Journal of Chemical Information and Modeling

Article

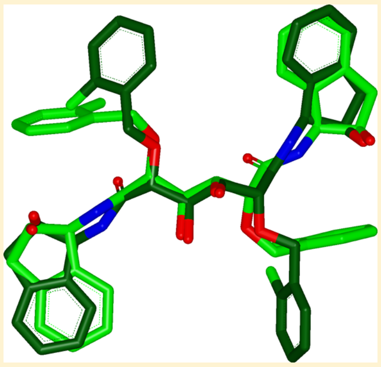
pubs.acs.org/jcim

Conformer Generation with OMEGA: Learning from the Data Set and the Analysis of Failures

Paul C. D. Hawkins*,[†] and Anthony Nicholls[†]

[†]OpenEye Scientific Software, 9 Bisbee Court, Suite D, Santa Fe, New Mexico 87508, United States

Ⓢ *Supporting Information*



ABSTRACT: We recently published a high quality validation set for testing conformer generators, consisting of structures from both the PDB and the CSD (Hawkins, P. C. D. et al. *J. Chem. Inf. Model.* **2010**, *50*, 572.), and tested the performance of our conformer generator, OMEGA, on these sets. In the present publication, we focus on understanding the suitability of those data sets for validation and identifying and learning from OMEGA's failures. We compare, for the first time we are aware of, the coverage of the applicable property spaces between the validation data sets we used and the parent compound sets to determine if our data sets adequately sample these property spaces. We also introduce the concept of torsion fingerprinting and compare this method of dissimilation to the more traditional graph-centric diversification methods we used in our previous publication. To improve our ability to programmatically identify cases where the crystallographic conformation is not well reproduced computationally, we introduce a new metric to compare conformations, RMSTanimoto. This new metric is used alongside those from our previous publication to efficiently identify reproduction failures. We find RMSTanimoto to be particularly effective in identifying failures for the smallest molecules in our data sets. Analysis of the nature of these failures, particularly those for the CSD, sheds further light on the issue of strain in crystallographic structures. Some of the residual failure cases not resolved by simple changes in OMEGA's defaults present significant challenges to conformer generation engines like OMEGA and are a source of new avenues to further improve their performance, while others illustrate the pitfalls of validating against crystallographic ligand conformations, particularly those from the PDB.

■ INTRODUCTION

Conformer generation has been a topic of considerable interest to the modeling community for a number of years, and a large number of methods for conformer generation have been presented.[1−8] In a recent publication,[9] we presented a validation of our conformer generator, OMEGA, on a set of carefully selected ligands from the PDB[10] and the CSD.[11] In selecting the ligands for the PDB-derived data set, we concentrated on identifying ligand structures that were well solved and well supported by the experimental data (the electron density), while the CSD set was selected by others as part of an effort to explore fragment conformational preferences.[12] The previous work therefore concentrated primarily on obtaining high quality sets of ligand structures against which to test OMEGA's conformation sampling algorithm and on presenting the aggregate performance of OMEGA on these sets. In the current analysis, we attend more closely to the individual performance of OMEGA on the particular molecules in the test sets and examine the utility of different metrics for identifying cases where OMEGA fails to acceptably reproduce a crystallographic structure. It is from this analysis of failures that we attempt to understand more fully the strengths and weaknesses of different aspects of the OMEGA algorithm with an eye to improving its performance in the future.

An unfortunate trait of publications in this area is for the authors to concentrate entirely on the successes they find in their results and to simply present any poor results, rather than trying to learn from them.[13] In many papers in this area, results are presented as an aggregate of the cases examined (mean RMSD over the data set being the most common measure), with outlying poor results being subsumed into the measure reported. Even in papers where the results are presented in their entirety,[14] no attention is paid to the failures. This is regrettable, as the cases in which a tool performs poorly or fails are precisely those cases from which the most can be learned.

There are a number of available methods for assessing the quality of pose reproduction (or finding failures) by conformer generators. In our previous publication, where we focused primarily on validating OMEGA's performance, we used two somewhat orthogonal measures (RMSD and TanimotoCombo). The well-known measure RMSD compares conformers atom by atom, while TanimotoCombo (TC) compares conformers by their overall shapes and chemical features in 3D.[9] For finding failures (or successes), no single measure is perfectly suited; to address some of the problems inherent in RMSD,[9,15] we sought another metric that would provide a measure of pose prediction that, like TC, is scaled across the

Received: July 6, 2012
Published: October 19, 2012

same range for all molecules but is based, like RMSD, on atom-to-atom alignments. The result of this search is a metric we term RMSTanimoto. The underlying basis for the calculation of RMSTanimoto and its application to failure identification is detailed in the Methods section. We find RMSTanimoto to be particularly well suited to identifying failures in small molecules (five or fewer rotatable bonds), thereby offering a very useful complement to RMSD.

In deriving the data sets used in this and the previous work, neither we nor any other authors in this area fully considered the relationship between the relevant properties of the validation set (a sample from the parent data set) and the properties of that parent data set as a whole (though for a start in this area see Kirchmair et al.[16]). For studies of this kind, relevant properties would include the types of flexible rings, the types of torsions, and the numbers of these groups present in the sample and parent sets. Therefore, we briefly examine some of the relevant properties of the selected validation sets with respect to their parent data sets and show that, for the most part, our validation sets do represent an adequate sampling of the parent sets. Having established this, we turn to an analysis of the success and failures in these sets.

The goal of failure analysis in this paper is not to tune OMEGA's parameters by exhaustively examining a large set of alternatives to determine which one produces the smallest number of failures. Rather, we wish to examine the default parameter set for simple (and hopefully sensible) changes that can resolve most or all of the initial failures. The nature of the adjustments to the default parameters are, in themselves, of use in understanding the nature of the problem that OMEGA is addressing, while the residual failures can be examined to understand why the experimental conformation proves difficult for OMEGA to reproduce. These failures, it is hoped, will provide some insights into the improvement of the OMEGA algorithm and/or its knowledge base.

As the OMEGA algorithm has already been outlined,[9] we will introduce only briefly the roles of the knowledge base (pregenerated information) in OMEGA's function. OMEGA utilizes two pregenerated inputs in its construction of conformers. First, 3D fragments of the input structure are identified in a prebuilt library of fragments (fragments not contained in the fragment library are generated on the fly). These 3D fragment conformations are generated using a distance geometry method against a modified version of the MMFF94 force-field.[17] In the subsequent step, torsion driving on the 3D structure assembled from these fragments is performed to generate a large initial "raw" conformer ensemble. This torsion driving is governed by a set of rules for torsion angles contained in what we term a torsion library. In the final stage, this raw conformer ensemble produced from torsion driving is pruned by geometric diversity and conformational strain energy (calculated using the same modification to MMFF94 used to calculate fragment geometries) to produce the output conformer ensemble. It is worth emphasizing here that at no point in this process are the conformers optimized against a force-field or any other function; they are enumerated, processed, and returned to the user "as is".

