Supporting Information

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.

TABLE OF CONTENTS

Table S1. Empirically Determined Weights for the MCOS.

Contribution	Weight
$w_{overlay}$	1.0
w_{bond}	1.0
w_{angle}	1.0
w_{limit}	500.0
$w_{torsion}$	0.1
$w_{torsion,conjugated}$	1.0
w_{clash}	1.0

Table S2. Mann-Whitney U Test Results of RMSD Values from Conformer Ensemble Generation for Platinum Diverse Dataset with a Maximum of 250 Conformers.[a]

Conformer ensemble generator		RDKit DG (UFF and clustering)	OMEGA (default)	Conformator Fast	Conformator Best
OMEGA (default)	**p**	< 0.001	-		
	Z	-5.05	-		
	U	3747852	-		
Conformator Fast	**p**	**0.03**	< 0.001	-	
	Z	-1.85	-7.25	-	
	U	3959194	3614564	-	
Conformator Best	**p**	< 0.001	**0.07**	< 0.001	-
	Z	-6.03	-1.48	-8.09	-
	U	3698854	3973330	3574312	-
CONFECT	**p**	< 0.001	< 0.001	< 0.001	< 0.001
	Z	-7.15	-11.62	-5.69	-12.47
	U	3387977	3116302	3478176	3076036

[a]The Mann-Whitney U test was used to test for statistical significance at $\alpha = 0.05$ and $\alpha = 0.01$, adjusted with the Holm−Bonferroni method to control the familywise error rate. Differences between Conformator Best and OMEGA, as well as Conformator Fast and the RDKit DG algorithm are not statistically significant (bold p values).

Table S3. Mann-Whitney U Test Results of RMSD Values from Conformer Ensemble Generation for Platinum Diverse Dataset with a Maximum of 50 Conformers.[a]

Conformer ensemble generator		RDKit DG (UFF and clustering)	OMEGA (default)	Conformator Fast	Conformator Best
OMEGA (default)	p	< 0.001	-		
	Z	-10.12	-		
	U	3399855	-		
Conformator Fast	p	**0.03**	< 0.001	-	
	Z	-1.94	-8.50	-	
	U	3917040	3537062	-	
Conformator Best	p	< 0.001	**0.02**	< 0.001	-
	Z	-7.89	-2.06	-6.18	-
	U	3549130	3937297	3693374	-
CONFECT	p	< 0.001	< 0.001	< 0.001	< 0.001
	Z	-3.78	-12.25	-5.47	-10.74
	U	3550478	3075448	3487404	3174824

[a]The Mann−Whitney U test was used to test for statistical significance at $\alpha = 0.05$ and $\alpha = 0.01$, adjusted with the Holm−Bonferroni method to control the familywise error rate. Differences between Conformator Best and OMEGA, as well as Conformator Fast and the RDKit DG algorithm are not statistically significant (bold p values).

Table S4. Percentage of Structures Successfully Reproduced within a Specified RMSD Threshold by Complete Sets of Conformers[a]

Setting	RMSD threshold [Å]		
	0.5	**1.0**	**1.5**
complete[b]	92	99	100
default	84	98	99

[a] On a subset of the Platinum Diverse Dataset of 987 molecules (with a maximum of 6 rotatable bonds)

[b] Conformator Best, no clustering, no maximum ensemble size, maximum runtime of 72 h per molecule

Algorithm S1 RMSD-Clustering of Conformers[a]

```
Input: List of conformers (Y)              //candidate conformers, partially presorted
Input: quality_level                       //1 = Fast, 2 = Best (default 2)
Input: max_ensemble_size                   //maximum ensemble size (default 250)
Output: List of cluster centers (Z)        //output conformer ensemble

rmsd_threshold ← 0.1                       //RMSD starting threshold in Å for Best (default)
rmsd_increase ← 0.05                       //RMSD threshold enlargement per round
if (quality_level == 1)
        rmsd_threshold ← 0.5               //RMSD starting threshold in Å for Fast
        rmsd_increase ← 0.5
while (Z.size() > max_ensemble_size)       //starting new clustering
        Z.clear                            //empty list of cluster centers
        candidate_conformer ← Y.begin()    //first conformation is the first cluster center
        while (candidate_conformer != Y.end())          //starting new clustering round
                for (cluster_center = Z.end() to cluster_center Z.begin()      //in reverse
                        rmsd = calculate_rmsd(candidate_conformer, cluster_center)
                        if (rmsd < rmsd_threshold)
                                tooclose ← true
                                break              //no further comparisons
                end for
                if (tooclose)
                        Y.erase(candidate_conformer)
                        //remove candidate conformer permanently
                else
                        Z.push_back(candidate_conformer)
                        //add candidate as cluster center
                        candidate_conformer = Y.next()
                if (Z.size() > max_ensemble_size)     //too many cluster centers
                        break                         //start new round (inner while loop)
        end while
        rmsd_threshold ← rmsd_threshold + rmsd_increase
end while
return Z                   //output list of cluster centers as the conformer ensemble
```

