JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation

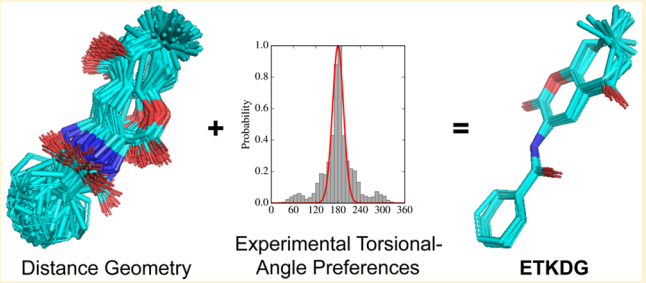
Sereina Riniker*,† and Gregory A. Landrum‡

†Laboratory of Physical Chemistry, ETH Zürich, Vladimir-Prelog-Weg 2, 8093 Zürich, Switzerland

‡Novartis Institutes for BioMedical Research, Novartis Pharma AG, Novartis Campus, 4056 Basel, Switzerland

Supporting Information

ABSTRACT: Small organic molecules are often flexible, i.e., they can adopt a variety of low-energy conformations in solution that exist in equilibrium with each other. Two main search strategies are used to generate representative conformational ensembles for molecules: systematic and stochastic. In the first approach, each rotatable bond is sampled systematically in discrete intervals, limiting its use to molecules with a small number of rotatable bonds. Stochastic methods, on the other hand, sample the conformational space of a molecule randomly and can thus be applied to more flexible molecules. Different methods employ different degrees of experimental data for conformer generation. So-called knowledge-based methods use predefined libraries of torsional angles and ring conformations. In the distance geometry approach, on the other hand, a smaller amount of empirical information is used, i.e., ideal bond lengths, ideal bond angles, and a few ideal torsional angles. Distance geometry is a computationally fast method to generate conformers, but it has the downside that purely distance-based constraints tend to lead to distorted aromatic rings and sp^2 centers. To correct this, the resulting conformations are often minimized with a force field, adding computational complexity and run time. Here we present an alternative strategy that combines the distance geometry approach with experimental torsion-angle preferences obtained from small-molecule crystallographic data. The torsional angles are described by a previously developed set of hierarchically structured SMARTS patterns. The new approach is implemented in the open-source cheminformatics library RDKit, and its performance is assessed by comparing the diversity of the generated ensemble and the ability to reproduce crystal conformations taken from the crystal structures of small molecules and protein−ligand complexes.



INTRODUCTION

The majority of small-molecule drugs and drug candidates are flexible molecules that can be expected to adopt multiple conformations in solution at room temperature. In addition, the "biologically active" conformation, i.e., that adopted inside the binding pocket, may correspond to a low-energy conformation in solution or may be a higher-energy structure induced by the protein binding.[1] The "shape" of a molecule therefore should be described by an ensemble of energetically accessible conformations. Methods like molecular dynamics simulations directly provide a Boltzmann-weighted conformational ensemble in solution (given an accurate force field), but such methods are generally too slow for drug design applications like structure-based virtual screening (VS),[2] three-dimensional (3D) ligand-based VS,[3,4] pharmacophore modeling,[5] and 3D quantitative structure−activity relationships (QSARs).[6] For these applications, fast conformer generation algorithms are used, with the drawback that solvation effects and energetics are often not explicitly considered. An important quality criterion for these algorithms is the diversity of the generated conformations, as it increases the likelihood of sampling the biologically relevant conformation(s).[7,8]

A large number of conformer generation approaches have been developed in the past decades (see ref 5 for a review). The algorithms can be broadly divided on the basis of the search strategy employed: (i) systematic and (ii) stochastic. In the first approach, conformations are generated by systematically scanning torsional-angle profiles. This ensures that all of the conformers are enumerated but restricts the methods to a relatively small number of rotatable bonds because of the combinatorial explosion. Despite this limitation, the maximum possible number of rotatable bonds is sufficient for most drug-like molecules. Stochastic algorithms, on the other hand, explore the conformational space in a random manner and can therefore be applied for molecules with any number of rotatable bonds. A popular example of a stochastic approach is distance geometry (DG)[9,10] as implemented in the open-source cheminformatics library RDKit[11] or in BALLOON.[12] In BALLOON, DG is used in combination with a genetic algorithm. DG is based on the assumption that purely geometric constraints can describe all of the possible conformations of a molecule. To this end, lower and upper

Received: October 28, 2015
Published: November 17, 2015

DOI: 10.1021/acs.jcim.5b00654
J. Chem. Inf. Model. 2015, 55, 2562−2574

distance bounds for all pairs of atoms in the molecule are determined and represented in a distance bounds matrix. In the RDKit implementation, a small amount of empirical information—ideal bond lengths, ideal bond angles, and a few ideal torsion angles—is used to construct the distance bounds matrix.

In rule- or knowledge-based methods, experimental information—typically from crystal structures—is used for conformer generation in the form of predefined libraries of torsional angles and ring conformations. Molecules are decomposed into smaller fragments and rebuilt by searching libraries with known conformations of these fragments in a systematic or stochastic way. Examples of rule/knowledge-based approaches include Corina,[13] OMEGA,[14] Confab,[15] ConfGen,[16] TrixX (TCG),[17] CAESAR,[18] FROG2,[19,20] CONFECT,[21] COSMOS[22] and BCL::CONF.[23]

The DG method as implemented in the RDKit has been found to perform better than some of the knowledge-based approaches for a test set of 708 molecules.[8] Though DG can perform quite well, the generated conformations can have distorted aromatic rings and sp^2 centers as well as torsional-angle values that are outside the ranges observed in experimental crystal structures. A remedy for the first issue has been proposed by Crippen et al.[24] in the form of a general linearized embedding algorithm. In this study, we present an approach to resolve both issues simultaneously. Recently, Schärfer et al.[25] developed a hierarchical series of SMARTS patterns encoding torsional substructures, which are at the core of the CONFECT[21] conformer generator. For each pattern, the authors obtained torsional-angle distributions from the Cambridge Structural Database (CSD)[26,27] and used these experimental torsional-angle preferences in conformer generation. The positions and widths of the peaks of the distributions were published for two different tolerance levels.[25]

In this study, we investigate the inclusion of experimental torsional-angle preferences in the RDKit's DG conformer generation, termed Experimental-Torsion Distance Geometry (ETDG). In addition, a variant termed ETKDG with additional "basic knowledge" (K) terms such as flat aromatic rings and linear triple bonds is presented. The first idea was to use the torsional-angle ranges from ref 25 to set improved upper and lower bounds for 1,4-neighbors in the distance bounds matrix. For this, 1,4-distances are calculated using ideal bond lengths and ideal bond angles together with the desired torsional angle. However, any deviations of bond lengths and angles from their ideal values during the embedding step invalidates these calculations, and the 1,4-distance bounds end up corresponding to different torsional-angle ranges. We therefore pursued an alternative approach and introduced an additional minimization step using torsional-angle potentials (torsional-angle restraints did not yield satisfactory results). The potentials were fitted to torsional-angle distributions obtained from the CSD for the patterns described in ref 25 and fine-tuned by monitoring the resulting torsional-angle distributions from the generated conformations of more than 1500 molecules. The performance of the new approach in reproducing crystal structures is compared with that of standard DG and CONFECT for two validation sets consisting of 1290 molecules from the CSD and 238 molecules from the RCSB Protein Data Bank (PDB).[28]

The ETDG and ETKDG methods were implemented in the open-source cheminformatics toolkit RDKit and are freely available under the BSD license starting with the 2015.09.1 release of the RDKit.

METHODS AND MATERIALS

Torsional Potentials. From the 392 SMARTS patterns for central bonds C−O, N−C, S−N, C−S, C−C, and S−S described in ref 25, 387 were considered here (the other five patterns use the lone pair of a nitrogen atom as a fourth atom). For each pattern, torsional-angle statistics were collected from the CSD. The same filters as used in ref 25 were applied to select entries from the CSD, i.e., only molecules containing H, C, N, O, F, Cl, Br, I, S, and P were considered. Ions, powder structures, organometallic compounds, and structures with an R factor larger than 10% were omitted. Each acyclic single bond of a molecule could be the target of multiple patterns. A list of the 387 patterns is given in the Supporting Information.

Parametrization. From the collected torsional angles, a histogram was calculated in the range [0°−360°] with a stepsize of 10° for each pattern. Thirteen patterns were omitted because insufficient data were available for fitting or the experimentally observed angles covered the whole range. Two different functional forms were fitted to the histograms of the remaining 374 torsions:

$$V(\phi) = K[1 + \cos(d)\cos(m\phi)] \quad (1)$$

and

$$V(\phi) = \sum_{i=1}^{6} K_i[1 + \cos(d_i)\cos(i\phi)] \quad (2)$$

where K is the force constant, d is the phase shift, and m is the multiplicity. For both potentials, the phase shift is restricted to 0 or π. The multiplicity in eq 1 can take values from 1 to 6. In eq 1 there are three parameters to fit, and eq 2 has 12 parameters. The final potentials for the patterns were selected manually. Whenever possible, the potential of eq 1 was used because of the smaller number of parameters and a straightforward way to fine-tune the force constants.

The force constants were further fine-tuned manually to obtain reasonable torsional-angle distributions from the generated conformers of the three data sets described below. A list of the parameters of the final potentials can be found in the RDKit implementation.[11]

As the SMARTS patterns from Schärfer et al.[25] are only for acyclic single bonds, a generic pattern for double bonds (`[*:1][X3,X2:2]=[X3,X2:3][*:4]`) was added to help keep double bonds flat (i.e., torsional angles at 0° or 180°). The torsional-angle potential for the double-bond pattern uses eq 1 with K = 7.0, $\cos(d) = -1.0$, and m = 2.

