extent, a function of ensemble size.[54] This is because ensembles are generally designed to consist of diverse conformers, which means that chances for one of these conformers to closely resemble the experimentally observed conformation generally increase with the number of generated conformers. Unless stated otherwise, all results presented in the following sections refer to ensembles with a maximum of 250 conformers.

Conformator Best represented the protein-bound ligand conformations with a median RMSD of 0.47 Å at a median ensemble size of 187. Its accuracy was significantly better than that of the RDKit DG algorithm (median RMSD 0.52 Å), even though the RDKit DG algorithm produces larger ensembles (median 229 conformers). The accuracy of Conformator Best was also competitive with that of OMEGA (RMSD 0.47 vs. 0.46 Å; difference not statistically significant), at, however, the expense of a substantially larger median ensemble size (187 vs. 74 conformers). Run at a maximum ensemble size of 250, Conformator Best tends to produce larger ensembles than OMEGA for molecules with four or fewer rotatable bonds (Figure 3a). The opposite trend is observed for more flexible molecules, for which OMEGA generally produces more conformers than Conformator Best. Whereas only 0.8% of all ensembles generated with Conformator Best consisted of the maximum allowed number of conformers (i.e. 250), this figure was 34% for OMEGA. The R^2 for the correlation between the number of rotatable bonds and the size of conformer ensembles was 0.27 for Conformator Best. This weak correlation is a result of the rules for sampling torsion angles for rotatable bonds and of the clustering algorithm, both of which bias the ensembles towards more diversity, meaning that even if for a rotatable bond multiple preferred torsion angles are known, few representative torsion angles are utilized to comply with the maximum allowed ensemble size.

For a maximum ensemble size of 50 conformers, Conformator Best produced smaller ensembles (median 42 conformers) than OMEGA (median 50 conformers) and the RDKit DG algorithm (median 49 conformers). In this setup, no statistically significant difference in the accuracy of Conformator Best (median 0.58 Å) and OMEGA (median 0.51 Å) was observed (Table S3). Again, the accuracy of Conformator Best was significantly higher than that of RDKit DG (median 0.64 Å). At a maximum ensemble size of 50 conformers, Conformator Best generated

larger ensembles than OMEGA for molecules with less than four rotatable bonds but smaller-sized ensembles for molecules with more than four rotatable bonds (Figure 3b). Only 7% of all conformers generated with Conformator Best but 56% of all conformers generated with OMEGA had the maximum ensemble size of 50 conformers (Figure 2a).

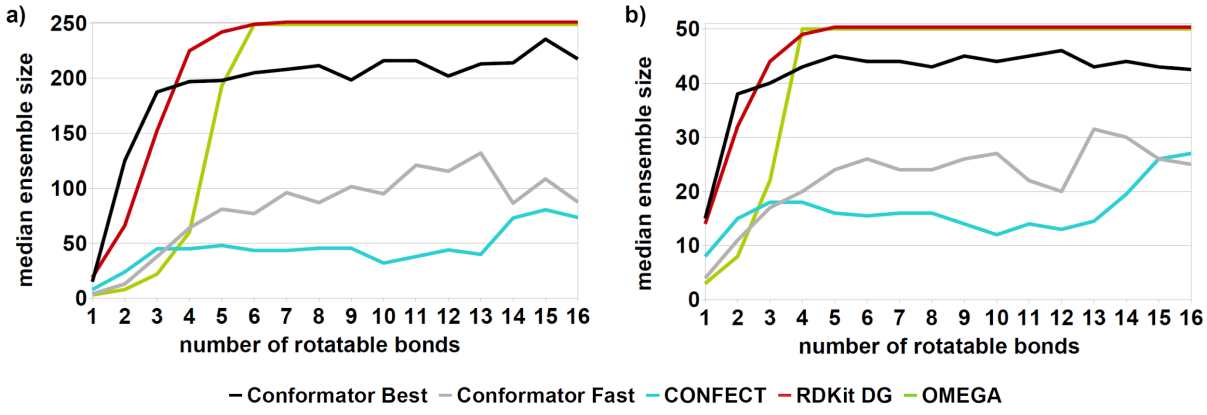

Figure 3. Median ensemble size vs number of rotatable bonds for ensembles of a maximum of a) 250 and b) 50 conformers. Lower curves indicate better performance with respect to ensemble size.

At a maximum ensemble size of 250 conformers, Conformator Fast reproduced the experimentally observed conformations with equal accuracy as the RDKit DG algorithm (median RMSD 0.53 vs. 0.52 Å; difference not statistically significant), despite much smaller ensembles (median 54 vs. 229 conformers). CONFECT produced the smallest ensembles but also was the least accurate among all tested algorithms (median 38 conformers per ensemble; median RMSD 0.67 Å).