[a]Note that the representation with two separate lists (Y and Z) was chosen for didactic reasons. The algorithm should be implemented with a single array of conformers running in place with indices marking the current end of the cluster center set and the beginning of the unprocessed conformer list.

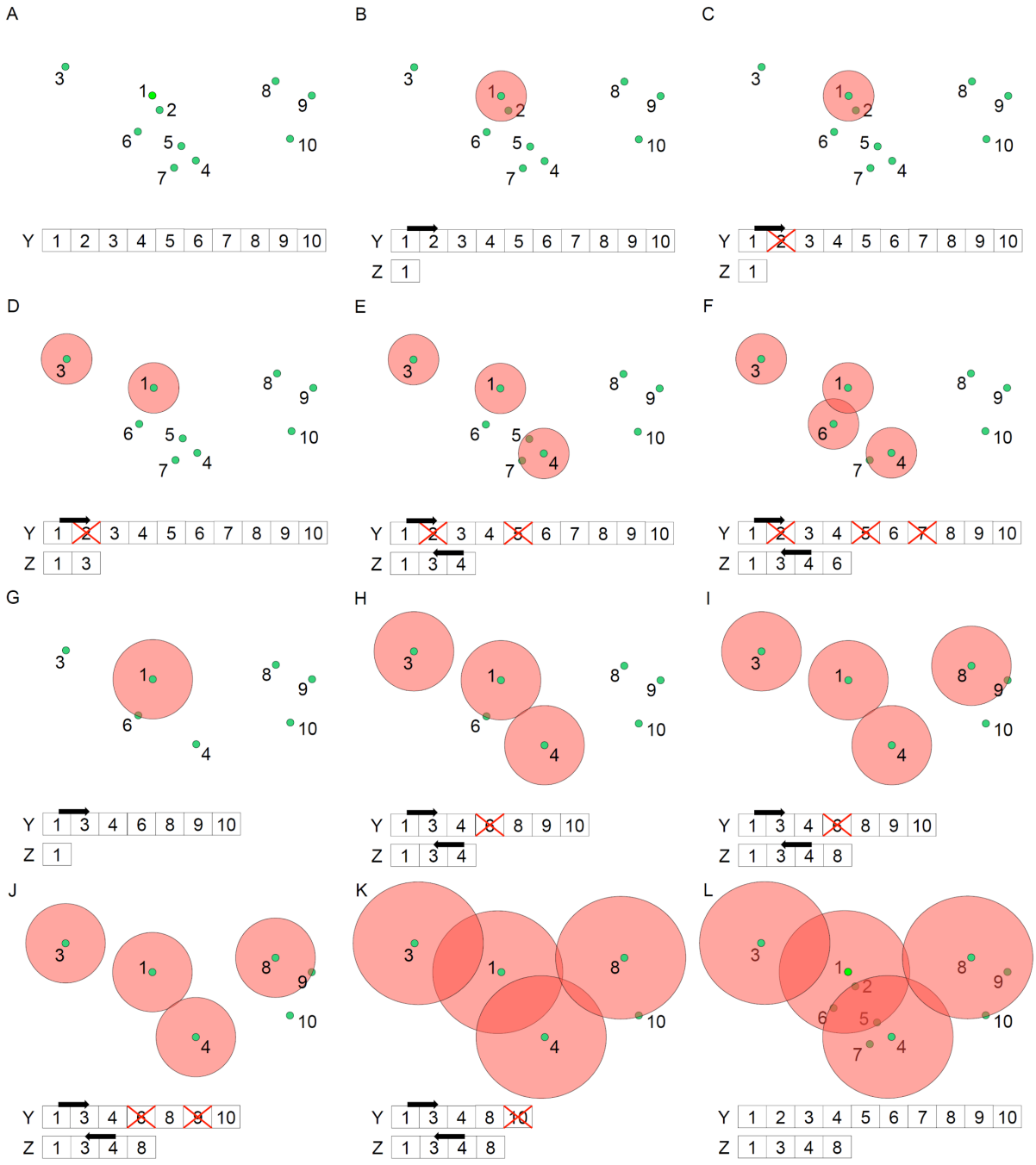


Figure S1. Visualization of Conformator's clustering algorithm by the example of the generation of an ensemble of four representative conformers starting from a set of ten candidate conformers. The green dots represent the candidate conformations. Their distances in 2D space is indicative of their RMSD. The increasing RMSD thresholds are illustrated by the red spheres. Arrows indicate the directions in which the lists of all remaining candidate conformers (Y, top list) and cluster centroids (Z, bottom list) are accessed. Crossed-out numbers indicate conformers that have been removed by the clustering algorithm from the list of candidate conformers. (a) Clustering starts from a list of ten candidate conformations generated with Conformator.

Importantly, these lists are partially presorted, meaning that sequentially generated conformers are likely similar. (b) The first conformer (usually based on very likely torsion angles; see Conformer Generation Algorithm) in the list of candidate conformers is always the first cluster center. (c) The candidate conformers are compared to any of the existing cluster centers. If they are within the RMSD radius (like it is the case for conformer 2) they are removed from the list of candidate conformers. (d) Outliers such as conformer 3 become cluster centers. This behavior is desired as it assures that a sufficiently large part of the relevant conformational space is covered. (e) To take advantage of the fact that conformers generated sequentially with Conformator are likely similar, the list of cluster centers is reversed when comparing candidate conformers to existing cluster centers. While this has no effect on conformer 