The torsion library, derived mostly from inspection of experimental crystal structures from the PDB, is knowledge-based, while the fragment library ensemble can be considered to be generated from first principles (in that it does not rely on direct experimental knowledge of fragment conformations). Our failure analysis will therefore address three questions in OMEGA's design:

1. Are OMEGA's defaults satisfactory for conformer generation for small molecules?

2. Does the OMEGA knowledge base (the torsion library and the fragment library) adequately cover the conformational space required for reasonable reproduction of crystallographic structures?

3. Is the MMFF94 force-field suitable to describe small molecule conformational energetics?

■ METHODS

All code was written in Python (version 2.5), and statistical calculations were performed with scipy version 0.1. An example Python script illustrating the computation of confidence intervals by bootstrapping is included in the Supporting Information. Cheminformatics functions (including RMSD, RMSTanimoto, property, and similarity calculations) were performed using OpenEye's OEChem (version 1.7.3) and GraphSim toolkits (version 1.1), and shape calculations were performed using the OpenEye Shape toolkit (version 1.7).[18] Protein−ligand cocrystal structures were downloaded from the PDB and filtered according to the criteria laid out previously.[9] Methods for identifying suitable ligand structures from the PDB have already been presented in some detail.[9] These methods have also been applied to an in-house data set of high quality protein structures,[19] and a number of new ligands from this set were added to the set of 197 used in the original publication. This new set was checked for diversity at the torsion and graph levels (*vide infra*). If a pair of ligands were found to be too similar at either level, the one with the poorer model fit (as judged by RSCC, RSR, and occupancy-weighted B-factor) was removed. After additions and deduplication, the PDB data set finally comprised 200 ligands. A set of 481 CSD structures, derived from 483 used in our previous publication, was used in this study (two molecules were removed due to high similarity in their torsions). PDB and CSD codes for these molecules are given in the Supporting Information.

Molecules were converted into isomeric SMILES format using the OEChem toolkit before being used for conformer generation and property calculation. Conformer databases were generated using OMEGA version 2.3, with an energy window for acceptable conformers (*ewindow*) of 10 kcal/mol above the ground state using the modified version of the MMFF94 force-field mentioned above, a maximum number of conformations per molecule (*maxconfs*) of 200 and an RMSD cutoff (*rmsd*) of 0.5Å (the default settings in OMEGA 2.3). A so-called complete conformer ensemble under a given energy window is produced by setting the rmsd and *maxconfs* parameters to 0, so that OMEGA does not perform its geometric diversification step nor limit the size of the ensemble produced. In the following discussion, it should be remembered that OMEGA outputs conformers in energy-sorted order—that is, the lowest energy conformers are added to the output set until the *maxconfs* limit is reached or no more conformers are available in the raw ensemble. Thus, increasing *maxconfs* can allow exploration of higher energy conformational space under the limit set by *ewindow*.

To generate druglike subsets of the PDB and CSD, the following filters were applied: Molecular weight 150−800; Heavy atom count 8−70; Donor count 0−6; Acceptor count 0−9; logP −5.0−6.0; only the elements H, C, N, O, F, Cl, Br, I, S, and P allowed. Optimization of experimental ligand structures was carried out against the MMFF94 force-field as implemented in OpenEye's SZYBKI toolkit (version 1.5), allowing only torsions to change (thereby avoiding inclusion of energy changes resulting from relaxation of bond lengths and angles in the experimental structure).

Quantum mechanical torsion scanning calculations were carried out at the MP-2/6-31G** level using GAMESS.[20]

Suitable Metrics for Comparing Conformers. Two metrics for comparing conformer ensembles to an experimental conformation were used in the previous publication: a geometric measure (RMSD) and a shape- and chemical feature-matching metric (TC). In this study we add a third metric, which we term RMSTanimoto. RMSTanimoto is computed by comparing atom positions, as does RMSD; but, like TC, RMSTanimoto is scaled over a defined range for molecules of all sizes. This offers a potential advantage over RMSD, which has no general upper bound, so the same RMSD has different significance for molecules of different sizes. To generate an RMSTanimoto in this study, the two conformations to be compared are first overlain by minimizing their heavy atom RMSD, and then RMSTanimoto is computed for this overlay. In this way, RMSTanimoto can be considered a method to rescore an RMSD-optimized alignment. In the following paragraphs we outline the calculation of RMSTanimoto.

RMSTanimoto. Algorithms that calculate the best RMSD between two molecules are specific cases of more general numerical procedures for calculating the minimal sum of the pairwise square distances between two ordered sets of ordered atoms. For example

Given N vectors, $\vec{x}$ and $\vec{y}$, min imize:

$$\sum_{i=1}^{N} (\mathbf{R}\vec{x}_i + \vec{D} - \vec{y}_i)^2$$

where **R** is a rotation matrix, and *D* is a displacement vector. The displacement vector turns out to be simply that which superimposes the geometric centers of the two sets of points, so the problem is actually to minimize

$$\sum_{i=1}^{N} (\mathbf{R}\vec{x}_i - \vec{y}_i)^2$$

Expanding, we get

$$\sum_{i=1}^{N} (\vec{x}_i^{\,2} + \vec{y}_i^{\,2} - 2\vec{y}_i\mathbf{R}\vec{x}_i)$$

Since only the last term in the bracket depends on the rotation matrix, the procedure is equivalent to maximizing

$$\sum_{i=1}^{N} \vec{y}_i\mathbf{R}\vec{x}_i = \sum_{i=1}^{N} \vec{y}_i\vec{x}'_i$$

where the prime denotes rotated coordinates. As such, there is a similarity with the maximization of overlap in shape (as used in the generation of the TC metric), wherein the goal is to find the rotation matrix (and displacement vector) that maximizes the overlap (inner product) of two volumes. By analogy, it is interesting to consider the quantity

$$T_{RMSD} = \frac{\sum_{i=1}^{N} \vec{y}_i\vec{x}'_i}{\sum_{i=1}^{N} \vec{y}_i\vec{y}_i + \sum_{i=1}^{N} \vec{x}_i\vec{x}_i - \sum_{i=1}^{N} \vec{y}_i\vec{x}'_i}$$

This quantity, which we term the RMSTanimoto, has similar properties to a shape or chemical similarity Tanimoto. When the molecular coordinates are aligned well, the inner product numerator is similar in magnitude to the first two terms in the denominator—that is, the Tanimoto approaches unity. When a pair of conformers are poorly aligned, the RMSTanimoto will be small. Note that this general form for RMSTanimoto can also be negative, with a possible minimum value of −1/3. This is unlikely when the molecules are optimally aligned, but possible if the comparison is between arbitrarily aligned molecules (for example, when comparing results of pose prediction by docking to the crystallographic coordinates of the ligand).

The considerations that led us to develop RMSTanimoto are very similar to those expressed by the developers of the GARD metric.[21] In the case of RMSTanimoto, however, no weighting scheme for different atom types is used, nor is an empirical function used to penalize atom-to-atom deviations. RMSTanimoto is therefore an unbiased or naïve method of computing a bounded measure of geometric deviation between, in the cases examined here, conformers of the same molecule. Obviously, RMSTanimoto can also be used to rescore molecular overlays or alignments generated by any means, including other ligand-based methods and docking.