In addition, we quantified the accuracy of conformer ensemble generators as the percentage of experimentally observed conformations represented below RMSD thresholds of 0.5, 1.0, 1.5 and 2.0 Å (Table 3). In this assessment, Conformator Best and OMEGA showed comparable

performance, with 53% and 56% of all experimental conformations represented with an RMSD below 0.5 Å, and 97% and 96% represented with an RMSD below 1.5 Å, respectively (maximum ensemble size 250 conformers). The success rates of Conformator Fast were comparable with those of the RDKit DG algorithm. For ensembles of a maximum of 50 conformers at an RMSD threshold below 0.5 Å, the success rate of OMEGA was higher than that of Conformator Best (49% vs. 42%) and any other tested algorithm.

Table 3. Percentage of Structures of the Platinum Diverse Dataset Successfully Reproduced within a Specified RMSD Threshold[a]

Algorithm	**Maximum ensemble size 50**				**Maximum ensemble size 250**			
	RMSD threshold [Å]							
	0.5	**1.0**	**1.5**	**2.0**	**0.5**	**1.0**	**1.5**	**2.0**
Conformator Best	42	78	**94**	**98**	53	86	**97**	**99**
Conformator Fast	37	73	91	**98**	46	83	95	**99**
CONFECT	32	60	76	85	37	62	82	88
RDKit DG	38	71	89	96	47	82	95	98
OMEGA	**49**	**80**	92	97	**56**	**87**	96	**99**

[a] The values of the best-performing algorithms per column are marked in bold.

As a third way of assessing the accuracy of conformer ensemble generators, we quantified the percentage of molecules represented with an RMSD below 0.6 (the maximum positional uncertainty for atoms in the Platinum Dataset)[11] and below 1.0 Å (below which docking poses are commonly deemed sufficiently accurate) with respect to the complexity of their conformational space, represented (in part) by the number of rotatable bonds (Figure 4). At both RMSD thresholds (maximum ensemble size 250 conformers), Conformator Best performed comparably to OMEGA and Conformator Fast comparably to the RDKit DG algorithm. Both Conformator Best and OMEGA, however, performed substantially better than Conformator Fast, the RDKit DG algorithm and CONFECT at both RMSD thresholds. The success rates of representing experimental structures below an RMSD of 0.6 Å were 63 to 96% for Conformator Best, 64 to 95% for OMEGA and 58 to 98% for the RDKit DG algorithm. Likewise, the success rates of representing experimental structures below an RMSD of 1.0 Å were 86 to 99% for Conformator Best, 87 to 98% for OMEGA and 82 to 99% for the RDKit DG algorithm.

Among all tested algorithms, the accuracy of ensembles generated with OMEGA was least dependent on the number of rotatable bonds. At an RMSD cutoff of 0.6 Å, OMEGA successfully represented 88% of all molecules with up to four rotatable bonds and 71% of all molecules with up to eight rotatable bonds. These figures were 89% and 69% for Conformator Best, respectively.

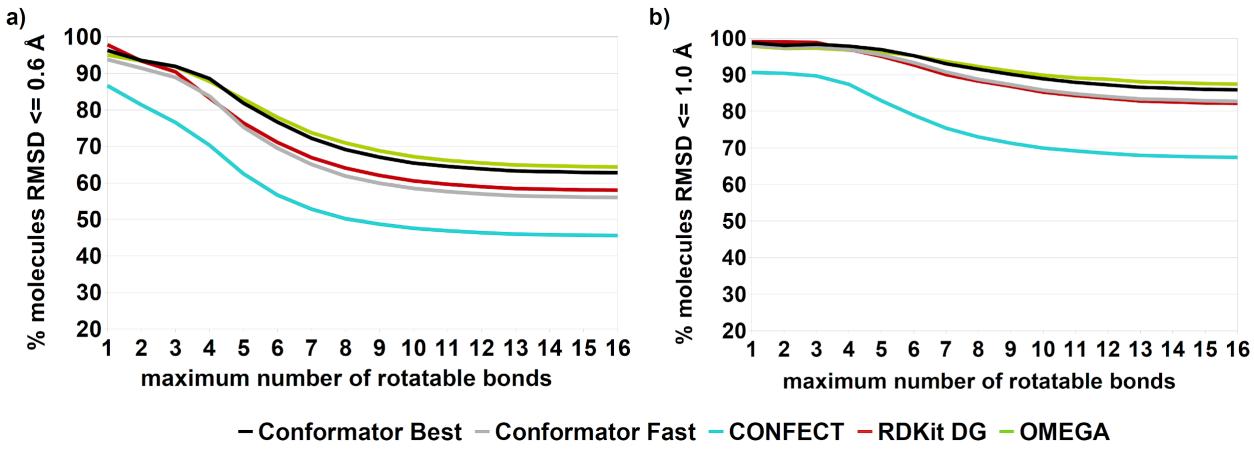

Figure 4. Percentage of molecules of the Platinum Diverse Dataset reproduced by the tested algorithms with a maximum RMSD of (a) 0.6 Å and (b) 1.0 Å as a function of the maximum number of rotatable bonds. The maximum ensemble size was set to 250 conformers.