4 (it is compared against all cluster centers, is dissimilar to all of them and thus becomes a new cluster center), most candidate conformers can be excluded from extensive pairwise comparison, such as conformer 5, which is only compared to conformer 4 before it is removed. (f) Conformer 6 is defined as a new cluster center and conformer 7 is removed from the list of candidate conformers because it is too similar to conformer 4. Conformer 8 is sufficiently distant to any of the existing cluster centers and hence would become a new cluster center. However, this would exceed the maximum ensemble size (which is 4 in this example), for which reason (g) the clustering is repeated with larger RMSD threshold, an empty list of cluster centers and the list of remaining candidates (in other words, previously removed conformers are not considered again). Over several iterations this process determines an appropriate RMSD threshold for each individual molecule. The final threshold depends on the maximum ensemble size and quality level, as well as the size and flexibility of the molecule. (h) Conformers 1, 3 and 4 are again defined as cluster centers but the former cluster center “conformer 6” is removed since it is closer to conformer 1 than the increased distance value allows. (i) Conformer 8 is another cluster center and conformer 9 is removed from the list of candidate conformers. Conformer 10 would become the next cluster center but this would exceed the maximum ensemble size. (j) Once more the clustering process is restarted with a larger RMSD threshold, an empty list of cluster centers and the list of remaining candidate conformers. Conformers 1, 3, 4 and 8 are still far enough apart to become cluster centers but conformer 10 is now too similar to conformer 8 and removed. (k) Now all conformers have been successfully assigned to a cluster center and the ensemble size is equal to (or below) the maximum ensemble size. The final list of cluster centers is then reported as the conformer ensemble.