■ RESULTS AND DISCUSSION

Understanding the Metrics. Since RMSTanimoto in effect rescores a pose pregenerated by optimizing the RMSD between two conformations, it is of interest to compare these two methods of scoring the same overlay. In Figure 1, we compare the lowest RMSD of any conformer in the OMEGA ensemble versus the experimental conformation to the RMSTanimoto computed from that same lowest RMSD overlay for the 200 ligands in the PDB-derived set. As expected, in the cases where the RMSD is low (<0.25 Å), the RMSTanimoto is always high (close to 1.0). As RMSD increases, the relationship to RMSTanimoto becomes increasingly less linear, in that there are several examples where constant RMSD gives widely varying RMSTanimoto and vice versa. The two metrics clearly assess the same overlay in different ways. Note that in this data set the RMSTanimotos are always greater than 0, even though this is possible given the formalism used to calculate it. This is because the conformer to conformer alignment is first optimized using RMSD and then rescored with RMSTanimoto, so the conformers are aligned in a close to optimal manner before the RMSTanimoto is

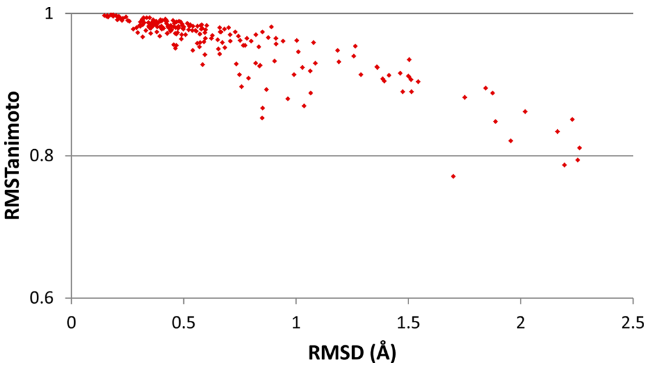


Figure 1. Relationship between RMSD and RMSTanimoto for reproduction of the 200 ligand structures from the PDB.

calculated. The significant advantage that RMSTanimoto provides over RMSD is that it is bounded by −1/3 and 1 for molecules of all sizes. This size-independence potentially allows RMSTanimoto to identify poor reproductions of conformations of molecules with few heavy atoms that pass a RMSD cutoff (vide infra). Another approach to adjusting for the size dependence of RMSD is to normalize it by the number of heavy atoms in the molecule. Plots illustrating the relationship of this ratio to RMSTanimoto, along with one illustrating the relationship between RMSTanimoto and heavy atom count, for the PDB ligands are contained in the Supporting Information. We find that even dividing RMSD by the heavy atom count of the molecule does not correct for RMSD's inability to find failures in this set of molecules (vide infra).

In our previous publication, we compared overlays optimized by RMSD to those optimized by TC. The difference in the two overlays is shown in Figure 2 for the ligand from the 1V2N PDB structure (ligand code BBA).

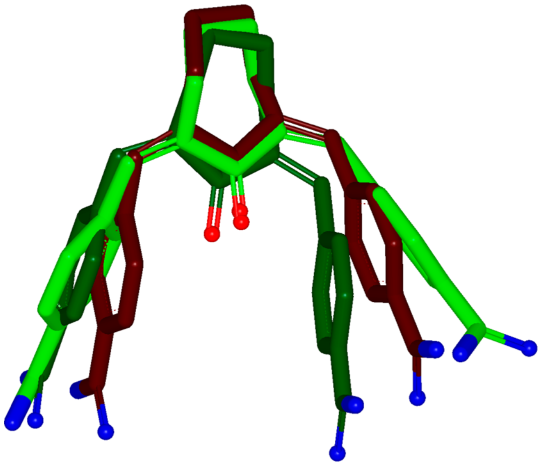


Figure 2. Comparison of the experimental conformation of the ligand BBA (light green) with the best overlay as provided by RMSD (red, RMSD = 1.06 Å) and the best overlay as provided by TC (dark green, TC = 0.89).

As can easily be seen in Figure 2, the RMSD overlay symmetrically (but poorly) matches both benzamidine functional groups, while the TC overlay matches one benzamidine group very well but not the other (giving the low TC). This illustrates a consistent difference between the two methods of comparison: RMSD tends to produce overlays that are equally good (or bad) everywhere across the molecule, while TC tends to produce overlays that are very close in some portions but can be much more divergent in others. This property of TC-driven overlays frequently makes them very straightforward to interpret, as they highlight the positions of maximum divergence between the two conformers. The RMSD optimized overlays, however, tend to minimize any divergence between the two.

Analysis of the Data Sets. In our previous paper we focused on identifying good quality crystal structures against which to validate OMEGA, assuming that the molecules we selected would be good representatives of the parent databases from which they were drawn. We did not, however, explicitly consider the relationship between the properties of our validation sets and the properties of the parent data sets. As such, the validation sets we used might inadvertently cover only a small fraction of the relevant properties in the parent data sets and therefore would not be appropriate guides to OMEGA's performance on other molecules drawn from the same parent data sets. In the context of this experiment, the parent data sets are not all ligands in the PDB and all small molecules in the CSD but rather only the druglike molecules or ligands (however loosely druglike is construed) present in these collections, since these criteria were applied in selecting the molecules in the validation sets. The criteria for druglikeness in this context are given in the Methods section and were applied to ligands downloaded from LigandExpo[22] and all molecules in the CSD (version 5.1). Duplicate molecules were removed from the compounds surviving the filters to give unique, druglike sets of molecules that are the appropriate parent sets from which our validation sets can be considered to have been drawn.

The properties of the data sets that we considered the most relevant to this study were those governing the available conformational space for a given molecule: the torsions and flexible rings present in the molecule. The method we chose to evaluate the suitability of our validation sets was to find all the unique flexible rings and all the unique torsions in the parent data sets. Our definition of flexible rings includes substituents, so the same core ring structure with different substituents is treated as a different ring. We then determine what fraction of those rings and torsions were present in our validation sets. This provides an estimation of the coverage of the relevant property spaces. The results for the torsions present in the druglike subsets of the PDB and the CSD with the corresponding validation set are shown in Table 1, while the results for flexible rings are in Table 2.