Data Sets. Three different compound collections were used in this study. The first test set consisted of 1290 distinct small molecules from the CSD. Of these, 469 were used before by Ebejer et al.[8] and are based on the work of Hawkins et al.,[14] whereas the other 821 molecules were selected from the CSD following the same procedure as described by Hawkins and co-workers. Two additional compounds found by the procedure were not considered: the CSD entry CUMHOH was discarded because a chiral carbon was planar in the crystal structure, and the CSD entry FIZNIL was ignored because the current DG implementation in the RDKit uses an angle for the sulfur in the fused ring system that is too large.

The second test set consisted of 238 crystal structures of drug-like molecules bound to proteins from the PDB, as used in ref 8. Of these structures, 79 were taken from the Astex Diverse Set,[29] and the other 159 molecules were selected by Ebejer et al.[8] from the 197 structures of the original data set in ref 14.

DOI: 10.1021/acs.jcim.5b00654
J. Chem. Inf. Model. 2015, 55, 2562−2574

The PDB entry 1PCK was discarded because a chiral carbon was planar in the crystal structure.

The two test sets described above were used for parametrization of the torsional potentials as well as validation of the ETDG and ETKDG approaches. In the combined set of 1528 compounds, 277 of the 374 torsional patterns were present. In order to include more patterns, a third set was collected from the CSD by selecting compounds with at least one of the missing patterns, a molecular weight between 100 and 800 g/mol, less than 11 rotatable bonds, and a crystallographic R factor smaller than 0.1. For each torsional pattern, a maximum of 10 compounds were considered (the 10 smallest based on molecular weight). This resulted in 431 additional compounds in the third set, which was used only for parametrization of the torsional potentials. In the combined set of 1959 molecues, 302 of the 374 torsional patterns were covered.

A list of CSD identifiers and a list of PDB identifiers for all three data sets are given in the Supporting Information.

Conformer Generation. For each compound, conformations were generated using four different methods: (i) standard DG[9,10] as implemented in the RDKit, (ii) ETDG, (iii) ETKDG, and (iv) CONFECT.[21] The target number of conformers was 100 with a root-mean-square (RMS) threshold for pruning of 1.0 Å. The source code for ETDG and ETKDG is open and freely available under the BSD license starting with the 2015.09.1 release of the RDKit. The results described here were obtained using a prerelease version.[30]

DG. For standard DG, the implementation in the RDKit[11] was used. The maximum number of attempts for embedding was set to 200. The tests to enforce correct chirality were applied.

ETDG. The ETDG approach has been implemented in the RDKit as an optional feature of the normal DG method. The steps of the algorithm are shown in Figure 1. It should be noted that only steps 2 and 8 distinguish ETDG from standard DG.

The patterns in ref 25 are hierarchically ordered with decreasing specificity. We therefore process the patterns in the same order, and a given path of four atoms is allowed to match only one pattern (i.e., the most specific one). Each bond can have only one torsional potential applied.

In step 7, the refinement with a fixed maximum number of iterations of 200 is repeated until the reduction of the error function is below a certain threshold. In step 8, there is only a single minimization with a fixed maximum number of iterations of 300. The choice of the maximum number of iterations is a balance between accuracy and computational cost.

The other parameters are set as used for DG.

ETKDG. The ETKDG approach follows the same procedure as ETDG. For the minimization in step 7, the following additional "basic knowledge" (K) terms are added:

- Torsional-angle potentials for bonds in aromatic rings (maximum one potential per bond) using eq 1 with K = 7.0, $\cos(d) = -1.0$, and $m = 2$.
- UFF[31] inversion terms for all sp^2 atoms that are N, O, or C.
- Angles involving triple bonds are constrained to 180° using the RDKit's UFF angle constraint terms.

The other parameters are set as used for ETDG.

CONFECT. A command-line version of CONFECT (version 2.0.0) was obtained from BioSolveIT.[32] The quality level was set to 21, and RMS clustering with a cluster threshold of 0.1

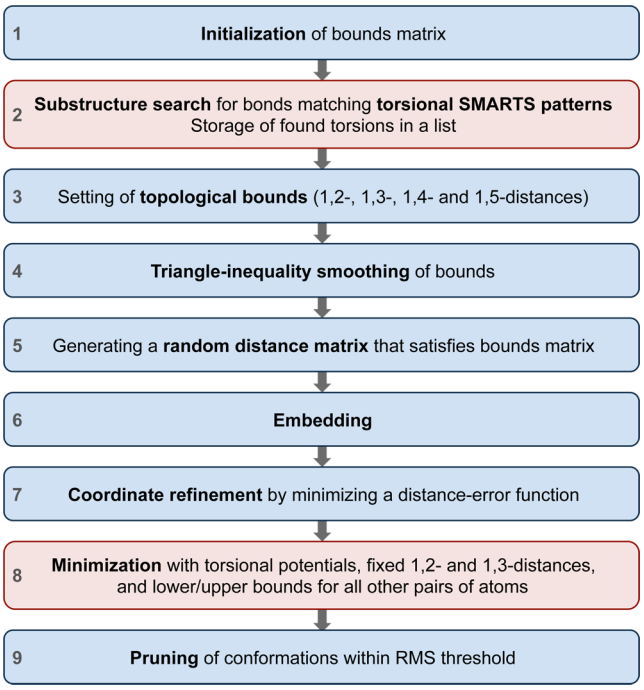


Figure 1. Steps of the standard DG algorithm (blue rectangles) as implemented in the RDKit with the two additional steps introduced by the ETDG approach shown in red.

was used. Conformers were optimized, and ring conformations were generated. Coordinates of hydrogens were regenerated, and no default protonation was calculated.

Analysis. *Root-Mean-Square Deviation.* The atom-positional root-mean-square deviation (RMSD) between the crystal structure and a generated conformation of a molecule was determined as

$$\mathrm{RMSD} = \sqrt{\frac{1}{N_{\mathrm{atoms}}} \sum_{i=1}^{N_{\mathrm{atoms}}} (\mathbf{r}_i - \mathbf{r}_{i,\mathrm{ref}})^2} \tag{3}$$

where N_{atoms} is the number of non-hydrogen atoms considered, $\mathbf{r}_i$ is the position of atom i, and $\mathbf{r}_{i,\mathrm{ref}}$ is the position of atom i in the reference configuration (i.e., the crystal structure). It should be noted that the two sets of atom coordinates used in eq 3 are those generated by optimal rigid-body superposition (with symmetry taken into account). This is implemented in the `GetBestRMS()` function in the RDKit. The RMSD was calculated for all of the generated conformers, and the best value was recorded. To access the diversity of the generated conformers, the matrix of "interconformer" RMSD (icRMSD) values was determined and the median recorded. We consider only differences in RMSD greater than 0.5 as practically relevant. This cutoff is indicated in the plots by dashed lines.

Torsion Fingerprint Deviation (TFD). An alternative to RMSD for comparison of conformations is the torsion fingerprint deviation (TFD) approach developed by Schulz-Gasch et al.[33] in 2012. The torsional angles of the nonterminal acyclic bonds and ring systems are extracted from two conformations and weighted according to their distance from the center of the molecule, and and the difference is recorded. The TFD can take values in the range [0, 1], where zero represents perfect alignment. The topological weighting ensures that changes of the torsional angle in the core of the molecule are more important than changes toward the edges.