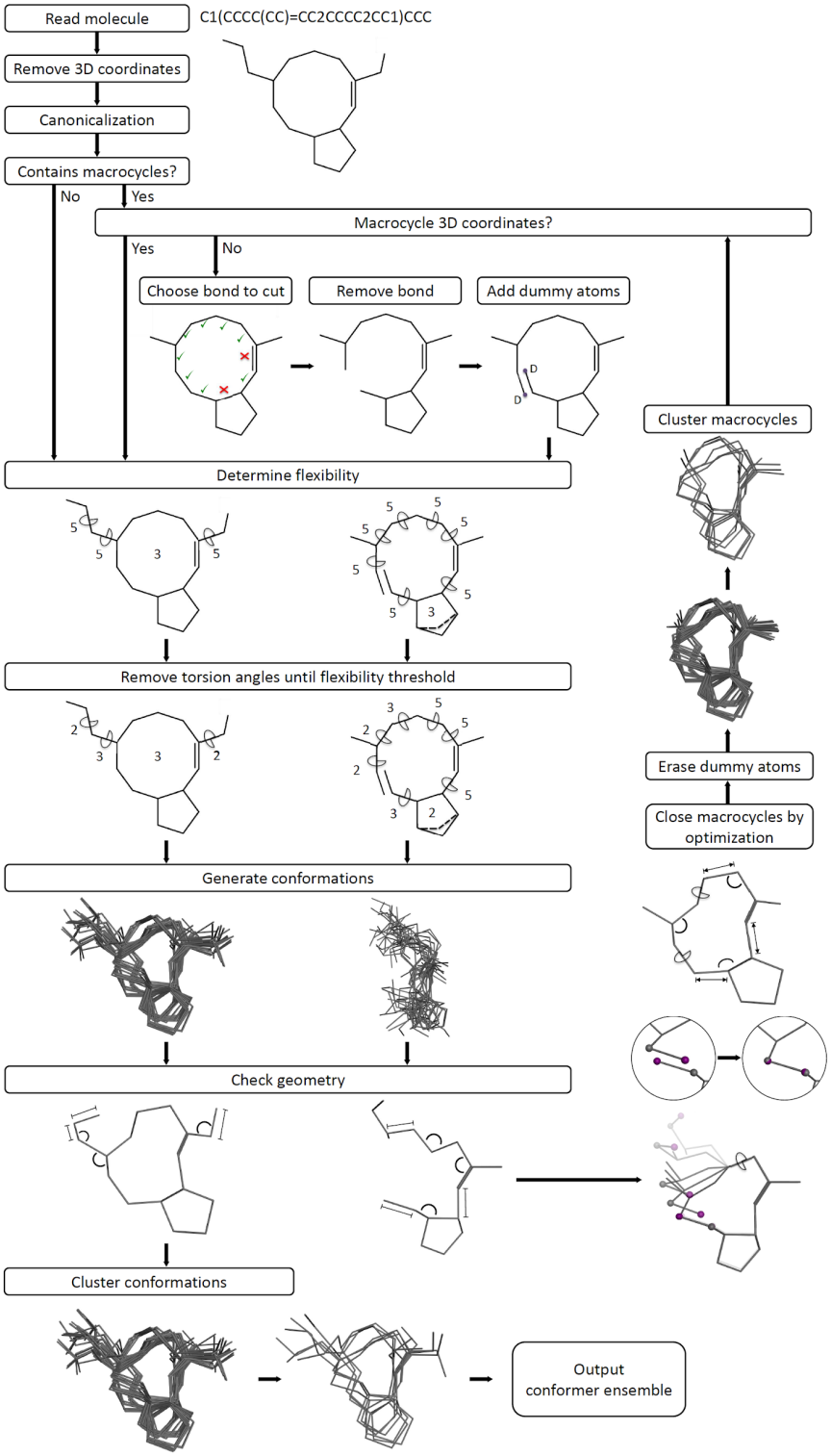
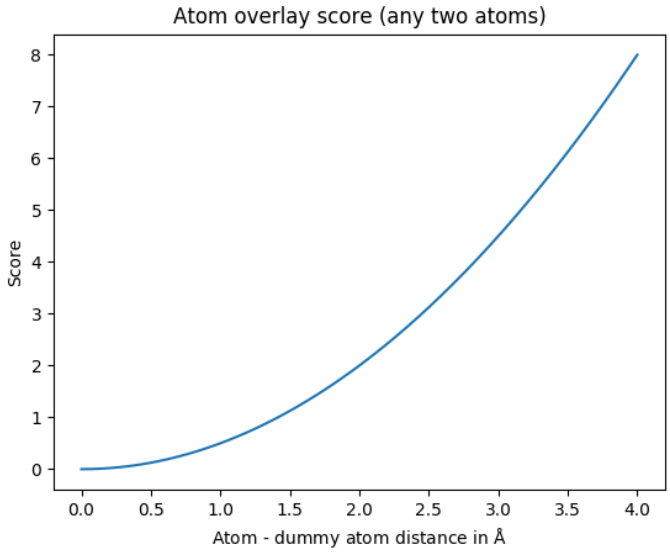
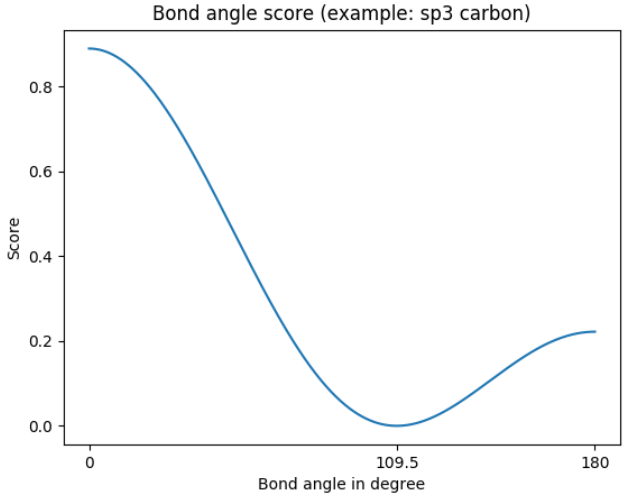