Table 1. Coverage of Unique Torsions between the Validation Sets and the Parent, Druglike Data Sets

	PDB	PDB_subset	CSD	CSD_subset
molecules	7722	200	70303	481
torsions	3829	814	9534	1082
coverage	100%	21.3%	100%	11.3%
enrichment	_	8.2	_	16.6

Table 2. Coverage of Unique Flexible Rings between the Validation Sets and the Parent, Druglike Data Sets

	PDB	PDB_subset	CSD	CSD_subset
molecules	7722	200	70303	481
flexible rings	547	46	10741	74
coverage	100%	8.4%	100%	0.7%
enrichment	_	3.2	_	1.0

The entry for enrichment in torsions found in Table 1 is calculated from the fraction of torsions found in both the parent data set and the corresponding validation set and the fraction of the parent data set that is the validation set (similarly for the flexible rings in Table 2). Both the validation sets—particularly the one from the CSD—are quite "enriched" in torsions found in the parent data set over the level expected from random selection of the validation sets. For example, our PDB-derived set contains 200 molecules (from a possible 7,722 from LigandExpo) that have 814 unique torsions (from a possible 3,829 unique torsions in the LigandExpo set), giving a torsion coverage of 21.3%. By way of comparison, the PDB-derived Vernalis set[23] contains 250 druglike molecules having

692 unique torsions (from a possible 3,829), giving a torsion coverage of 18.1%. By carefully selecting our ligands, we cover a greater fraction of the available torsion space (21.3% v. 18.1%) with fewer molecules (200 v. 250), leading to an “enrichment” of 8.2-fold for our data set, compared to 5.6-fold for the Vernalis set. The results for the flexible ring coverage, however, are much less encouraging; in particular, the CSD data set shows only random “enrichment” for flexible rings and the PDB set a slight 3.1-fold enrichment. We conclude that, as far as the sampling of torsion angles is concerned, the data sets we use are reasonably good, as they contain a pleasingly large proportion of the available torsions in small fractions of the size of the parent database (particularly so for the CSD). As such, OMEGA’s performance on a variety of torsion angles (the adequacy of the torsion library) is reasonably well tested by the data sets used here. The situation is less good with flexible rings, in that the data sets used do not explore more than a small fraction of the flexible rings available in the parent databases (particularly so for the CSD).

Data Set Diversity in Torsion and Graph Space. In the development of both the data sets used in this paper, a graph-based diversification step was used to ensure that sets of simple analogues were not contained in the final data set. In the original paper, we mentioned[9] that this use of graph-based methods is a surrogate for the desired property, which is diversity in the torsions and flexible rings in the data set. To examine the utility of the concept of “flexibility fingerprinting” for testing our data sets, we implemented a simple torsion fingerprinting method and compared the similarities from this method to those from traditional graph-based methods. The goal of this analysis is to determine whether there are molecules in our validation sets that are diverse in their graphs but possess essentially the same torsions—i.e., they have high torsional similarity but lower graph similarity. These molecules would be considered as possibly redundant and therefore candidates for removal from the data sets. Our method is akin to the MACCS keys approach,[24] but in torsion fingerprinting the bits of the fingerprint are set based upon the presence in the molecule of a torsion found in the OMEGA torsion library. There are 143 SMARTS patterns in the OMEGA torsion library, resulting in a fingerprint 143 bits long. From this fingerprint we can calculate a Tanimoto coefficient between any pair of molecules. Figure 3 shows the relationship between Tanimoto coefficients for two graph-based methods (path-based fingerprints, upper plot) or MACCS keys (lower plot) and the Tanimoto coefficients for the torsion fingerprinting method for all pairwise comparisons between the 200 PDB ligands.

The comparison between the MACCS key similarities and the torsion fingerprint similarities is particularly instructive, as these two methods are very similar in concept. It is clear that in general the two methods find rather different levels of similarity between the same molecules. It is also clear that, unsurprisingly, the torsion fingerprint is quite sparse (with a number of molecules having a pairwise similarity of 0.0 using torsion fingerprints, while no molecules have a pairwise similarity of 0.0 using MACCS keys).

Pairs of molecules with low graph similarity but high torsion similarity were visually inspected to determine if they offered significant enough diversity to be retained in the sets. Two pairs of molecules were identified as having identical torsions: P19 (from PDB structure 1D4L) and HBH (from PDB structure 1Z1R) and RDL (1NQU) and HDF (1QNF) (see Figure 4). These pairs of molecules have a torsion Tanimoto of 1.0, while their graphs are rather different (path Tanimotos P19:HBH = 0.70, RDL: HDF = 0.19). Close inspection showed that the torsion fingerprinting accurately designated these molecules as redundantly similar. Accordingly, one of each pair was removed from the final set, based on the criteria given in the Methods section. A similar analysis was conducted for the CSD set; again, two molecules were removed, giving a final data set of 481 molecules (data not shown). In this way we hoped to maximize the information content per molecule in our validation set and prevent a small set of torsions from dominating the data set, thereby skewing our estimations of OMEGA’s performance.

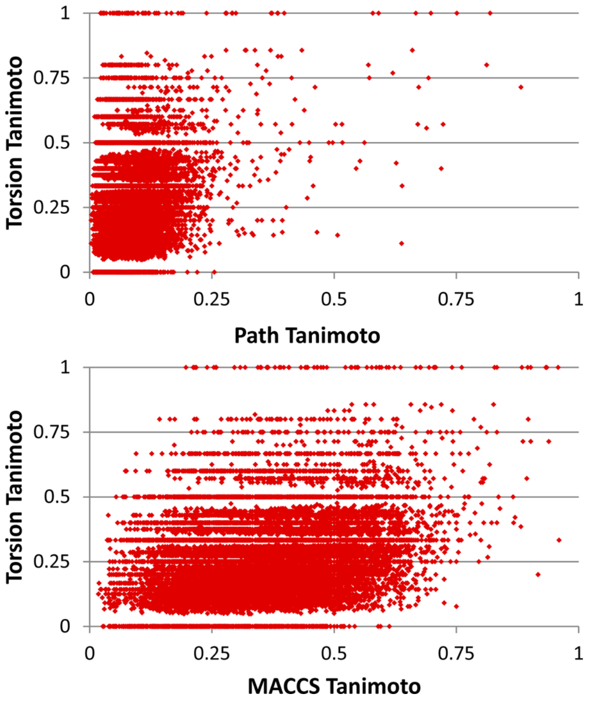


Figure 3. Relationship between path-based similarity and torsion-based similarity (upper) and MACCS key similarity and torsion-based similarity (lower).

From these data sets of structurally diverse molecules (in both torsion and fingerprint space) that are also good models of their experimental electron density, we generated conformers for each of the molecules using OMEGA’s defaults. These conformer ensembles were compared to the crystallographic conformation using the three metrics discussed above (RMSD, RMSTanimoto, and TC). Using these three metrics, we identified conformations that were not well reproduced by OMEGA and investigated why the reproduction may have been poor.