DOI: 10.1021/acs.jcim.5b00654
J. Chem. Inf. Model. 2015, 55, 2562−2574

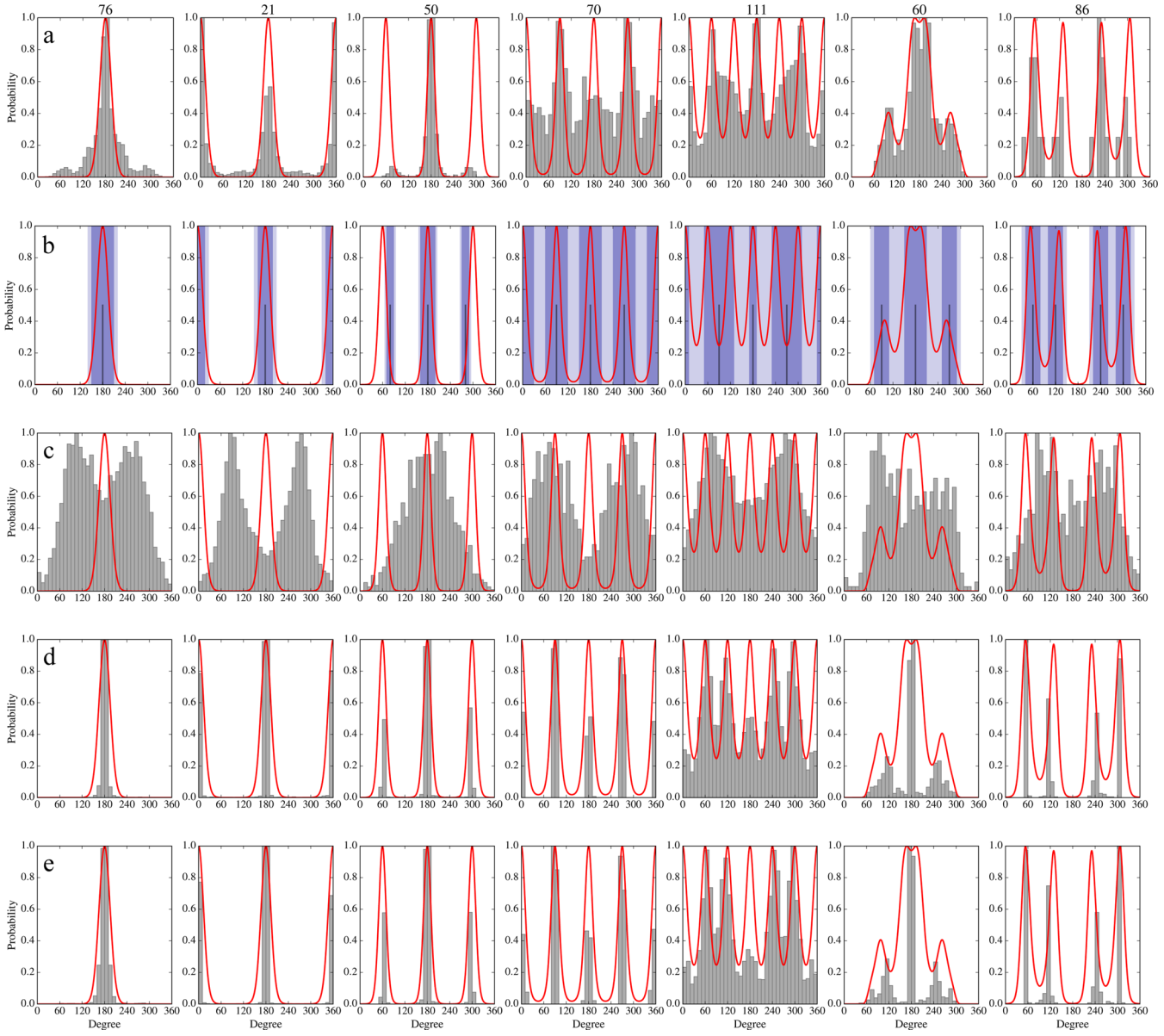


Figure 2. Comparison of the final torsional potentials (red lines) for seven selected patterns in the range [0°−360°] for the combined data sets (1959 molecules) with (a) normed torsional-angle distributions from the CSD used for fitting; (b) ranges from ref 25 (purple rectangles for tolerance 1, light-purple rectangles for tolerance 2, dark-purple lines for the peaks); (c) normed torsional-angle distributions in conformers generated with DG; (d) normed torsional-angle distributions in conformers generated with ETDG; and (e) normed torsional-angle distributions in conformers generated with ETKDG.

The TFD approach has been implemented in the RDKit. In TFD, a quad of atoms *a*−*b*−*c*−*d* is selected for each torsion on the basis of the 21 cases described in Figure 2 in ref 33. In the RDKit implementation, symmetric atoms *a* and/or *d* are determined on the basis of ECFP[34] atom types with a user-defined radius (default radius = 2). If multiple nonsymmetric atoms *a* and/or *d* are present, the atom with the smallest ECFP atom invariant is chosen for the torsional-angle calculation. For torsions involving symmetric atoms *a* and/or *d*, all possible deviations are determined, and the smallest one is used for the TFD calculation. Hydrogens are always excluded. The maximum possible deviation can either be set for each torsion type individually or set uniformly to 180° (default) as done by Schulz-Gasch et al.[33] Single bonds adjacent to triple bonds and allenes can be excluded (default), as done by Schulz-Gasch et al.,[33] or a "combined torsion" can be assigned, in which the first non-colinear atoms are used to define the torsion. The following parameters were used in this study: a fixed maximum deviation of 180°, a radius to determine symmetric atoms of 2, and "combined torsions" for single bonds adjacent to triple bonds and allenes. To access the diversity of the generated conformers, the matrix of "interconformer" TFD (icTFD) values was determined and the median recorded. We consider only differences in TFD greater than 0.2 as practically relevant.[33] This cutoff is indicated in the plots by dashed lines.

Timings. Introducing the ET and K terms in the force field used to refine the basic DG conformations adds computational complexity and has an impact on the amount of time required to generate conformations. To measure this, we carried out benchmark tests using 1665 of the molecules from our set of experimental crystal structures. For these tests we generated 100 conformations for each molecule without using RMSD

DOI: 10.1021/acs.jcim.5b00654
J. Chem. Inf. Model. 2015, 55, 2562−2574

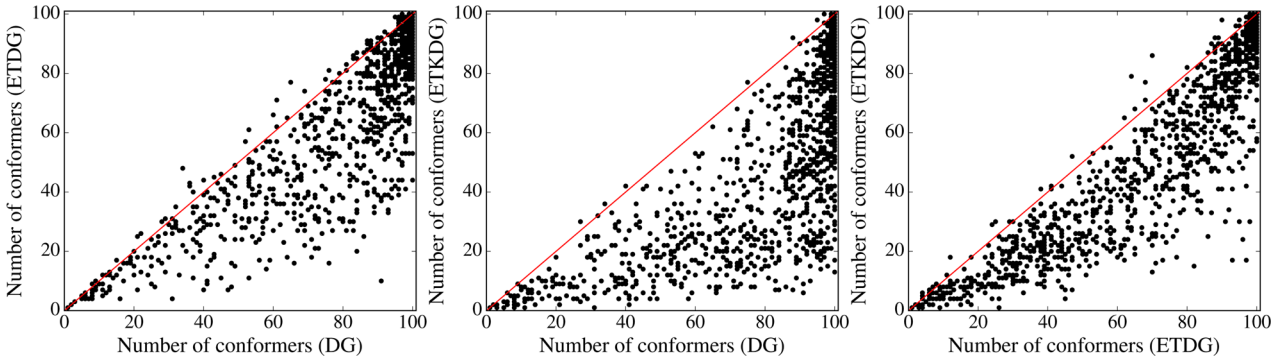


Figure 3. Comparison of the numbers of effectively generated conformers given a target number n = 100 and an RMS threshold of 1.0 Å for DG, ETDG, and ETKDG using the CSD data set (1290 molecules). The red lines represent $y = x$.

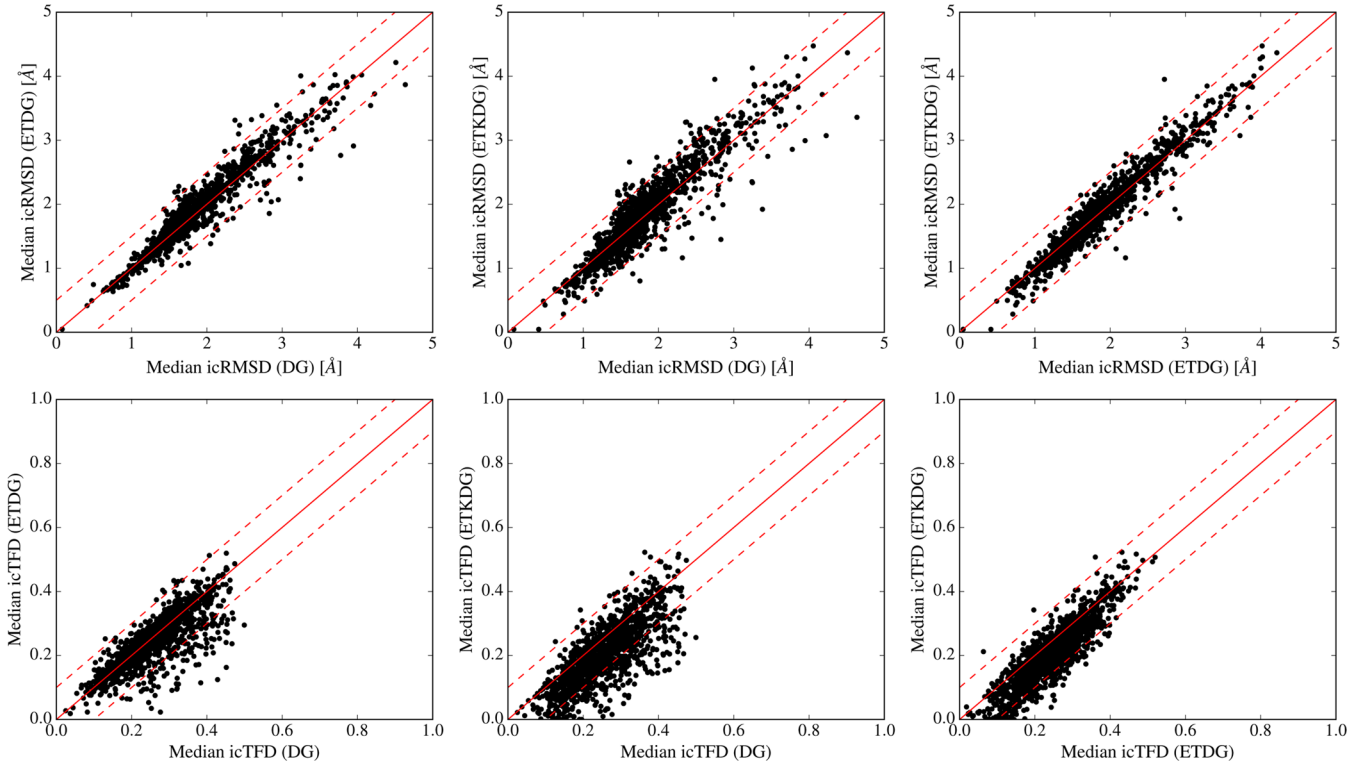


Figure 4. Comparison of (top) the median of interconformer RMSD (icRMSD) values and (bottom) the median of interconformer TFD (icTFD) values for DG, ETDG, and ETKDG. The CSD data set (1290 molecules) was used. In each panel, the red solid line represents $y = x$, and the red dashed lines indicate differences of ΔRMSD = 0.5 or ΔTFD = 0.2.

pruning. The same random-number seeds were used in all of the calculations, so the minimizations with the distance-error function (step 7 above) started from the same point for each molecule and conformer. The experiments were run on a Dell Dimension XPS workstation with a 3.6 GHz Intel Core i7-4790 CPU and 16 GB of RAM running Ubuntu Linux 15.04 and Python 3.4.3.