The diversity of the ensembles generated with Conformator strongly depends on the specific molecular structure in question. In general, the diversity of ensembles increases with the number of rotatable bonds. The R^2 for the correlation between the median pairwise RMSD of all conformers and the number of rotatable bonds was 0.60 (default settings; Figure S10). Two outliers were observed, which are the highly symmetrical ligands *B3P* (Figure S10A) and *5MY* (Figure S10B), for which the symmetry-corrected RMSD was lower than expected based on the number of rotatable bonds. The R^2 for the correlation between the minimum pairwise RMSD and the number of rotatable bonds was 0.50 (default settings; Figure S11). Note that the RMSD also depends on the size of the molecule and that the clustering threshold is not adjusted if the initially generated conformer ensemble is smaller than the maximum allowed ensemble size. Also, during each round of clustering, the radius is incrementally increased by a defined value

(i.e. 0.1 Å for Fast and 0.05 Å for Best), for which reason the maximum allowed ensemble size is often not reached.

For a subset of 987 molecules of the Platinum Diverse Dataset (all of them have a maximum of six rotatable bonds) we were able to generate complete conformer ensembles without clustering and without a set maximum ensemble size (maximum allowed runtime of 72 h per molecule; Table S4). For 92% of all molecules in this subset (84% with default settings) the complete ensembles included a conformer with an RMSD lower than 0.5 Å and for 99% (98% with default settings) a conformer with an RMSD lower than 1 Å. Use of complete conformer ensembles instead of the (default) ensembles of a maximum size of 250 improved the RMSD by 0.5 Å or more in only 14 out of 987 cases. The maximum ensemble size measured was 185,112 conformers; the mean ensemble size 12,024. These results demonstrate the efficiency of the clustering procedure implemented in Conformator.

Success Rates in Processing Molecules

With the exception of CONFECT (success rate 93.4%), all ensemble generators successfully produced ensembles for more than 99% of all tested molecules (Conformator Best and Fast 100.0%; OMEGA 99.6%; RDKit DG algorithm 99.9%). Conformator and OMEGA are designed to handle both 2D and 3D input and produce identical results with either type of information. In the case of SMILES input, Conformator was able to successfully process all molecules with the exception of three molecules with small, bridged rings (i.e. *HUX*, *SAW*, *TSA*). If valid input coordinates are given and the option to generate new 3D coordinates is not set, these three molecules can also be successfully processed by Conformator.

Runtimes

For ensembles consisting of a maximum of 250 conformers, the median runtimes for Conformator Fast and Best were 1 and 3 seconds, respectively (for individual molecules, repeated runtime measurements differed by less than 5%). Hence Conformator was much faster than the RDKit DG algorithm (median 14 seconds) and approximately as fast as OMEGA (median 2 seconds). For ensembles consisting of a maximum of 50 conformers, no substantial differences in the median runtimes were observed: calculations with Conformator Fast and Best had a median runtime of 1 second, with OMEGA 2 seconds and with the RDKit DG algorithm 3 seconds. Note that in previous tests[11] the RDKit ETKDG and DG algorithms produced conformers of comparable quality, with the ETKDG algorithm being 25% faster.

Case Studies on the Reproduction of Experimentally Observed Conformations of Macrocycles

In recent years, macrocycles have emerged as one of the most promising categories of drug candidates for multiple indications.[55–58] Macrocyclic systems are restricted in their rotational and conformational freedom. While this property is actively exploited in the design of highly effective and specific compounds, the interdependency of rotatable bonds and other features such as bridged rings pose significant challenges to conformer ensemble generation. New conformer ensemble generators and extensions, in particular to commercial algorithms, have recently been reported to specifically address these issues.[59–66]

The dedicated algorithm for macrocycle conformer generation, which is part of Conformator, cuts all macrocycles and generates conformers for these open ring structures with Conformator's

standard algorithm. In contrast to DG approaches (which usually start from random coordinates), the conformers used as a starting point for cyclization are already geometrically valid.

We tested the ability of Conformator to represent the experimentally observed, protein-bound conformations of macrocyclic compounds. For this purpose, we extracted from the Sperrylite Dataset all 49 structures of compounds including at least one ring formed by ten or more atoms (29 of these structures are also part of the Platinum Diverse Dataset). Seven of the molecules included in this dataset are represented by more than one experimental structure: latrunculin A (*LAR*; 6 conformers), 6-deoxyerythronolide B (*DEB*; 4 conformers), and geldanamycin (*GDM*), *LAB*, *LY4*, *PP9* and *S1A* (2 conformers). The dataset contains rings of eight different sizes (Figure 5a). It is dominated by 16 molecules (26 conformers) with rings consisting of twelve atoms and seven molecules (nine conformers) with rings consisting of 16 atoms.