Figure S2. Schematic representation of Conformator's macrocycle conformer generation algorithm (for a detailed description see "Conformer Generation for Macrocycles" in the main text).



$$s_{overlay}(d) = \frac{1}{2}d^2$$

Figure S3. MCOS function for the overlay score for the distance d between the dummy atoms and the atoms in the original macrocycle they replaced. Ideally, this distance should be close to 0. It ensures that the bond angle and bond length across the cut bond will be restored during local optimization and also supports the preservation of local stereochemistry.



$$s_{angle}(\theta, \theta_0) = \frac{1}{2}(\cos\theta - \cos\theta_0)^2$$

Figure S4. MCOS function for the bond angle score. It uses a harmonic potential that is calculated on the cosine of the bond angle θ, to account for deviations from the ideal values θ_0. The bond angle score is calculated only for bond angles directly altered during optimization (i.e. angles that are optimization parameters) and the angles involving the cut bonds.

$$s_{limit}(\theta) = \begin{cases} \frac{1}{2}(\cos\theta - \cos 30°)^2 & \cos\theta > \cos 30° \\ \frac{1}{2}(\cos\theta - \cos 150°)^2 & \cos\theta < \cos 150° \\ 0 & \text{otherwise} \end{cases}$$

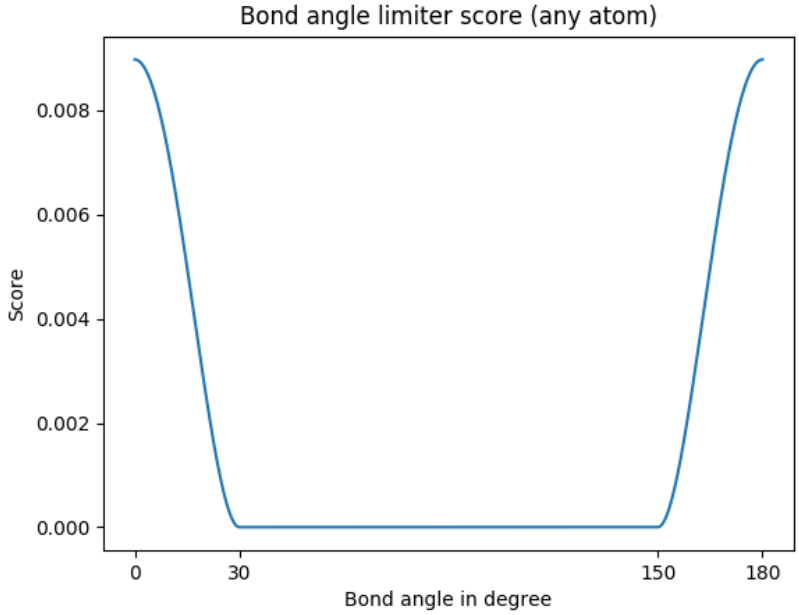
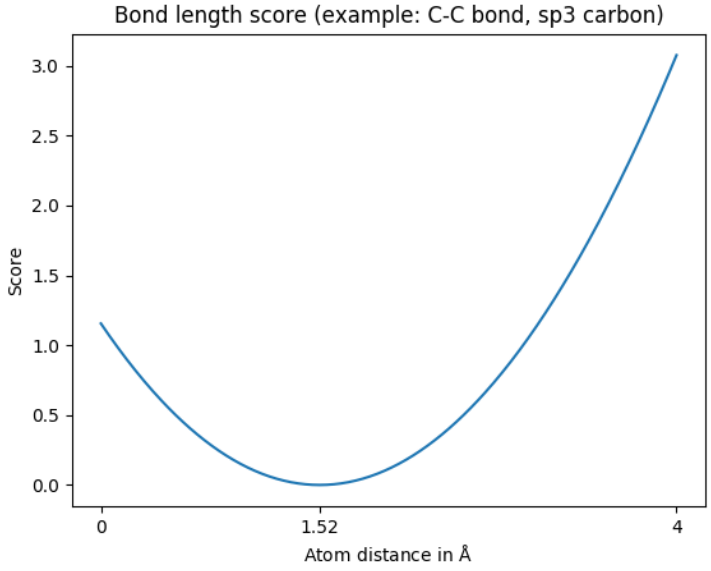