RESULTS AND DISCUSSION

Fitted Torsional Potentials. The torsional-angle distributions from the CSD used for fitting, together with the final fine-tuned potentials implemented in the RDKit, are shown in Figure 2a for a selected subset of the torsion patterns. The subset was chosen to represent the different phase shift and multiplicity combinations observed for eq 1 and different examples of eq 2. The distributions for all 387 patterns are shown in Figure S1 in the Supporting Information. In general, potentials using eq 2 allow a close fitting of the distributions (see, e.g., patterns 60 and 86 in Figure 2b). However, when fine-tuning of the force constants was required, potentials using eq 1 were found to be more robust and easier to tune. Thus, those potentials were preferentially used (see, e.g., patterns 76, 21, 50, 70, and 11 in Figure 2b).

A comparison of the final potentials with the torsion ranges described in ref 25 is shown in Figure 2b for the selected subset of patterns and for all patterns in Figure S2 in the Supporting Information. With a few exceptions, the peaks of the ranges match the minima of the potentials. For 20 patterns (10, 12, 85, 97, 146, 150, 152, 222, 167, 252−254, 295, 298, 313, 327, 361, 366, 368, and 379), we decided to include smaller peaks that were not considered in ref 25.

The lower three panels in Figure 2 show the torsional-angle distributions for the subset of patterns observed in the generated conformers of DG, ETDG, and ETKDG. The

DOI: 10.1021/acs.jcim.5b00654
J. Chem. Inf. Model. 2015, 55, 2562−2574

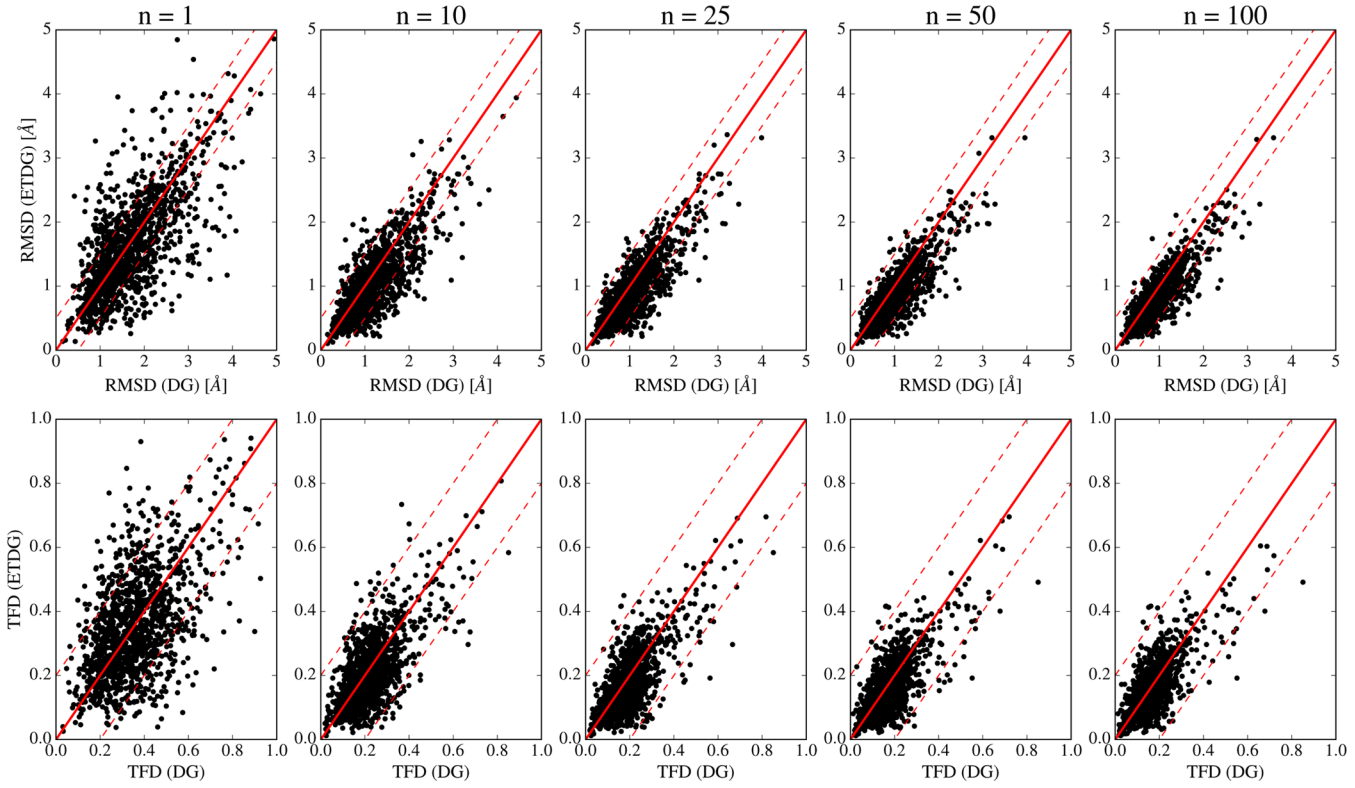


Figure 5. Comparison of the best (top) RMSD and (bottom) TFD values (using the crystal conformation as the reference) for DG and ETDG as a function of the number of conformers considered, *n*, for the CSD data set (1290 molecules). In each panel, the red solid line represents $y = x$, and the red dashed lines indicate a difference of ΔRMSD = 0.5 or ΔTFD = 0.2.

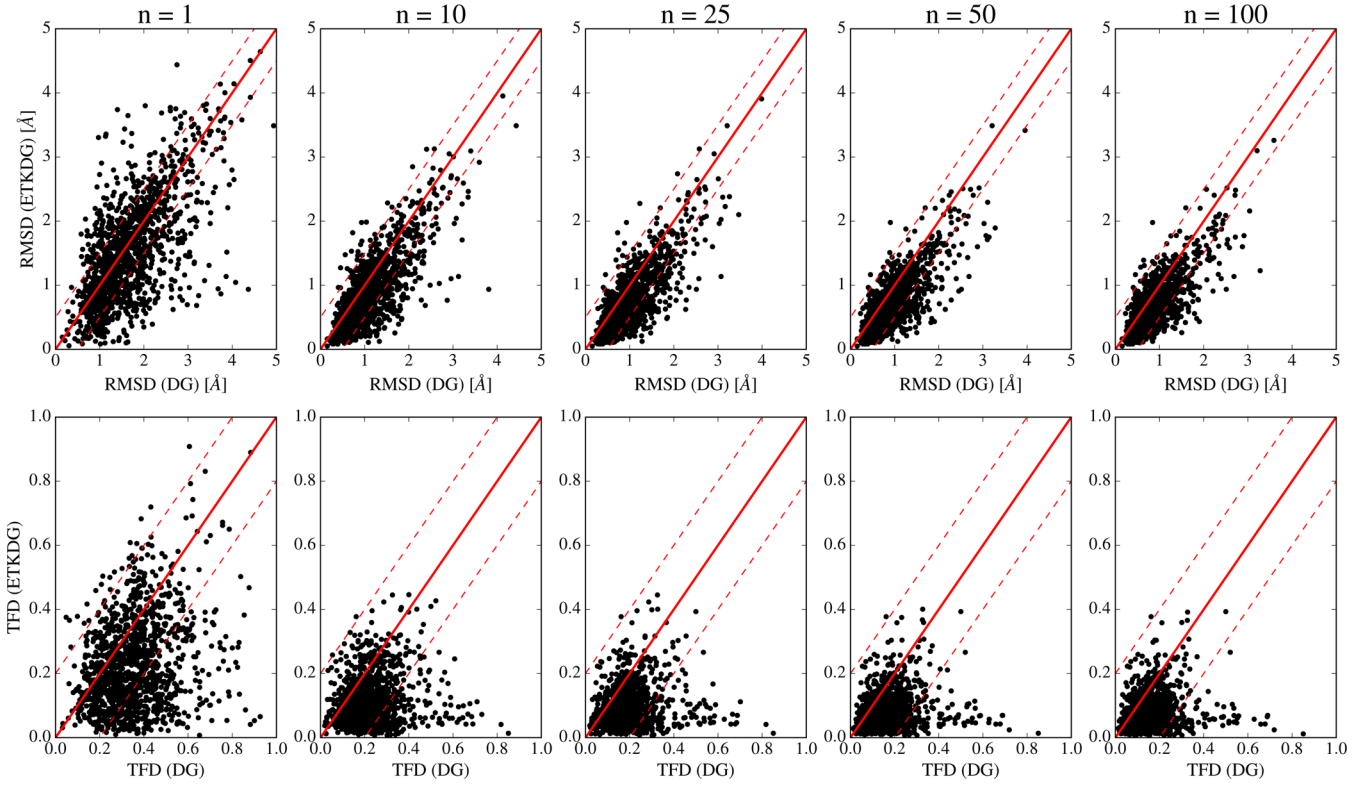


Figure 6. Comparison of the best (top) RMSD and (bottom) TFD values (using the crystal conformation as the reference) for DG and ETKDG as a function of the number of conformers considered, *n*, for the CSD data set (1290 molecules). In each panel, the red solid line represents $y = x$, and the red dashed lines indicate a difference of ΔRMSD = 0.5 or ΔTFD = 0.2.