Conformator Best successfully processed all 49 macrocyclic structures and obtained a median RMSD of 1.0 Å (Figure 5b). The maximum RMSD measured was 2.3 Å for both structures of geldanamycin (PDB complexes 3C11 and 4XDM; Figure 6). Geldanamycin is a particularly challenging molecule. It consists of 40 heavy atoms and a macrocycle formed by 19 atoms. Its conformation is strongly bent and includes several torsion angles that according to Conformator's torsion angle library are unlikely.

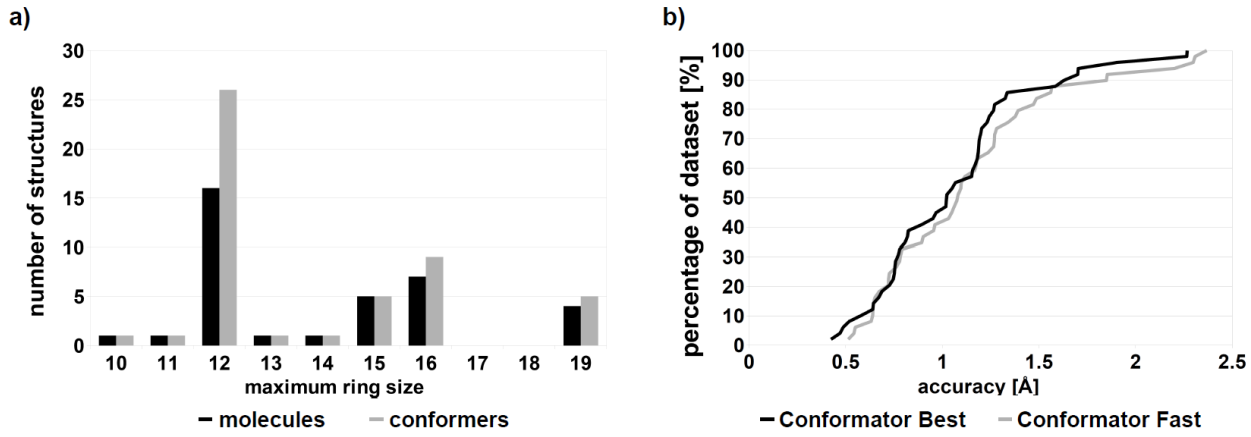

Figure 5. The Sperrylite Dataset contains 49 protein-bound structures of compounds including at least one macrocycle formed by ten or more atoms. (a) Distribution of the maximum ring sizes (number of atoms in a ring) of these macrocycles and their conformations. (b) Cumulative percentage of these structures reproduced by Conformator below a defined maximum RMSD threshold (maximum ensemble size 250 conformers).

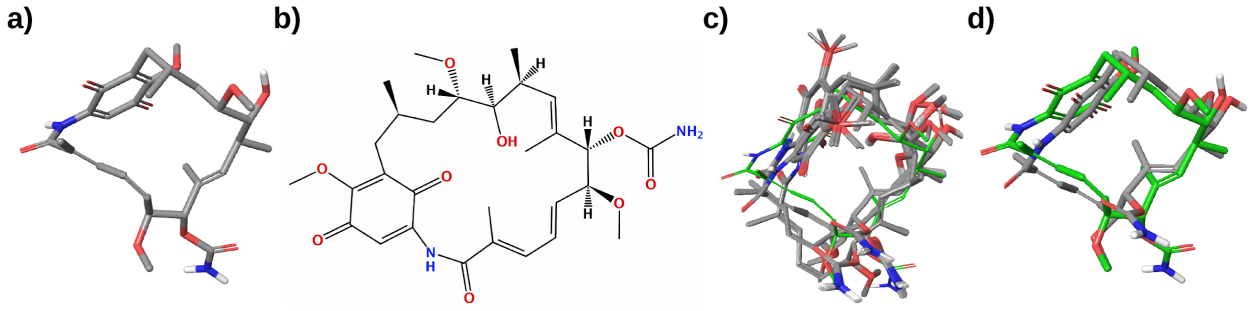

Figure 6. Visualization of structures of geldanamycin. (a) The conformer from the Sperrylite Dataset (*GDM* in 3C11; input for the validation of Conformator), (b) 2D representation of geldanamycin, (c) an ensemble of conformers generated by Conformator Best and superposed with original conformer (green carbon atoms), and (d) the closest conformer generated with Conformator Best and superposed with the original conformer (green carbon atoms).

All further (47) macrocyclic structures were reproduced with RMSD values of less than 2.0 Å. Conformator Best reproduced the experimentally observed conformation of macbecin (*BC2*; 2VWC) and valerjesomycin (*VJ6*; 4JQL), both including macrocycles formed by 19 atoms, with RMSDs of 1.9 Å and 0.8 Å, respectively. For 27 macrocyclic structures (55%), Conformator Best generated at least one conformer with an RMSD not higher than 1.0 Å. At a maximum ensemble size of 250 conformers, the median size of ensembles generated with Conformator Best for the 49 macrocycles was 197 conformers and the average runtime was 104 s (median 88 s) per molecule. Given the limited amount of high-quality structural data on protein-bound macrocycles available to date, no statistically sound conclusions can be drawn on which of the two algorithms performs better.