Figure S5. MCOS penalty function for limiting bond angles θ to guide the optimization of bond angles in macrocycles away from 0 and 180 degrees (if the atom does not have linear VSEPR geometry). It leads to a preference of bond angles between 30 and 150 degrees.



$$s_{bond}(d, d_0) = \frac{1}{2}(d - d_0)^2$$

Figure S6. The MCOS bond length term uses a harmonic potential to account for deviations of the bond length d from ideal values d_0. Only the bond lengths of the cut bonds are scored.

$$s_{distance\_factor}(d) = \begin{cases} 1 - g(d) & d \leq 0.5\text{Å} \\ 1 & \text{otherwise} \end{cases}$$

$$g(x) = \begin{cases} a_0 + a_1 \cdot x + a_2 \cdot x^2 + a_3 \cdot x^3 & 0 \leq x < 0.25 \\ b_0 + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3 & 0.25 \leq x < 0.375 \\ c_0 + c_1 \cdot x + c_2 \cdot x^2 + c_3 \cdot x^3 + c_4 \cdot x^4 & 0.375 \leq x < 0.5 \\ d_0 & 0.5 \leq x < \infty \end{cases}$$

	0	1	2	3	4
a	1	0	0	−6.5641	0
b	2.23077	−14.7692	59.0769	−85.3333	0
c	−100.923	960	−3337.85	5060.92	−2835.69
d	0	0	0	0	0

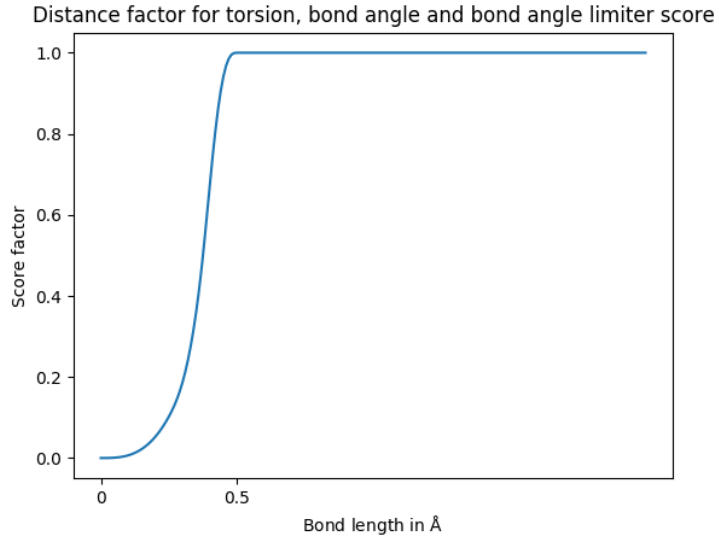


Figure S7. The MCOS distance factor by which the torsion angle potential, the bond angle potential and the bond angle limiter score are multiplied to reduce the respective score to 0 in cases where any bond length is close to 0 Å. This is necessary to ensure the continuity of the score contributions that depend on torsion angles or bond angles. It is a piecewise polynomial approximation to a plateau function that is twice continuously differentiable. The function $g(x)$ was modeled by fixing function and derivative values at defined points and solving the system of equations for the coefficients of the polynomials (the coefficients are shown in the table).

$$s_{angle\_factor}(\theta) = \begin{cases} 1 - g(1 - \cos\theta) & \theta \leq 20° \\ 1 & 20° < \theta < 160° \\ 1 - g(\cos\theta + 1) & \theta \geq 160° \end{cases}$$

$$g(x) = \begin{cases} a_0 + a_1 \cdot x + a_2 \cdot x^2 + a_3 \cdot x^3 & 0 \leq x < 0.0151922 \\ b_0 + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3 & 0.0151922 \leq x < 0.0377498 \\ c_0 + c_1 \cdot x + c_2 \cdot x^2 + c_3 \cdot x^3 + c_4 \cdot x^4 & 0.0377498 \leq x < 0.0603074 \\ d_0 & 0.0603074 \leq x < \infty \end{cases}$$

	0	1	2	3	4
a	1	0	0	−10046.7	0
b	0.987639	2.44088	−160.666	−6521.48	0
c	−11.3122	1115.49	−36828.1	507524	$-2.52014 \cdot 10^6$
d	0	0	0	0	0