DOI: 10.1021/acs.jcim.5b00654
J. Chem. Inf. Model. 2015, 55, 2562−2574

Table 1. p Values from One-Sided Paired t Tests of RMSD and TFD Values for DG, ETDG, and ETKDG for Different Numbers of Conformers, n, Using the CSD and PDB Data Sets[a]

			p values				
methods	data set	measure	$n = 1$	$n = 10$	$n = 25$	$n = 50$	$n = 100$
DG−ETDG	CSD	RMSD	2.3×10^{-6}	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
		TFD	3.9×10^{-2}	1.42×10^{-10}	$<2.2 \times 10^{-16}$	1.6×10^{-15}	1.4×10^{-10}
	PDB	RMSD	$\underline{1.8 \times 10^{-1}}$	2.7×10^{-8}	3.5×10^{-14}	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
		TFD	$\underline{1.3 \times 10^{-1}}$	7.8×10^{-4}	8.2×10^{-5}	1.2×10^{-4}	4.3×10^{-2}
DG−ETKDG	CSD	RMSD	2.7×10^{-16}	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
		TFD	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
	PDB	RMSD	8.3×10^{-3}	3.6×10^{-16}	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
		TFD	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
ETDG−ETKDG	CSD	RMSD	2.8×10^{-5}	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
		TFD	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
	PDB	RMSD	$\underline{5.9 \times 10^{-2}}$	1.2×10^{-4}	1.4×10^{-5}	1.2×10^{-5}	5.5×10^{-6}
		TFD	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$

[a]The null hypothesis states that there is no difference between the RMSD values ($\mu = \mu_0 = 0$), and the alternative hypothesis is that the difference is greater than zero ($\mu > \mu_0$), i.e., the second method performs better than the first method. If the p value is smaller than 0.05, the null hypothesis is discarded. Cases where the null hypothesis *cannot* be discarded are underlined.

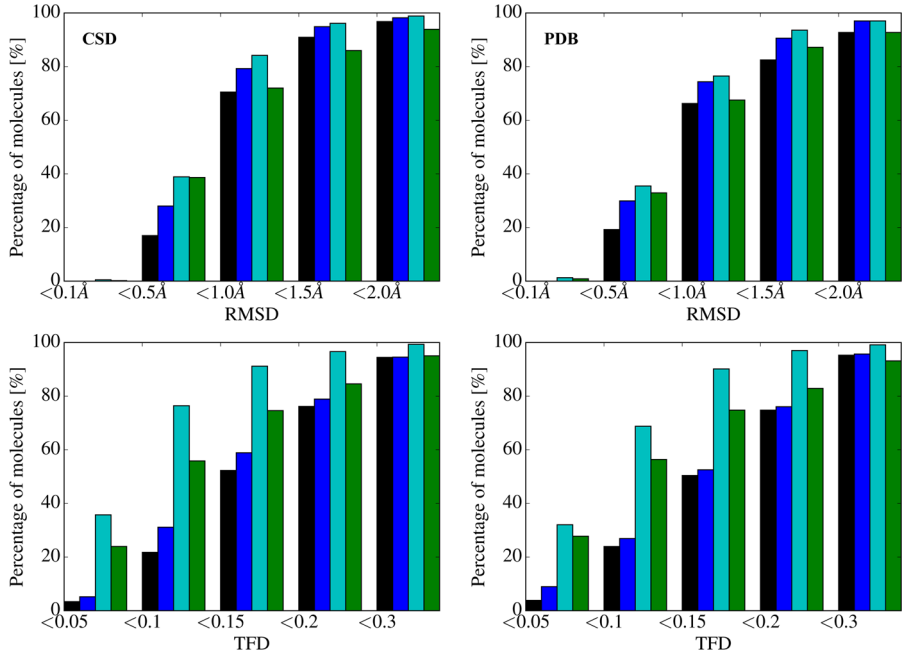


Figure 7. Percentages of molecules where the crystal conformation was reproduced within a certain RMSD cutoff (0.1, 0.5, 1.0, 1.5, or 2.0 Å) or TFD cutoff (0.01, 0.05, 0.1, 0.2, of 0.3) using (left) the CSD data set (1290 molecules) and (right) the PDB data set (238 molecules). DG is shown in black bars, ETDG in blue, ETKDG in cyan, and CONFECT in green. $n = 100$ was used for all methods.

distributions of all 302 patterns present in the data sets are shown in Figures S3, S4, and S5 in the Supporting Information, respectively. For DG, the torsion angles in the generated conformers are often very different from the ones observed in the crystal structures (Figures 2c and S3). The use of the torsion potentials in ETDG and ETKDG results—as expected—in torsional-angle distributions matching the experimental preferences, although the distributions in the generated conformers tend to be more narrow. ETDG and ETKDG were found to give the same distributions (Figures 2d,e, S4, and S5). This indicates that the K terms and the torsion terms are independent.

Diversity Analysis. Diversity is an important quality criterion for conformer generators, as it indicates better sampling of the possible conformational space and increases the chances of finding the biologically relevant or crystal conformation(s) among the generated conformers. A first indication of diversity is the number of generated conformers given a target number $n = 100$ and an RMS pruning threshold of 1.0 Å. The numbers of conformers for DG, ETDG, and ETKDG are compared in Figure 3. The use of the experimental torsional-angle preferences alone in ETDG leads on average to a decrease in the number of conformers (i.e., average number of conformers = 70.8 for ETDG vs 82.6 for DG) because of the narrower ranges for certain torsions. However, the decrease is smaller than the one observed with ETKDG (i.e., average number of conformers = 56.2) because the K terms enforce flat aromatic rings and linear triple bonds. A decrease in the number of generated conformers is not automatically a

DOI: 10.1021/acs.jcim.5b00654
J. Chem. Inf. Model. 2015, 55, 2562−2574

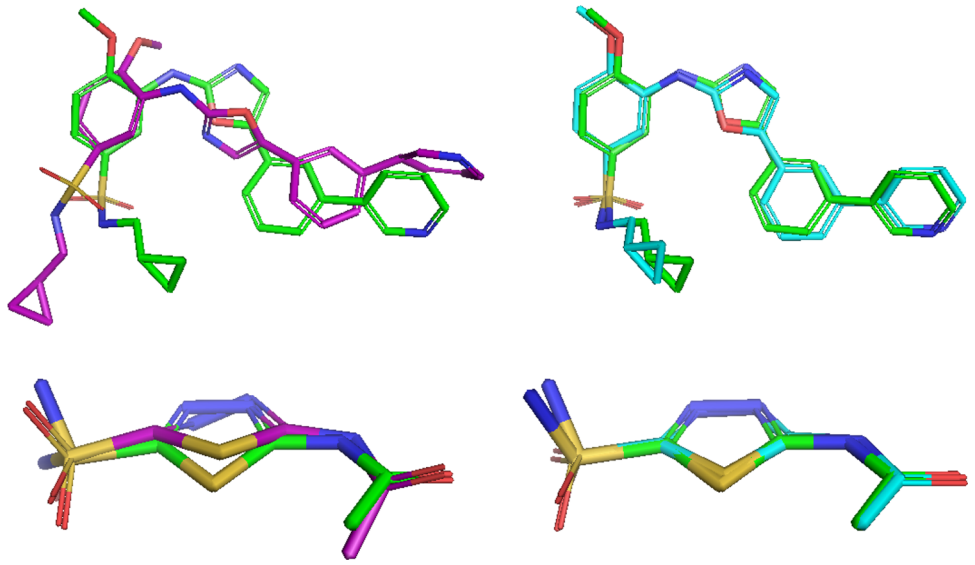


Figure 8. Cases in the PDB data set where ETKDG outperforms DG (n = 100). Shown are overlays of the crystal structures (green) with the best generated conformers of DG (purple) and ETKDG (cyan) for (top) PDB entry 1Y6B (best RMSD value of DG = 2.66 Å, best RMSD value of ETKDG = 0.71 Å) and (bottom) PDB entry 1JD0 (best TFD value of DG = 0.58, best TFD value of ETKDG = 0.01). Alignment of the structures was done using the `AlignMols()` function in the RDKit. Figures were generated using PyMol.[35]

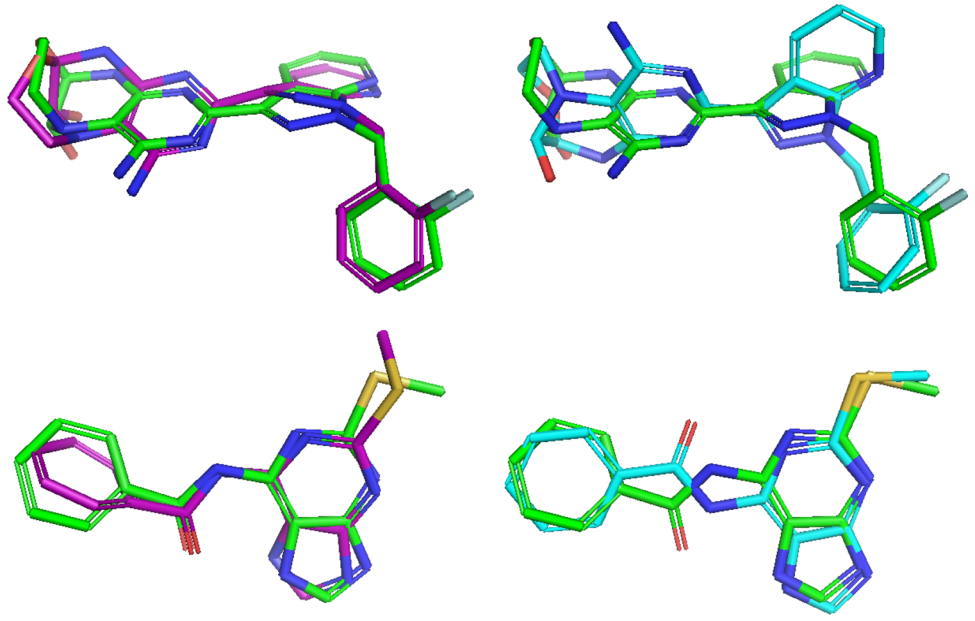


Figure 9. Cases in the CSD data set where DG outperforms ETKDG (n = 100) by ΔRMSD > 0.5 Å or by ΔTFD > 0.2. Shown are overlays of the crystal structures (green) with the best generated conformers of DG (purple) and ETKDG (cyan) for (top) CSD entry EWEKAS (best RMSD value of DG = 0.84 Å, best RMSD value of ETKDG = 1.98 Å) and (bottom) CSD entry MTBPNP (best TFD value of DG = 0.16, best TFD value of ETKDG = 0.38). Alignment of the structures was done using the `AlignMols()` function in the RDKit. Figures were generated using PyMol.[35]

disadvantage, as many conformers from DG and ETDG may not be chemically reasonable.