Comparison of Conformator's Clustering Algorithm with K-Medoids Clustering

In order to assess the performance of the new clustering algorithm implemented in Conformator we produced a version of Conformator Best with the new clustering algorithm replaced by the k-medoids clustering algorithm (the partitioning around medoids method).[67,68] With a maximum of 25 iterations, Conformator in combination with the k-medoids clustering algorithm reached median and mean accuracy values identical to those of the original version of Conformator (median RMSD 0.47 Å; mean RMSD 0.57 Å). However, the median and mean runtimes were substantially longer for the k-medoids clustering algorithm variant (14 s and 272 s per molecule, respectively) as compared to the original version of Conformator (median 3 s; mean 7 s per molecule, respectively). The longest runtime observed for the k-medoids clustering variant was

12.1 h as compared to 512 s for the original version of Conformator. The ensembles generated by the k-medoids clustering variant had a median ensemble size of 250 conformers (mean ensemble size 205) as compared to 187 conformers (mean ensemble size 166) for the original version of Conformator. With k-medoids clustering, 58% of all generated ensembles were of the maximum allowed size (250) whereas this was the case for only 7% of all ensembles generated with the original version of Conformator. The high percentage of large ensembles generated by the k-medoids clustering variant is not surprising since reaching the maximum ensemble size is a defined objective of this clustering algorithm.

CONCLUSION

Conformator is an efficient knowledge-based algorithm for the generation of conformer ensembles of small molecules. One of the key features of Conformator is its new clustering algorithm for the compilation of representative conformer ensembles that exploits the partial presorting of consecutively generated conformers. Conformer ensembles generated with Conformator are independent of input geometries and formats, because the input coordinates are not considered, the new cluster algorithm introduced here is deterministic and the atom order of the molecule is canonized prior to conformer generation. Furthermore, we present a novel algorithm for the generation of conformations for macrocyclic ring systems. The algorithm is robust, widely applicable and makes use of the sophisticated technology for acyclic conformer generation. A novel numeric optimizer working hand in hand with a differentiable scoring function MCOS is responsible for low-energy conformations even in complex, macrocyclic ring systems.

Conformator reaches a level of accuracy and efficiency that is comparable to that of OMEGA. The new algorithm performs particularly well with molecules composed of five or more rotatable bonds, for which it reaches competitive performance while keeping ensemble sizes low. OMEGA, on the other hand, is still ahead in sampling molecules with fewer than five rotatable bonds (which account for more than half of all molecules of the benchmarking dataset), for which it obtains the best accuracy among all tested algorithms even with small ensembles. Preference for either algorithm will depend on the specific application, such as the composition and size of the molecular libraries to be processed. From the outcomes of this study, however, it is clear that in direct comparison with other free algorithms, Conformator obtains very good performance and is the only algorithm for which no significant geometric errors were detected in any of the generated conformations. Conformator successfully processes more than 99% of all input structures, is capable of handling different types of 2D and 3D input and requires only moderate computing resources. In contrast to many other approaches, Conformator does not use any PDB data for deriving geometric parameters like bond lengths, bond angles, torsion angles or ring conformations. Therefore, the performance measured on the basis of the Platinum Dataset gives a realistic picture of the algorithm's practical performance.

Software Availability

Conformator is free for academic use. It is part of the software tool UNICON, a universal converter able to create 2D and 3D conformations on the fly. Conformator and UNICON are standalone command-line tools within the NAOMI ChemBio Suite[36] available from https://software.zbh.uni-hamburg.de.

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI:

Additional figures and tables: Results of the Mann–Whitney U tests and p-values adjusted with the Holm–Bonferroni method for ensembles with a maximum of 250 conformers. Pseudo code for the cluster algorithm. Visualization of Conformator's clustering algorithm by an example. Empirically determined weights for the MCOS and functions of its individual score contributions.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rarey@zbh.uni-hamburg.de. Tel.: +49 40 42838 7351.

Author Contributions

NF, JK and MR conceived the work. NF developed the algorithmic concepts, implemented the software and tested it. FF contributed to the development of the algorithmic concepts, implemented the macrocycle optimization and contributed to the testing of Conformator. KS contributed to the implementation and testing of the algorithm and provided improvements. AM developed the tailored equation for kappa of the Mises function as kernel for curve approximation for the (continuous) torsion angle potential in the calculation of the torsion angle score. MR supervised the method development and JK the validation of Conformator. All

authors contributed to writing of the manuscript and have given approval to the final version of the paper.