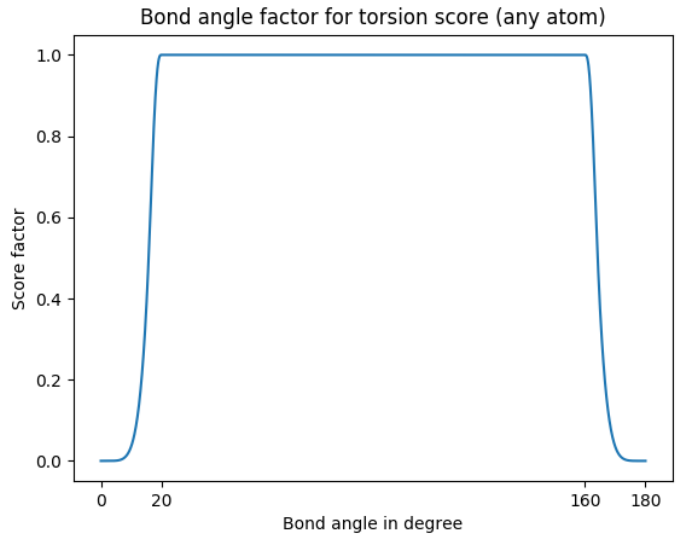


Figure S8. The MCOS bond angle factor by which the torsion angle potential is multiplied to reduce the torsion angle score to 0 in cases where any bond angle θ along that torsion bond is either close to 0 or 180 degrees. This is necessary because the torsion angle, as a function of the four atom coordinates, has a discontinuity when three consecutive atoms are collinear. It is a piecewise polynomial approximation to a plateau function that is twice continuously differentiable. The function $g(x)$ was modeled by fixing function and derivative values at defined points and solving the system of equations for the coefficients of the polynomials (the coefficients are shown in the table).

$$s_{clash}(d, d_{vdw}) = \begin{cases} \frac{1}{2}(d - 0.7 \cdot d_{vdw})^2 & d < 0.7 \cdot d_{vdw} \\ 0 & \text{otherwise} \end{cases}$$

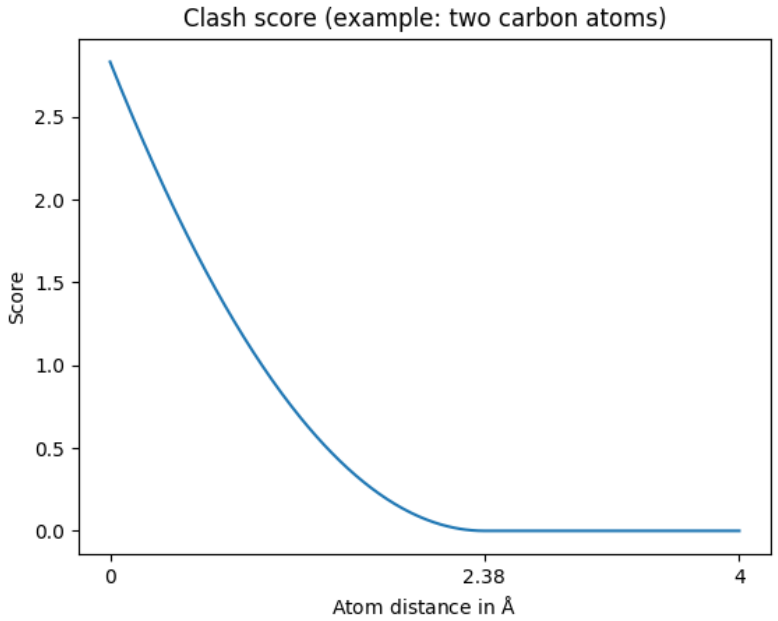


Figure S9. The MCOS clash score prevents intramolecular clashes. It is a quadratic function depending of the atomic distance d and the sum of the van der Waals radii of the atoms d_{vdw} and penalizes van der Waals overlaps between 1-4-connected (or further away) heavy atoms that exceed the threshold level of 30%.