Analysis of Failures. In the following analysis, we focus on those cases where OMEGA does not produce a conformer sufficiently close to the crystallographic structure. We identify these failures using the following cutoffs for the three different metrics:

- RMSD > 2.0Å
- TC < 1.0
- RMSTanimoto < 0.8

There is, of course, a degree of arbitrariness in choosing the magnitude of a cutoff. The 2Å cutoff for RMSD has been widely used as a criterion of success in docking studies and conformer generation and is thus blessed by both usage and history. The cutoffs for the two other metrics have been set for

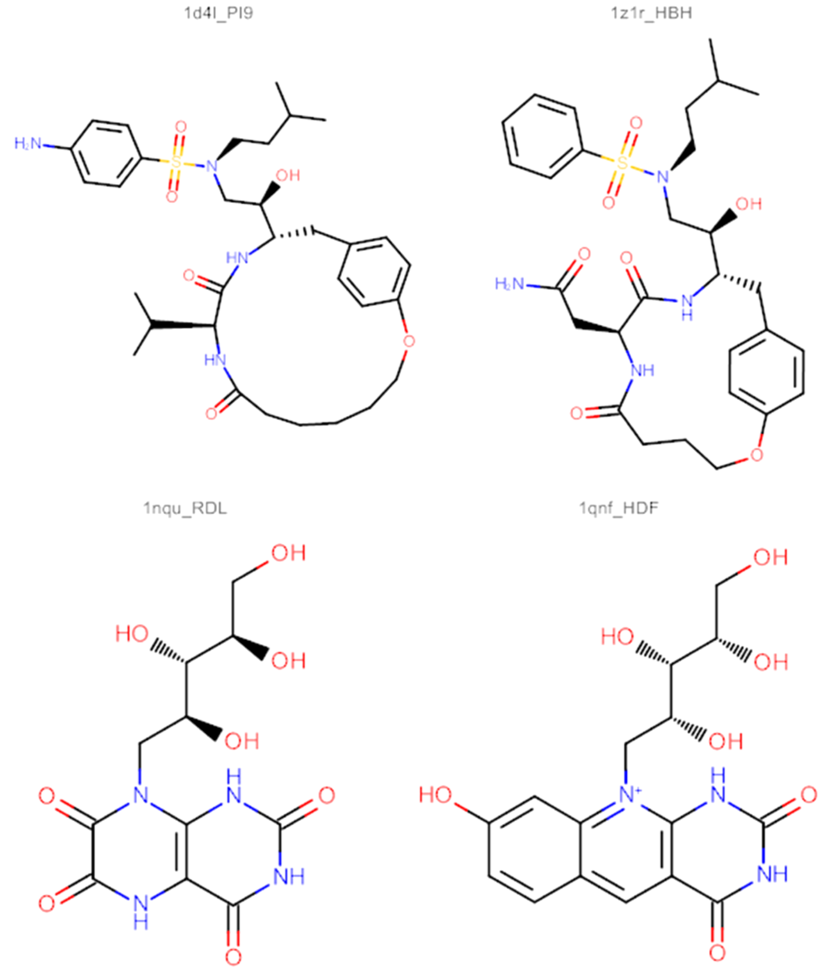


Figure 4. Two pairs of molecules identified as being too similar in the torsions they carry.

this study based on visual inspection of overlays; in general, overlays that fail these cutoffs have one or more poorly matching features or a poor match in shape. Our main motivation for using multiple numerical metrics of failure is to allow automated identification of all the cases that might require closer inspection. Simple numerical criteria for success/failure can never substitute for visual analysis, but using several independent numerical criteria can ensure that all the cases with possible problems are highlighted for more careful attention. In addition, although visual inspection of the alignments will eventually find all failures, for hundreds of alignments this procedure is time-consuming, quite possibly arbitrary, and almost certainly order-dependent. Metrics of quality for overlays based on judgment and visual inspection alone may have undefined distributions, making unbiased and statistically reliable comparisons difficult or meaningless. As such, a combination of numerical failure identification (using metrics that are unbiased and permit statistical analysis) and visual inspection and interpretation offers the best compromise between a purely automated analysis and a fully subjective approach.

Having gathered a set of failures, they now may be analyzed to understand why the given conformation was not reproduced successfully. There are a number of possible reasons why OMEGA might fail on a given molecule:

1. Insufficient sampling of the low-energy conformational space. This is particularly likely for molecules with a large number of rotatable bonds/flexible rings, since their accessible conformational space is very large, even at the low strain energy levels that OMEGA is designed to sample.

2. The experimental structure is high in strain energy as judged by the MMFF94 force-field. The goal of OMEGA is to produce low strain energy conformers. Therefore, high-energy conformers will not, by design, be among those output from OMEGA. The ligand structures in this set were carefully chosen to be well fit to their electron density, so an experimental conformation with genuinely high strain was not considered likely at the outset.

3. The knowledge base used by OMEGA (its torsion library and database of 3D fragment geometries) is incomplete. If an experimental torsion or ring geometry differs substantially from those in the OMEGA knowledge base, then OMEGA may be unable to match the experimental structure closely.

4. A deficiency or error in the implementation of the OMEGA algorithm.

Our main focus in this paper is to determine if deficiencies in the OMEGA implementation, knowledge base or the MMFF94 force-field cause failures, or if the failures are more easily resolved by changes to the default parameters. In those cases where a failure is due to a deficiency in some aspect of OMEGA's function, we can use this knowledge to improve OMEGA in future versions.

Failures among PDB Ligands. The 200 ligands from the PDB set were submitted to OMEGA for conformer generation, and the resulting ensembles were compared to the experimental conformation. The aggregate results are shown in Table 3, and the complete results are given in Figure 5.

Table 3. Mean, Standard Deviation (StdDev), and Median for TC, RMSD, and RMSTanimoto for 200 Ligands from the PDB

	TC	RMSD (Å)	RMSTanimoto
mean	1.55	0.66	0.96
StdDev	0.29	0.44	0.04
median	1.64	0.50	0.98

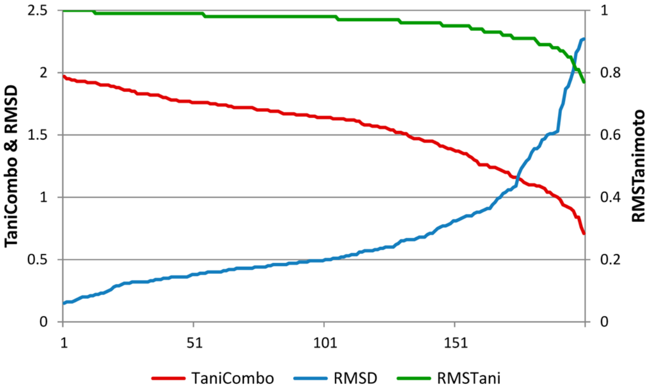


Figure 5. Distribution of TanimotoCombo (red) and RMSD (blue) (vertical scale at left), RMSTanimoto (green, vertical scale at right) between the closest conformer from OMEGA and the experimental conformation for 200 ligands from the PDB.

In an effort to take appropriate account of experimental error or variance in calculating aggregate statistics for performance, we use here—for the first time in this area to our knowledge—the variance weighted mean. The goal of variance weighting is to give more weight in the calculation of aggregate statistics to prediction of those measurements (here ligand models) with the lowest experimental error. In the case of the PDB ligands, a measure of the variance of the experimental data (ligand coordinates) is available through the diffraction coordinate precision index (DPI) as calculated by Blow.[25] The equation expressing the variance weighted mean, $\overline{m}$, is

$$\overline{m} = \sum_{i=1}^{n} w_i y_i \div \sum_{i=1}^{n} w_i$$

where the weight, w, is the coordinate error for the model, sqrt(3) × DPI, and y is the lowest RMSD for each ligand.