More important than the number of conformers is the diversity among them. To this end, the medians of the interconformer RMSD (icRMSD) and interconformer TFD (icTFD) values were recorded for DG, ETDG, and ETKDG (Figure 4). They describe the diversity within the generated conformational ensemble independent of its size. Conformers with an RMSD smaller than 1.0 Å were pruned during generation, and therefore, the icRMSD values should, by construction, be greater than 1.0 Å. The comparison in Figure 4 shows that the different approaches yield similarly diverse conformers in terms of icRMSD. For icTFD, ETKDG yields less diverse conformers because the K terms reduce deviations in the ring conformations, which are also part of TFD. The experimental torsional-angle preferences have a smaller effect on the TFD values than the K terms do.

Reproducing Crystal Conformations. The performance of conformer generators is typically tested by assessing their ability to reproduce crystal conformations from either small-molecule crystal structures or protein−ligand complexes. To this end, sets of conformations (upper limit n = 5, 10, 25, 50, or 100) were generated, and the best RMSD and TFD values were recorded. The performance of ETDG and ETKDG was

DOI: 10.1021/acs.jcim.5b00654
J. Chem. Inf. Model. 2015, 55, 2562−2574

compared to that of standard DG using the CSD data set (Figures 5 and 6) and the PDB data set (Figures S6 and S7 in the Supporting Information). ETDG performed better than DG for more than 10 conformers in terms of both RMSD and TFD values, as confirmed by one-sided paired *t* tests (Table 1). The number of molecules that could be reproduced below a certain RMSD or TFD cutoff increased (Figure 7). Interestingly, the TFD values did not improve as much as could be expected from applying torsional-angle preferences that were derived from crystal structures. The reason for this is that TFD includes ring conformations. In ETDG, the rings are not affected by the experimental torsional-angle preferences but remain as distorted as in DG.

The performance could be further improved by applying the additional K terms (Figure 6 and Figure S7 in the Supporting Information). ETKDG performed statistically significantly better than both DG and ETDG for more than 10 conformers (Table 1). With ETKDG, the crystal conformations of 38% of the molecules in the CSD data set could be reproduced within 0.5 Å and 84% within 1.0 Å (Figure 7). Using the K terms alone, however, was not sufficient to match the performance of ETKDG (Figures S8 and S9 and Table S1 in the Supporting Information). When the average best RMSD was plotted as a function of the number of rotatable bonds (Figure S10 in the Supporting Information), we found the same overall trend of increasing RMSD with increasing number of rotatable bonds as Ebejer et al.,[8] with ETKDG outperforming the other methods.

Case Studies. To highlight the benefit of the new approach, Figure 8 shows for PDB entries 1Y6B and 1JD0 the overlays between the crystal conformation and the best conformers of DG and ETKDG. For PDB entry 1Y6B, the improvement with ETKDG was biggest in terms of RMSD. Whereas the best DG conformer is much more elongated compared with the crystal conformation, the best ETKDG conformer differs only in a few torsions of the side chains. The biggest improvement in terms of TFD was found for PDB entry 1JD0. The combination of the experimental torsion-angle preferences and the K terms leads to an almost perfect reproduction of the crystal conformation by ETKDG.

In the PDB data set, the best ETKDG conformer was closer to or at least no further away from the crystal conformation than the best DG conformer by both ΔRMSD > 0.5 Å and ΔTFD > 0.2. The biggest outliers in the CSD data set in terms of RMSD and TFD are shown in Figure 9. In the case of CSD entry EWEKAS, the discrepancy between ETKDG and the crystal is the torsion around the central bond connecting the two hetero ring systems. The pattern governing this bond is no. 275, which has a single peak at 180°. For this specific compound, pattern 275 would match both ways, but the one picked is the opposite to that in the crystal structure, leading to a large deviation. A similar reason is behind the discrepancy between ETKDG and the crystal conformation of CSD entry MTBPNP: pattern 77, which governs the bond connecting the amide moiety to the purine ring, has a single peak at 180°. The crystal conformation of MTBPNP is a rare example where this torsion is experimentally found around 0°. This example shows nicely the limitation of ETKDG and knowledge-based conformer generators in general: the only crystal conformations that can potentially be reproduced are those that exhibit "normal" torsional angles.

Comparison with the Knowledge-Based Conformer Generator CONFECT. As the purely knowledge-based conformer generator CONFECT uses the same torsional-angle patterns as ETKDG, its performance was compared with that of ETKDG. Interestingly, the number of conformers generated by CONFECT is not correlated with that generated by ETKDG (Figure 10). Compounds with a high number of ETKDG

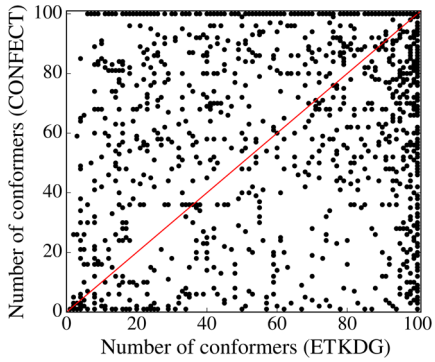


Figure 10. Comparison of the numbers of effectively generated conformers given a target number $n = 100$ and a RMS threshold of 1.0 Å for CONFECT and ETKDG using the CSD data set (1290 molecules). The red line represents $y = x$.

conformers and a low number of CONFECT conformers typically contain large fused ring systems (data not shown). The conformations of nonaromatic rings are also knowledge-based in CONFECT, resulting in only one or a few preferred conformations, whereas in ETKDG the nonaromatic ring conformations are derived from DG, which yields diverse conformations. On the other hand, compounds with a low number of ETKDG conformers and a high number of CONFECT conformers typically contain torsions that exhibit sharp peaks in the CSD distributions (data not shown). These cases illustrate the difference between the use of torsional-angle ranges and torsional-angle potentials. For example, a range of ±20° around 0° leads to more diverse conformations compared with a potential with a minimum at 0°.

The comparisons of the best RMSD and TFD values from CONFECT and ETKDG are shown in Figure 11 for the CSD data set and in Figure S11 in the Supporting Information for the PDB data set. The results from the one-sided paired *t* tests are given in Table S2 in the Supporting Information. For $n = 1$, CONFECT performs better than or similar to ETKDG, whereas for $n > 1$ ETKDG reproduces the crystal conformations better. While CONFECT reproduces more compounds with RMSD < 0.5 Å compared with DG and ETDG and outperforms these two methods in terms of TFD (Figure 7), the performance of ETKDG is superior to CONFECT, which indicates that the use of torsional-angle potentials is advantageous.

Comparison with Force-Field-Minimized DG. Because of the distorted aromatic rings and sp^2 centers that are often obtained from DG, this method is generally used in conjuction with an additional minimization using a force field (FF) such as UFF[31] or the Merck Molecular Force Field (MMFF).[36,37] This FF-optimization step is relatively slow and is performed subsequent to the RMS pruning, thus requiring an additional clustering step to remove close conformations. One motivation behind ETKDG is to present a computationally more efficient alternative to the DG + FF-optimization approach. As the minimization step using the ETK terms is within the DG conformer generation, it is directly considered by the RMS pruning. The comparison of the best RMSD and TFD values