ORCID

Nils-Ole Friedrich: 0000-0002-8983-388X

Florian Flachsenberg: 0000-0001-7051-8719

Agnes Meyder: 0000-0001-8519-5780

Kai Sommer: 0000-0003-1866-8247

Johannes Kirchmair: 0000-0003-2667-5877

Matthias Rarey: 0000-0002-9553-6531

FUNDING INFORMATION

JK is supported by the Bergen Research Foundation (BFS) grant no. BFS2017TMT01.

ACKNOWLEDGMENTS

The authors thank Christina de Bruyn Kops from the Center for Bioinformatics (ZBH) of the University of Hamburg for discussion and proofreading of the manuscript. Conformator is built on the NAOMI platform developed by many colleagues at BioSolveIT GmbH and the ZBH. The authors thank the whole team for their help in software development.

REFERENCES

(1) Güner, O.; Clement, O.; Kurogi, Y. Pharmacophore Modeling and Three Dimensional Database Searching for Drug Design Using Catalyst: Recent Advances. *Curr. Med. Chem.* **2004**, *11*, 2991–3005.

(2) Wolber, G.; Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.* **2005**, *45*, 160–169.

(3) Chen, I.-J.; Foloppe, N. Conformational Sampling of Druglike Molecules with MOE and Catalyst: Implications for Pharmacophore Modeling and Virtual Screening. *J. Chem. Inf. Model.* **2008**, *48*, 1773–1791.

(4) Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Friesner, R. A. PHASE: A New Engine for Pharmacophore Perception, 3D QSAR Model Development, and 3D Database Screening: 1. Methodology and Preliminary Results. *J. Comput. Aided Mol. Des.* **2006**, *20*, 647–671.

(5) Friedrich, N.-O.; Simsir, M.; Kirchmair, J. How Diverse Are the Protein-Bound Conformations of Small-Molecule Drugs and Cofactors? *Front. Chem.* **2018**, *6*, 68.

(6) Perola, E.; Charifson, P. S. Conformational Analysis of Drug-like Molecules Bound to Proteins: An Extensive Study of Ligand Reorganization upon Binding. *J. Med. Chem.* **2004**, *47*, 2499–2510.

(7) Iuzzolino, L.; Reilly, A. M.; McCabe, P.; Price, S. L. Use of Crystal Structure Informatics for Defining the Conformational Space Needed for Predicting Crystal Structures of Pharmaceutical Molecules. *J. Chem. Theory Comput.* **2017**, *13*, 5163–5171.

(8) Govind, N.; Petersen, M.; Fitzgerald, G.; King-Smith, D.; Andzelm, J. A Generalized Synchronous Transit Method for Transition State Location. *Comput. Mater. Sci.* **2003**, *28*, 250–258.

(9) Jacobson, L. D.; Bochevarov, A. D.; Watson, M. A.; Hughes, T. F.; Rinaldo, D.; Ehrlich, S.; Steinbrecher, T. B.; Vaitheeswaran, S.; Philipp, D. M.; Halls, M. D.; et al. Automated Transition State Search and Its Application to Diverse Types of Organic Reactions. *J. Chem. Theory Comput.* **2017**, *13*, 5780–5797.

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

(11) Friedrich, N.-O.; Meyder, A.; de Bruyn Kops, C.; Sommer, K.; Flachsenberg, F.; Rarey, M.; Kirchmair, J. High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, *57*, 529–539.

(12) Friedrich, N.-O.; de Bruyn Kops, C.; Flachsenberg, F.; Sommer, K.; Rarey, M.; Kirchmair, J. Benchmarking Commercial Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, *57*, 2719–2728.

(13) Tosco, P.; Stiefl, N.; Landrum, G. Bringing the MMFF Force Field to the RDKit: Implementation and Validation. *J. Cheminform.* **2014**, *6*.

(14) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562–2574.

(15) O'Boyle, N. M.; Vandermeersch, T.; Flynn, C. J.; Maguire, A. R.; Hutchison, G. R. Confab - Systematic Generation of Diverse Low-Energy Conformers. *J. Cheminform.* **2011**, *3*, 8.

(16) Miteva, M. A.; Guyon, F.; Tufféry, P. Frog2: Efficient 3D Conformation Ensemble

Generator for Small Compounds. *Nucleic Acids Res.* **2010**, *38* (Web Server issue), W622–W627.

(17) Sauton, N.; Lagorce, D.; Villoutreix, B. O.; Miteva, M. A. MS-DOCK: Accurate Multiple Conformation Generator and Rigid Docking Protocol for Multi-Step Virtual Ligand Screening. *BMC Bioinformatics* **2008**, *9*, 184.

(18) Vainio, M. J.; Johnson, M. S. Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Model.* **2007**, *47*, 2462–2474.

(19) Watts, K. S.; Dalal, P.; Murphy, R. B.; Sherman, W.; Friesner, R. A.; Shelley, J. C. ConfGen: A Conformational Search Method for Efficient Generation of Bioactive Conformers. *J. Chem. Inf. Model.* **2010**, *50*, 534–546.

(20) ConfGenX, Version 2016−2, Part of the Schrödinger Small-Molecule Drug Discovery Suite; Schrödinger: New York, NY, 2016.