The weighted and unweighted means and their 95% confidence intervals using RMSD for this set are shown in Table 4 (confidence intervals were calculated using bootstrapping with replacement). It is more difficult to determine the appropriate level of experimental variance for use with RMSTanimoto or TC, so we do not compute a variance-weighted mean for these two measures. That the weighted mean for RMSD is significantly lower than the unweighted mean (the 95% confidence intervals do not overlap) implies that OMEGA performs better when reproducing structures with lower inherent experimental error, exactly as we would hope.

Table 4. Unweighted (UnWt) and Variance Weighted (Wt) Means and Their Respective 95% Confidence Intervals for Reproduction of the PDB-Derived Set Measured by RMSD

	UnWtMean	95%CI UnWt	WtMean	95%CI WtMean
RMSD	0.66	0.604, 0.728	0.51	0.457, 0.574

With aggregate statistics of performance in hand, we then turned to identification of the individual failures. Using the three cutoffs for successful reproduction of the experimental conformation detailed earlier, we found 13 failures in the PDB set (see Table 5). In particular, RMSD identified four failures (<2 Å), RMSTanimoto found three (<0.8), one of which passes the RMSD cutoff, and TC found eleven failures (<1.0), seven of which pass the RMSD cutoff (noting that the optimal TC overlay is often derived from a different conformer than the optimal RMSD overlay). The complementarity of these three metrics is already plain. The ligands that failed on one or more of these criteria are shown in Table 5, along with the number of rotatable bonds in the molecule and the DPI of the structure. Unsurprisingly, the great majority of these molecules are highly flexible (11/13 have nine or more rotatable bonds, though only two have flexible rings). The low-energy conformational space available to these very flexible molecules is quite large (many thousands of conformations), so it is no surprise that OMEGA had difficulty in generating a conformer close to the crystallographic, given the low level of sampling that it is permitted by its default parameters (a maximum of 200 conformers). By inspection there is no obvious relationship between the molecule being a failure and the DPI of the structure from which it was obtained.

Table 5. PDB code Ligand code TC RMSD (Å) RMST Rotors DPI (Å)[a]

PDB code	ligand code	TC	RMSD (Å)	RMST	rotors	DPI (Å)
1b6k	PI5	**0.82**	1.5	0.912	13	0.168
1b6m	PI6	**0.97**	1.19	0.948	12	0.152
1cvu	ACD	**0.94**	1.48	0.89	14	0.176
1d8d	FII	**0.91**	1.75	0.882	12	0.133
1ec0	BED	**0.87**	**2.23**	0.851	15	0.128
1fm9	570	**0.99**	1.54	0.904	12	0.335
1g2k	NM1	**0.95**	**2.25**	**0.794**	12	0.192
1s63	778	**1.22**	1.70	**0.771**	5	0.086
1t32	OHH	**0.74**	**2.16**	0.834	9	0.203
1v2n	BBA	**0.89**	1.06	0.919	4	0.130
1y6b	AAX	**1.26**	**2.20**	**0.787**	9	0.213
1zz2	B11	**0.95**	1.41	0.913	8	0.171
2i0a	MUI	**0.76**	1.88	0.888	15	0.141

[a] Failures are in bold. Rotatable bond count for the ligand and DPI for the structural model are also shown.

Given the high flexibility of many of the molecules under examination, an obvious solution to finding a conformation close to the crystallographic is to allow more conformers to be generated for these molecules (a more thorough sampling of the available low-energy conformational space). The default maximum number of conformations generated in OMEGA is 200 (set by the *maxconfs* parameter), irrespective of the flexibility of the molecule. At a *maxconfs* of 200, a molecule with five rotatable bonds has around 2.9 torsions per rotor—a reasonable, though not overly thorough, sampling of the conformational space available to each torsion. A molecule with eight rotatable bonds is, however, allowed only around 1.9 conformations per rotor at that same setting of *maxconfs*, quite a low sampling per rotor. To allow flexible molecules a more similar degree of sampling, *maxconfs* must be increased above the default of 200 while being mindful of the size of the conformational ensemble produced (e.g., a molecule with eight rotors will have an ensemble of 5,000 conformations at 2.9 conformations per rotor). After some preliminary experiments, we elected to allow more conformations to be generated for the more flexible molecules by increasing the *maxconfs* to 800 but only for molecules with eight or more rotatable bonds (there are 62 ligands in this set with eight or more rotors). This new setting for *maxconfs* translates to around 2.3 conformations per rotor in a molecule with eight rotors (such as ligand B11) and around 1.5 conformations per rotor in a molecule with 15 rotors (such as ligand MUI). Note that this is still a proportionately lower sampling per rotor than for the less flexible molecules. New conformational ensembles for the 200 ligands were generated with this single parameter change. Clearly the cost of this increased conformer sampling for more flexible molecules is the increased size of the conformer ensembles produced. The mean conformer count per molecule rises from 119 with the defaults to 270 with the new parameters, while the median is, unsurprisingly, not changed at 159. Since these new ensembles are either the same as, or are supersets of, the default ensembles, in no case was a ligand structure reproduced less accurately by the new ensemble than the default. The aggregate statistics for RMSD (mean and variance weighted mean) from using the defaults and the new parameter set are shown in Figure 6. As can easily be seen by comparing the 95% confidence interval for the mean RMSD (computed by bootstrapping), the new parameter set produces, on aggregate, only slightly better overall results than the defaults (while the failure rate is halved, vide infra). Similar trends are seen for TC and RMSTanimoto (data not shown).

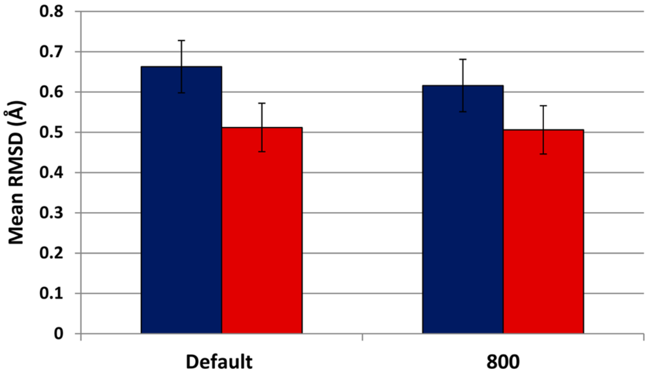


Figure 6. Comparison of mean RMSD and the 95% confidence intervals for the mean for reproduction of 200 PDB ligands using default and maxconfs = 800 parameter sets in OMEGA. Blue bars are for the unweighted mean, red bars for the variance weighted mean.

The effect of the change in *maxconfs* is most clearly seen when the list of failures is inspected (see Table 6). As we had hoped, this simple change in the defaults eliminates a large fraction of the failures. With the new parameters, there were only seven failures, down from 13 when using the defaults.