DOI: 10.1021/acs.jcim.5b00654
J. Chem. Inf. Model. 2015, 55, 2562−2574

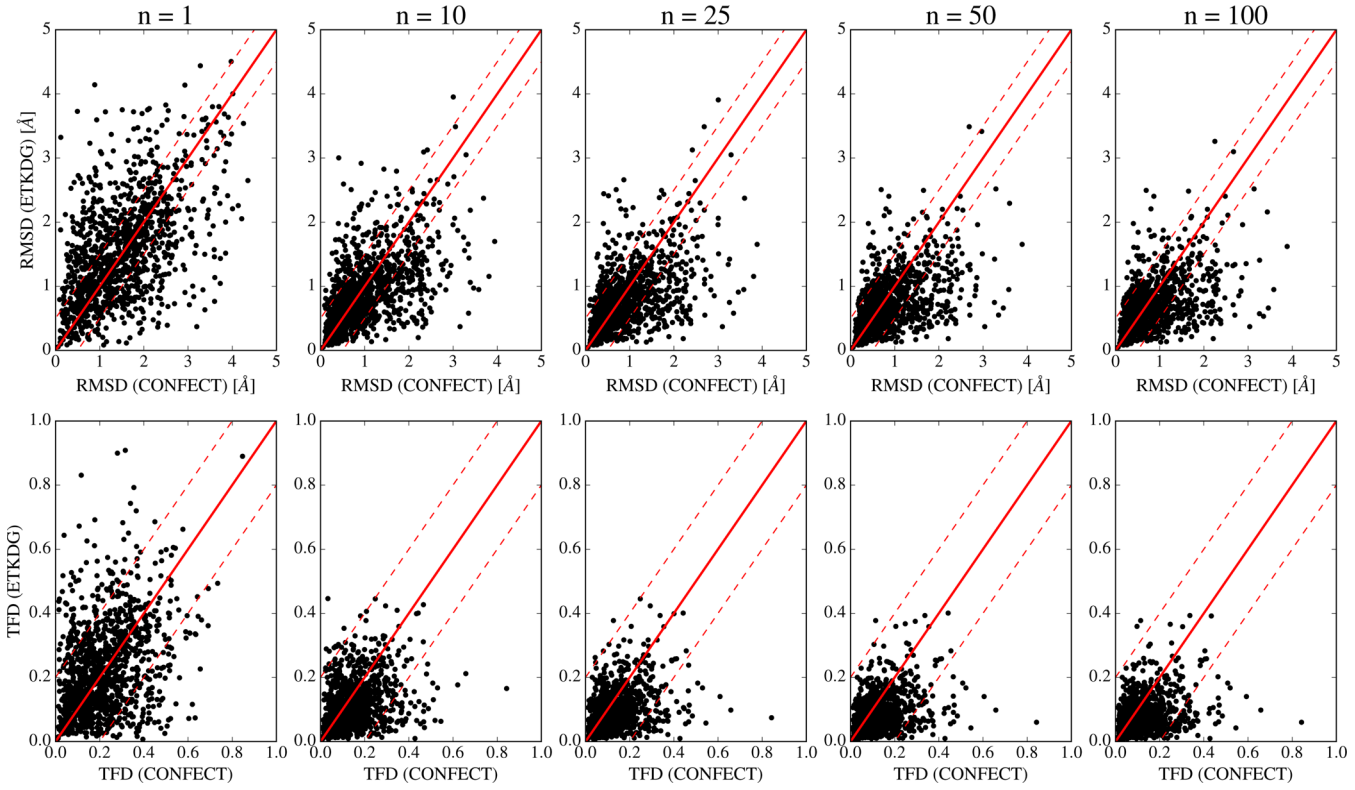


Figure 11. Comparison of the best (top) RMSD and (bottom) TFD values relative to the crystal structure for CONFECT and ETKDG as a function of number of conformers considered, *n*, for the CSD data set (1290 molecules). In each panel, the red solid line represents $y = x$, and the red dashed lines indicate a difference of ΔRMSD = 0.5 or ΔTFD = 0.2.

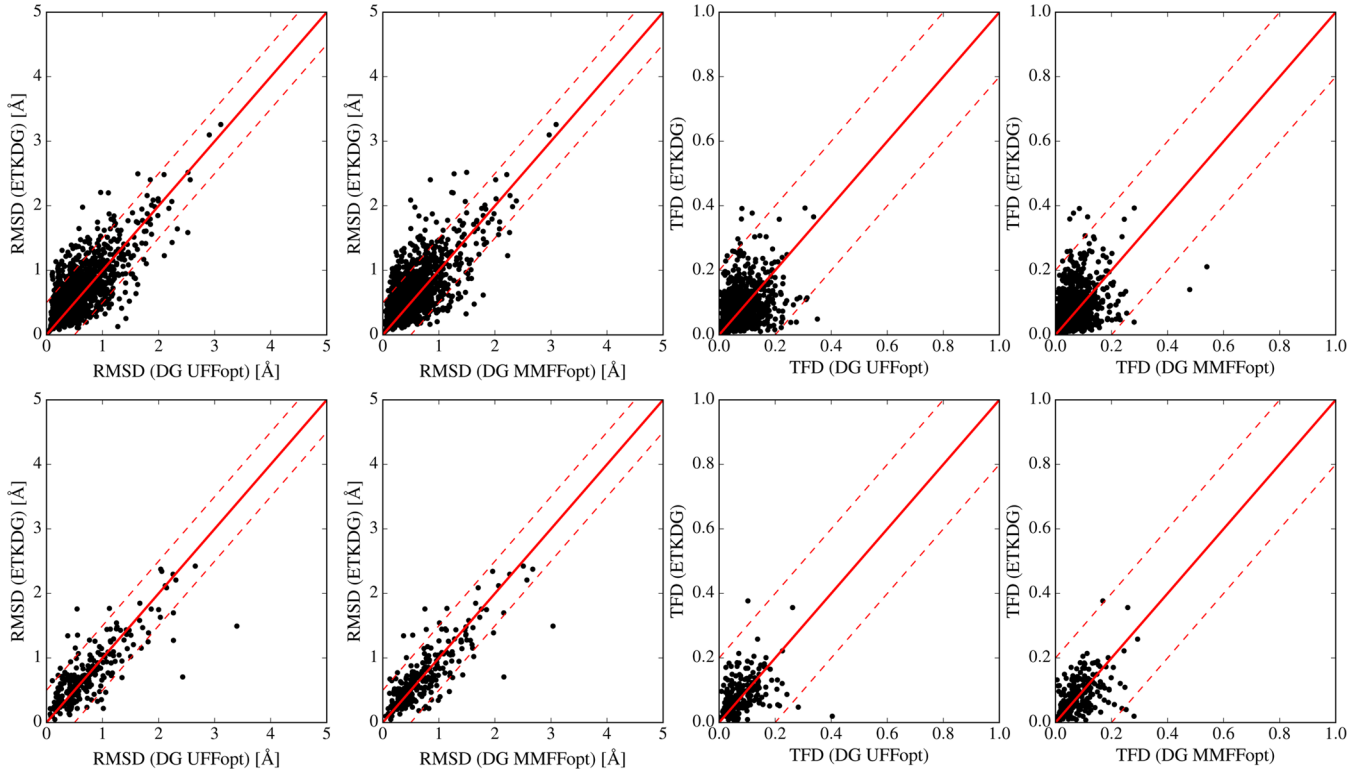


Figure 12. Comparison of the best RMSD (left two panels) and TFD (right two panels) values relative to the crystal structure for ETKDG and FF-minimized DG (UFF or MMFF) for $n = 100$: (top) CSD data set (1290 molecules); (bottom) PDB data set (238 molecules). In each panel, the red solid line represents $y = x$, and the red dashed lines indicate a difference of ΔRMSD = 0.5 or ΔTFD = 0.2.

DOI: 10.1021/acs.jcim.5b00654
J. Chem. Inf. Model. 2015, 55, 2562−2574

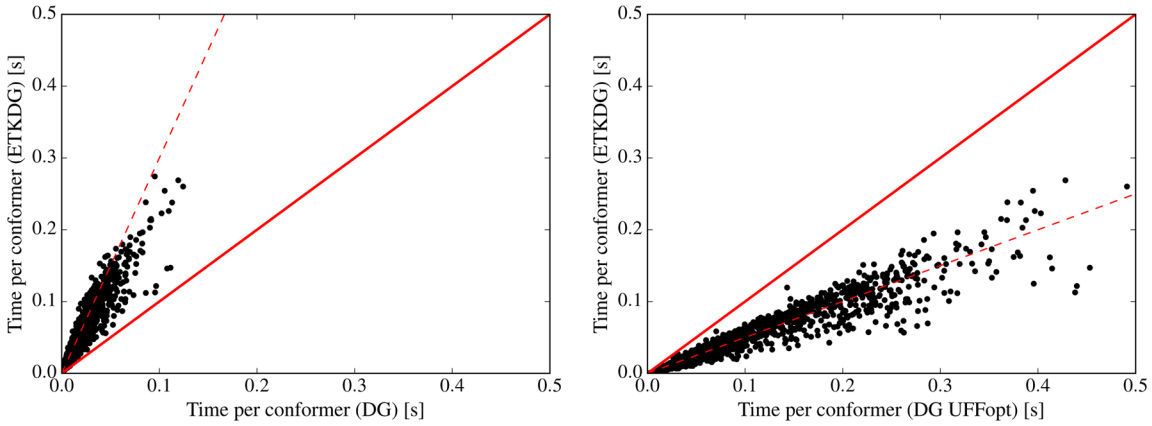


Figure 13. Comparison of the time (in s) per conformer for (left) DG and ETKDG and (right) DG + UFF optimization and ETKDG using the CSD data set. The red lines are to guide the eye. The red solid lines represent $y = x$, and the red dashed lines indicate (left) $y = 3x$ and (right) $y = 0.5x$.

for $n = 100$ are shown in Figure 12, and the results from the two-sided paired t tests are given in Table S3 in the Supporting Information. It should be noted that we use a two-sided test here because we are interested in whether the approaches perform the same (null hypothesis) or one of the two performs better (alternative hypothesis). While DG + FF-optimization is statistically significantly better than ETKDG for the CSD data set, the two approaches are comparable for the PDB data set. This could be due to the smaller size of the data set (i.e., 238 vs 1290 molecules), the different structure refinement methods used for the two different kinds of crystallography, or the inherent differences between small-molecule crystal structures and biologically active conformations from protein−ligand complexes. In order to resolve this question, more work will be necessary, including a comparison using a larger PDB-derived data set.