(21) Cxcalc, Version 15.8.31.0, Part of the Discovery Toolkit; ChemAxon: Budapest, Hungary, 2015.

(22) Poli, G.; Seidel, T.; Langer, T. Conformational Sampling of Small Molecules With iCon: Performance Assessment in Comparison With OMEGA. *Front. Chem.* **2018**, *6*, 229.

(23) Molecular Operating Environment (MOE), Version 2016.08; Chemical Computing Group: Montreal, QC, 2017.

(24) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572–584.

(25) Meyder, A.; Nittinger, E.; Lange, G.; Klein, R.; Rarey, M. Estimating Electron Density

Support for Individual Atoms and Molecular Fragments in X-Ray Structures. *J. Chem. Inf. Model.* **2017**, *57*, 2437–2447.

(26) Schärfer, C.; Schulz-Gasch, T.; Hert, J.; Heinzerling, L.; Schulz, B.; Inhester, T.; Stahl, M.; Rarey, M. CONFECT: Conformations from an Expert Collection of Torsion Patterns. *ChemMedChem* **2013**, *8*, 1690–1700.

(27) Guba, W.; Meyder, A.; Rarey, M.; Hert, J. Torsion Library Reloaded: A New Version of Expert-Derived SMARTS Rules for Assessing Conformations of Small Molecules. *J. Chem. Inf. Model.* **2016**, *56*, 1–5.

(28) O'Boyle, N. M. Towards a Universal SMILES Representation - A Standard Method to Generate Canonical SMILES Based on the InChI. *J. Cheminform.* **2012**, *4*, 22.

(29) Schärfer, C.; Schulz-Gasch, T.; Ehrlich, H.-C.; Guba, W.; Rarey, M.; Stahl, M. Torsion Angle Preferences in Druglike Chemical Space: A Comprehensive Guide. *J. Med. Chem.* **2013**, *56*, 2016–2028.

(30) Sutton, L. E. *Tables of Interatomic Distances and Configuration in Molecules and Ions : Supplement 1956-59*; 1965.

(31) Gillespie, R. J. The Electron-Pair Repulsion Model for Molecular Geometry. *J. Chem. Educ.* **1970**, *47*, 18.

(32) Ehrlich, H.-C.; Henzler, A. M.; Rarey, M. Searching for Recursively Defined Generic Chemical Patterns in Nonenumerated Fragment Spaces. *J. Chem. Inf. Model.* **2013**, *53*, 1676–1688.

(33) Floyd, R. W. Algorithm 97: Shortest Path. *Commun. ACM* **1962**, *5*, 345.

(34) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261*, 470–489.

(35) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Model.* **1989**, *29*, 97–101.

(36) Urbaczek, S.; Kolodzik, A.; Fischer, J. R.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI: On the Almost Trivial Task of Reading Molecules from Different File Formats. *J. Chem. Inf. Model.* **2011**, *51*, 3199–3207.

(37) Kolodzik, A.; Urbaczek, S.; Rarey, M. Unique Ring Families: A Chemically Meaningful Description of Molecular Ring Topologies. *J. Chem. Inf. Model.* **2012**, *52*, 2013–2021.

(38) Flachsenberg, F.; Andresen, N.; Rarey, M. RingDecomposerLib: An Open-Source Implementation of Unique Ring Families and Other Cycle Bases. *J. Chem. Inf. Model.* **2017**, *57*, 122–126.

(39) Hudson, B. D.; Hyde, R. M.; Rahr, E.; Wood, J.; Osman, J. Parameter Based Methods for Compound Selection from Chemical Databases. *Quant. Struct.-Act. Relat.* **1996**, *15*, 285–289.

(40) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747–750.

(41) Zhu, C.; Byrd, R. H.; Lu, P.; Nocedal, J. Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization. *ACM Trans. Math. Softw.* **1997**, *23*, 550–560.

(42) Morales, J. L.; Nocedal, J. Remark on "algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound Constrained Optimization." *ACM Trans. Math. Softw.* **2011**, *38*, 1–4.

(43) Nittinger, E.; Flachsenberg, F.; Bietz, S.; Lange, G.; Klein, R.; Rarey, M. Placement of Water Molecules in Protein Structures: From Large-Scale Evaluations to Single-Case

Examples. *J. Chem. Inf. Model.* **2018**, *58*, 1625–1637.

(44) McCabe, P.; Korb, O.; Cole, J. Kernel Density Estimation Applied to Bond Length, Bond Angle, and Torsion Angle Distributions. *J. Chem. Inf. Model.* **2014**, *54*, 1284–1288.

(45) Meyder, A.; Kampen, S.; Sieg, J.; Fährrolfes, R.; Friedrich, N.-O.; Flachsenberg, F.; Rarey, M. StructureProfiler: An All-in-One Tool for 3D Protein Structure Profiling. *Bioinformatics* **2018**. doi: 10.1093/bioinformatics/bty692

(46) Platinum Diverse Dataset (version 2017_01). http://www.zbh.uni-Hamburg.de/platinum_dataset. Accessed 27 Jun 2017.