Most of the ligands that were not reproduced satisfactorily with the default settings but passed our criteria with the single change in *maxconfs* were ligands with high rotatable bond counts, indicating that a simple increase in the amount of low-energy sampling for these molecules was an efficient solution. The residual failures are molecules with widely varying flexibility (see rotor counts in Table 6 and depictions in Figure 7), so other reasons for failure than simply a low sampling of a large conformational space are likely to be at work. Since these molecules proved challenging for OMEGA, we searched the PDB to find all other instances of the same ligands (crystallized with the same or a different protein) to compare OMEGA's performance on these to the original. This served two purposes: to determine if we had selected a particularly difficult instance of the ligand to reproduce and to investigate more thoroughly the conformational properties of these difficult ligands. The frequency of occurrence of these ligands varied widely; three ligands occur only once in the PDB (PI5, BED, and AAX), while FII occurs in 10 instances, in nine PDB structures (FII occurs twice in the 1D8D structure).

Table 6. RMSD, TC, and RMSTanimoto for the Seven Ligands from the PDB Data Set That Cannot Be Successfully Reproduced As Judged by One or More of These Metrics Using Modified OMEGA Parameters[a]

PDB	ligand code	TC	RMSD (Å)	RMSTanimoto	rotors
1b6k	PI5	**0.92**	1.27	0.94	13
1d8d	FII	**0.93**	1.49	0.92	12
1ec0	BED	1.08	**2.02**	0.87	15
1s63	778	1.22	1.7	0.77	5
1t32	OHH	**0.74**	**2.02**	0.87	9
1v2n	BBA	**0.89**	1.06	0.919	4
1y6b	AAX	1.26	**2.14**	0.81	9

[a] Failures are in bold.

The experimental conformational diversity exhibited by some of the failure ligands is quite large (see Table 7). For example, the ligand OHH, from the 1T32 structure, occurs in only one other instance in the PDB, as the ligand for the 1T31 structure. In spite of only appearing once in each structure, the differences between these two instances of OHH are significant (the RMSD is more than 3 Å). The geometric differences among the 10 instances of FII, on the other hand, are quite small (the maximum pairwise RMSD is only 1.24 Å).

With some alternate cases in hand, we undertook a close analysis of the reasons for OMEGA's failures. In some cases, a failure is easily explained as a feature of the metric used to judge reproduction quality. The failure of reproduction of the 1D8D ligand, PI5, as judged by TC, arises from the properties of TC as an alignment mechanism, as was shown for the 1V2N ligand BBA *vide supra* (the RMSD is quite satisfactory, especially for a molecule of this size). The optimal overlays from TC and RMSD are compared in Figure 8. The RMSD overlay is good overall; the low-scoring TC overlay aligns the phosphate groups perfectly and the hydroxamate carbonyl well, but the conformation of the hydrocarbon side chain matches poorly. Since the conformational diversity of the experimental instances of the FII ligand is low, it was not surprising that OMEGA performed very similarly on all the other nine instances of FII (data not shown). OMEGA succeeded in all cases to reproduce the experimental instance by RMSD and RMSTanimoto, yet failed to reproduce it by TC. A similar analysis for the 1B6K ligand, PI5, can be found in the Supporting Information.

Three of the residual seven failures are thus explained by the choice of metric: TC produces a low score for reproduction due to its tendency to drive overlays to match perfectly in one region of the molecule, while RMSD correctly rates the reproduction as satisfactory. The inverse situation applies to the 1Y6B ligand, as the optimal overlay by TC is informative and

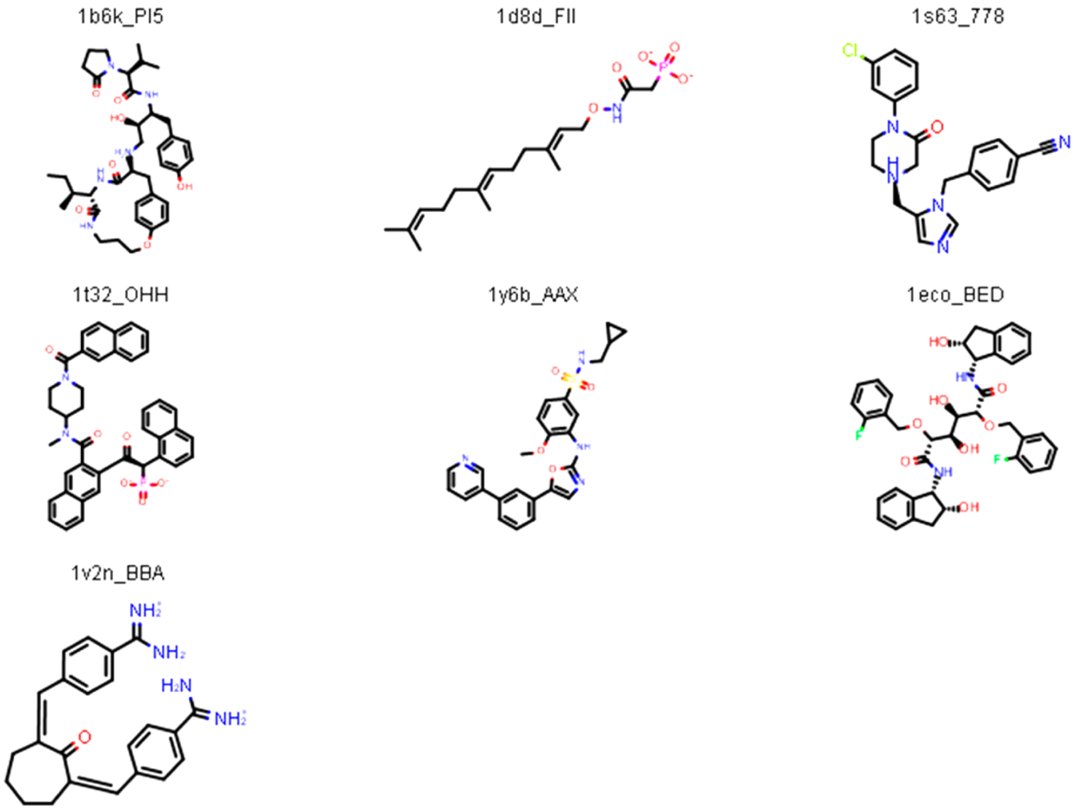


Figure 7. 2D structures of the seven ligands from the PDB set unsuccessfully reproduced with modified OMEGA settings.