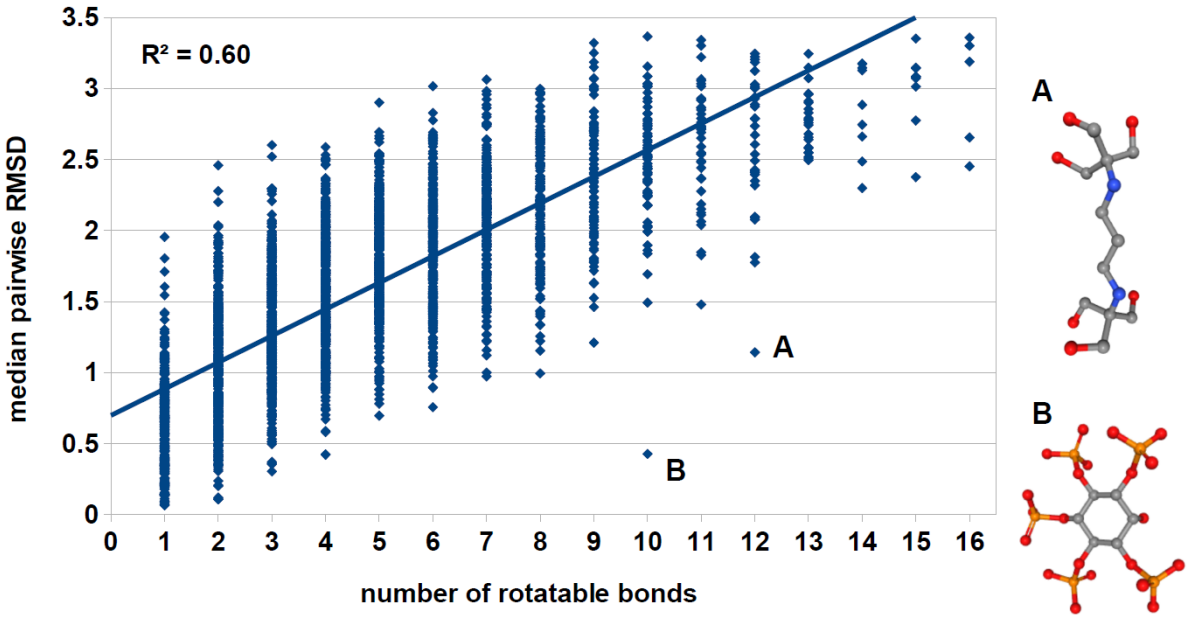


Figure S10. Median pairwise RMSD of all-against-all comparisons for each conformer ensemble generated for the Platinum Diverse Dataset with Conformator (default settings) plotted versus the number of rotatable bonds. The two labeled outliers are the highly symmetrical ligands *B3P* (A) and *5MY* (B). The R^2 for the correlation between median pairwise RMSD of all conformers and the number of rotatable bonds was 0.60.

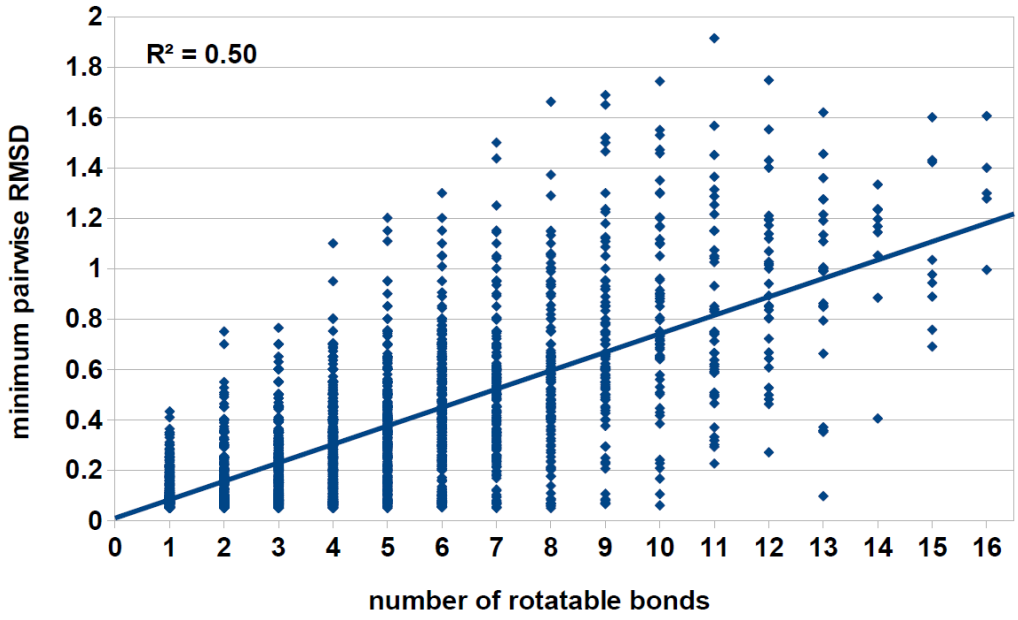


Figure S11. Minimum pairwise RMSD of all-against-all comparisons for each ensemble generated for the Platinum Diverse Dataset with Conformator (default settings) plotted versus the number of rotatable bonds. The R^2 for the correlation between median pairwise RMSD of all conformers and the number of rotatable bonds was 0.50.