(47) The Platinum Datasets http://www.zbh.uni-hamburg.de/platinum_dataset (accessed Jun 27, 2017).

(48) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.

(49) C. Schärfer, Personal Communication, May, 2014.

(50) Schulz-Gasch, T.; Schärfer, C.; Guba, W.; Rarey, M. TFD: Torsion Fingerprints as a New Measure to Compare Small Molecule Conformations. *J. Chem. Inf. Model.* **2012**, *52*, 1499–1512.

(51) Clark, M.; Cramer, R. D.; Van Opdenbosch, N. Validation of the General Purpose Tripos 5.2 Force Field. *J. Comput. Chem.* **1989**, *10*, 982–1012.

(52) Holm, S. A. Simple Sequentially Rejective Multiple Test Procedure Scand. J. Stat. **1979**, 6, 65–70.

(53) Platinum Diverse Dataset (version 2016_01). http://www.zbh.uni-Hamburg.de/platinum_dataset. Accessed 27 Jun 2017.

(54) Griewel, A.; Kayser, O.; Schlosser, J.; Rarey, M. Conformational Sampling for Large-Scale Virtual Screening: Accuracy versus Ensemble Size. *J. Chem. Inf. Model.* **2009**, *49*, 2303–2311.

(55) Marsault, E.; Peterson, M. L. Macrocycles Are Great Cycles: Applications, Opportunities, and Challenges of Synthetic Macrocycles in Drug Discovery. *J. Med. Chem.* **2011**, *54*, 1961–2004.

(56) Krahn, D.; Ottmann, C.; Kaiser, M. Macrocyclic Proteasome Inhibitors. *CMC* **2011**, *18*, 5052–5060.

(57) Churpek, J. E.; Pro, B.; van Besien, K.; Kline, J.; Conner, K.; Wade, J. L., 3rd; Hagemeister, F.; Karrison, T.; Smith, S. M. A Phase 2 Study of Epothilone B Analog BMS-247550 (NSC 710428) in Patients with Relapsed Aggressive Non-Hodgkin Lymphomas. *Cancer* **2013**, *119*, 1683–1689.

(58) Dougherty, P. G.; Qian, Z.; Pei, D. Macrocycles as Protein-Protein Interaction Inhibitors. *Biochem. J* **2017**, *474*, 1109–1125.

(59) Bonnet, P.; Agrafiotis, D. K.; Zhu, F.; Martin, E. Conformational Analysis of Macrocycles: Finding What Common Search Methods Miss. *J. Chem. Inf. Model.* **2009**, *49*, 2242–2259.

(60) Chen, I.-J.; Foloppe, N. Tackling the Conformational Sampling of Larger Flexible Compounds and Macrocycles in Pharmacology and Drug Discovery. *Bioorg. Med. Chem.* **2013**, *21*, 7898–7920.

(61) Watts, K. S.; Dalal, P.; Tebben, A. J.; Cheney, D. L.; Shelley, J. C. Macrocycle Conformational Sampling with MacroModel. *J. Chem. Inf. Model.* **2014**, *54*, 2680–2696.

(62) Coutsias, E. A.; Lexa, K. W.; Wester, M. J.; Pollock, S. N.; Jacobson, M. P. Exhaustive Conformational Sampling of Complex Fused Ring Macrocycles Using Inverse Kinematics.

J. Chem. Theory Comput. **2016**, *12*, 4674–4687.

(63) Cleves, A. E.; Jain, A. N. ForceGen 3D Structure and Conformer Generation: From Small Lead-like Molecules to Macrocyclic Drugs. *J. Comput. Aided Mol. Des.* **2017**, *31*, 419–439.

(64) 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.

(65) Kamenik, A. S.; Lessel, U.; Fuchs, J. E.; Fox, T.; Liedl, K. R. Peptidic Macrocycles - Conformational Sampling and Thermodynamic Characterization. *J. Chem. Inf. Model.* **2018**, *58*, 982–992.

(66) OMEGA v3.0.0 released https://www.eyesopen.com/news/omega-v3.0.0-released (accessed Sep 7, 2018).

(67) Kaufman, L.; Rousseeuw, P. Clustering by Means of Medoids. In *Statistical Data Analysis Based on the L1–Norm and Related Methods*, 1987, 405–416.

(68) Jin X., Han J. K-Medoids Clustering. 2011. In *Encyclopedia of Machine Learning*; Sammut, C., Webb, G. I., Eds.; Springer US: Boston, MA, 2010, 22.

For Table of Contents Use Only

Conformator: A Novel Method for the Generation of Conformer Ensembles

Nils-Ole Friedrich, Florian Flachsenberg, Agnes Meyder, Kai Sommer, Johannes Kirchmair, Matthias Rarey