Timings. The additional minimization step required for ETKDG relative to DG does have an impact on runtime performance. This can be seen in Figure 13 (left). The median ratio of the ETKDG and DG runtimes is 3.0, i.e., ETKDG takes 3 times as long as DG. The addition of the K terms, i.e., generating ETKDG embeddings instead of ETDG embeddings, increases runtime by only 10% over ETDG (results not shown). Despite the longer runtime per conformer, ETKDG requires on average one-quarter of the number of conformers to achieve performance similar to DG (Figure S12 and Table S4 in the Supporting Information). This results in a net performance improvement, at least when it comes to reproducing crystal conformers.

As measured by performance in reproducing experimental crystal structures, ETKDG is a viable alternative to plain DG followed by a UFF-optimization, so it is of interest how their runtimes compare. Figure 13 (right) plots the runtime for ETKDG versus the runtime for DG + UFF-optimization. The median ratio of the DG + UFF optimization and ETKDG runtimes is 1.97, i.e., DG + UFF optimization takes almost twice as long as ETKDG. Thus, although ETKDG is significantly slower than DG on a per-conformer basis, when higher-quality conformations are required it can provide structures that are the equivalent of those obtained using DG + UFF-optimization in about half the time.

CONCLUSIONS

The 3D structures of flexible small organic molecules cannot be described by a single conformation but require a conformational ensemble that covers the accessible phase space. The quality of a conformer generator should therefore be measured both by the diversity of the generated ensembles and how well crystal conformations can be reproduced by these ensembles. Distance geometry (DG) is a computationally efficient approach that samples conformational space in a random manner. In contrast to knowledge-based conformer generators, DG uses a small amount of empirical information in the form of ideal bond lengths, ideal bond angles, and a few ideal torsional angles. Unfortunately, the nature of the distance terms used in the DG procedure tends to lead to distorted aromatic rings and sp^2 centers that cannot be fixed by the refining step without adding additional terms. Therefore, DG is typically used in combination with a minimization using a molecular force field, a step that adds computational complexity and run time.

In this study, we have presented an alternative, computationally more efficient approach termed ETKDG that uses empirical knowledge in the form of experimental torsional-angle preferences (ET) from the CSD together with a set of "basic knowledge" (K) terms to improve the generated conformers. ETKDG was found to outperform standard DG and the knowledge-based conformer generator CONFECT in reproducing crystal conformations from both small-molecule crystals (CSD data set) and protein−ligand complexes (PDB data set). With ETKDG, 84% of a set of 1290 small-molecule crystal structures from the CSD could be reproduced within an RMSD of 1.0 Å and 38% within an RMSD of 0.5 Å. The experimental torsional-angle preferences or the K terms alone each performed better than standard DG but were not sufficient to obtain the full performance of ETKDG.

Comparison of ETKDG with the DG conformers optimized using either the Universal Force Field (UFF) or the Merck Molecular Force Field (MMFF) showed different results for the two data sets. While FF-optimized DG performed better on the CSD data set, the two approaches were comparable for the PDB data set. In order to rule out effects from the smaller size of the PDB data set, the comparison should be repeated with a larger data set of biologically active conformations.

ETKDG is implemented in the open-source cheminformatics toolkit RDKit and is freely available. Future work will focus on extending the experimental torsional-angle preferences to

DOI: 10.1021/acs.jcim.5b00654
J. Chem. Inf. Model. 2015, 55, 2562−2574

nonaromatic rings, which are currently not covered by the patterns developed by Rarey and co-workers.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00654.

Figures S6−S12, Tables S1−S4, and captions of Figures S1−S5 (PDF)
Figure S1 (ZIP)
Figure S2 (ZIP)
Figure S3 (ZIP)
Figure S4 (ZIP)
Figure S5 (ZIP)
File list_torsion_patterns.txt containing the list of torsion SMARTS patterns taken from ref 25 (ZIP)
File list_PDB_entries.txt containing the list of PDB entries used in this study (ZIP)
File list_CSD_entries.txt containing the list of CSD entries used in this study (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: sriniker@ethz.ch.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Jean-Paul Ebejer for his help with the data sets and Matthias Rarey, Christin Schäfer, and Christian Lemmen for their help with CONFECT and useful discussions on the TFD implementation.

■ REFERENCES

(1) 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.
(2) Lyne, P. D. Structure-Based Virtual Screening: An Overview. *Drug Discovery Today* **2002**, *7*, 1047−1055.
(3) Venkatraman, V.; Perez-Nueno, V. I.; Mavridis, L.; Ritchie, D. W. Comprehensive Comparison of Ligand-Based Virtual Screening Tools Against the DUD Data Set Reveals Limitations of Current 3D Methods. *J. Chem. Inf. Model.* **2010**, *50*, 2079−2093.
(4) Hu, G.; Kuang, G.; Xiao, W.; Li, W.; Liu, G.; Tang, Y. Performance Evaluation of 2D Fingerprint and 3D Shape Similarity Methods in Virtual Screening. *J. Chem. Inf. Model.* **2012**, *52*, 1103−1113.
(5) Schwab, C. H. Conformations and 3D Pharmacophore Searching. *Drug Discovery Today: Technol.* **2010**, *7*, e245−e253.
(6) Verma, J.; Khedkar, V. M.; Coutinho, E. C. 3D-QSAR in Drug Design - a Review. *Curr. Top. Med. Chem.* **2010**, *10*, 95−115.
(7) Agrafiotis, D. K.; Gibbs, A. C.; Zhu, F.; Izrailev, S.; Martin, E. Conformational Sampling of Bioactive Molecules: A Comparative Study. *J. Chem. Inf. Model.* **2007**, *47*, 1067−1086.
(8) Ebejer, J.-P.; Morris, G. M.; Deane, C. M. Freely Available Conformer Generation Methods: How Good Are They? *J. Chem. Inf. Model.* **2012**, *52*, 1146−1158.
(9) Blaney, J. M.; Dixon, J. S. Distance Geometry in Molecular Modeling. *Rev. Comput. Chem.* **1994**, *5*, 299−335.
(10) Havel, T. F. Distance Geometry: Theory, Algorithms, and Chemical Applications; In *Encyclopedia of Computational Chemistry*; John Wiley & Sons: New York, 2002.
(11) The RDKit: Open-Source Cheminformatics Software, version 2015.03.1. http://www.rdkit.org (accessed November 2015).
(12) Vainio, M. J.; Johnson, M. S. Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Model.* **2007**, *47*, 2462−2474.
(13) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic Generation of 3D-Atomic Coordinates for Organic Molecules. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537−547.
(14) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B.; 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.
(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. Cheminf.* **2011**, *3*, 8.
(16) 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.
(17) 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.
(18) Li, J.; Ehlers, T.; Sutter, J.; Varma-O'Brien, S.; Kirchmair, J. CAESAR: A New Conformer Generation Algorithm Based on Recursive Buildup and Local Rotational Symmetry Consideration. *J. Chem. Inf. Model.* **2007**, *47*, 1923−1932.
(19) Leite, T. B.; Gomes, D.; Miteva, M.; Chomilier, J.; Villoutreix, B.; Tufféry, P. Frog: A FRee Online DruG 3D Conformation Generator. *Nucleic Acids Res.* **2007**, *35*, W568−W572.
(20) Miteva, M. A.; Guyon, F.; Tufféry, P. Frog2: Efficient 3D Conformation Ensemble Generator for Small Compounds. *Nucleic Acids Res.* **2010**, *38*, W622−W627.
(21) 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.
(22) Sadowski, P.; Baldi, P. Small-Molecule 3D Structure Prediction Using Open Crystallography Data. *J. Chem. Inf. Model.* **2013**, *53*, 3127−3130.
(23) Kothiwale, S.; Mendenhall, J. L.; Meiler, J. BCL::CONF: Small Molecule Conformational Sampling Using a Knowledge Based Rotamer Library. *J. Cheminf.* **2015**, *7*, 47.
(24) Crippen, G. M.; Smellie, A. S.; Richardson, W. W. Conformational Sampling by a General Linearized Embedding Algorithm. *J. Comput. Chem.* **1992**, *13*, 1262−1274.
(25) 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.
(26) Allen, F. H. The Cambridge Structural Database: A Quarter of a Million Crystal Structures and Rising. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 380−388.
(27) Groom, C. R.; Allen, F. H. The Cambridge Structural Database in Retrospect Andprospect. *Angew. Chem., Int. Ed.* **2014**, *53*, 662−671.
(28) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(29) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, High-Quality Test Set for the Validation of Protein-Ligand Docking Performance. *J. Med. Chem.* **2007**, *50*, 726−741.
(30) The pre-lease version was tagged as *ETKDG_Paper_Oct2015* in Github, DOI: 10.5281/zenodo.32402.
(31) 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.
(32) BioSolveIT. *TorsionAnalyzer*, version 2.0.0. http://www.biosolveit.de/TorsionAnalyzer/index.html (accessed November 2015).

DOI: 10.1021/acs.jcim.5b00654
J. Chem. Inf. Model. 2015, 55, 2562−2574

(33) 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.
(34) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(35) Schrödinger, LLC. The PyMOL Molecular Graphics System, version 1.7.4. http://pymol.org (accessed November 2015).
(36) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parametrisation, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.
(37) Halgren, T. A. Merck Molecular Force Field. II. MMFF94 Van Der Waals and Electrostatic Parameters for Intermolecular Interactions. *J. Comput. Chem.* **1996**, *17*, 520−552.

DOI: 10.1021/acs.jcim.5b00654
J. Chem. Inf. Model. 2015, 55, 2562−2574