Table 7. Worst RMSD (Highest), TC (Lowest), and RMSTanimoto (Lowest) for All Pairwise Comparisons of All the Instances of the Seven Failure Ligands Found in the PDB

ligand	instances	min TC	max RMSD (Å)	min RMSTani
PI5	1	-	-	-
FII	10	0.93	1.24	0.92
BED	1	-	-	-
778	8	1.50	1.49	0.81
OHH	2	0.57	3.21	0.63
BBA	4	0.894	1.063	0.919
AAX	1	-	-	-

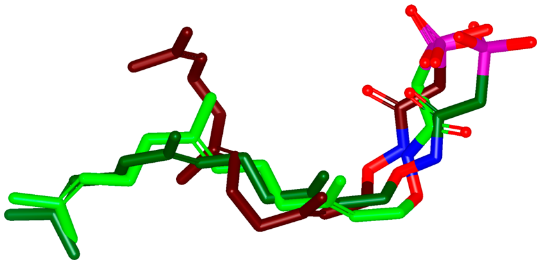


Figure 8. Comparison of the experimental conformation of the 1D8D ligand FII (light green) with the best overlay as provided by TC (red, TC = 0.93) and by RMSD (dark green, RMSD = 1.49 Å).

relatively accurate (giving a good score) except for the misplacement of the 3-pyridyl ring, while the optimal RMSD overlay is poor in almost all areas (giving a poor RMSD), as shown in Figure 9. However, neither of these two conformations from OMEGA would be suitable for docking or shared feature detection, and thus OMEGA has failed to satisfactorily reproduce the experimental conformation.

Careful inspection of the closest matching OMEGA conformations to the crystallographic structure of the 1Y6B ligand, AAX (overlain by TC, not RMSD), revealed that the difficulty in reproducing this structure arose from setting the torsion connecting the aminooxazole to the phenyl group incorrectly. The correct setting exists in the OMEGA torsion library, so a closely matching conformation is in theory produced by OMEGA but was not observed in the default output. To explore the conformational space of AAX much more carefully, we regenerated a conformer ensemble using no geometric deduplication and with an *ewindow* of 25 kcal/mol. These settings represent a very thorough sampling of the available conformational space of a molecule, at a strain energy much higher than we would expect to find in a well-solved ligand structure. However, still no conformation acceptably close to the crystallographic was produced, again due to the misplacement of the pyridine group. Deeper investigation revealed that an internal setting controlling the size of the memory footprint occupied by OMEGA (and thus the size of the raw pool of conformers generated) was responsible for the loss of all those conformations close to the experimental; these conformers were only returned to the user if a much larger pool of raw conformers was generated. When OMEGA was allowed to occupy a larger memory footprint, increasing the size of the allowed pool of raw conformers, the final ensemble contained a conformer very close to the experimental (TC = 1.59, RMSD = 0.96 Å, RMSTanimoto = 0.94) at 324th in the energy ranked list, 3.9 kcal/mol above the lowest energy conformer found. So here, a platform-based problem arising from limitations on accessible memory on 32 bit platforms, rather than a fundamental issue with the OMEGA knowledge base or force

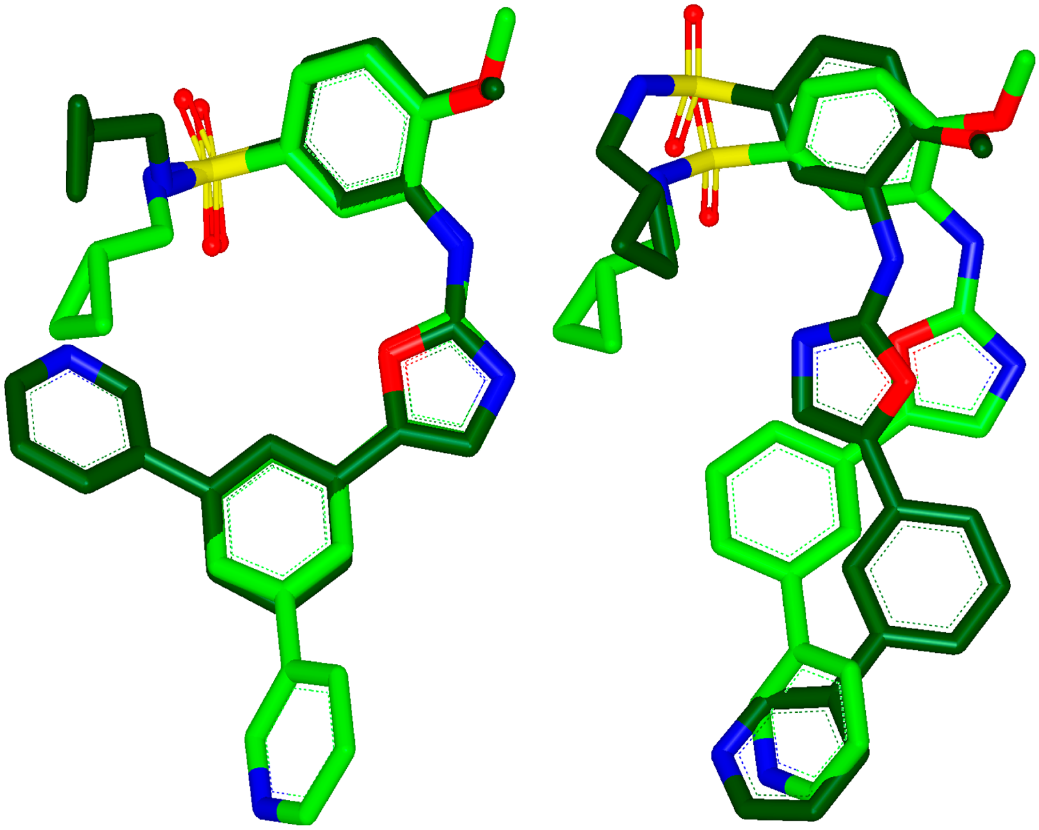


Figure 9. The best reproductions of the 1Y6B ligand AAX as provided by TC (TC = 1.26, left) and RMSD (RMSD = 2.14 Å, right). The PDB conformation is in light green, the closest OMEGA conformation in dark green.

field, has confounded us. Whether the geometric properties of the AAX ligand in the 1Y6B structure are unique is impossible to determine as AAX exists in only a single instance in the PDB, in the 1Y6B structure studied here.

The failure to reproduce the 1Y6B ligand was easily detectable by RMSD, the most common metric of quality in studies of this sort. However, other clear failures were not readily identified using RMSD but were readily illuminated by the complementary metric RMSTanimoto. It is in the case of the smallest ligand in the set of failures, the 1S63 ligand (het code 778), that the RMSTanimoto metric showed its utility most clearly. When judged either by RMSD or by TC, OMEGA is able to reproduce the crystallographic conformation of 778 reasonably well (see Table 7). However when judged by RMSTanimoto, OMEGA fails to reproduce the structure acceptably, which is confirmed by visual inspection of the alignments (see Figure 10). Molecule 778 is quite small, which allows a relatively low RMSD to arise from a fairly poor reproduction of the crystallographic conformation (vide infra for more striking examples of this observation in the CSD-derived set). As is also clear from Figure 10 the conformation of 778 is very compact or folded, which presents special problems for conformation generators.

An analysis of the difficulty of reproducing folded conformations, like that shown by 778 in the 1S63 structure, was reported by Chen and Foloppe,[22] with specific examples of rescuing some failures in reproduction of folded ligands by parameter manipulation. While not in the Chen set of folded ligands, 778 shares the low radius of gyration and compact geometry possessed by ligands in that set. Therefore, cases such as 778 in the 1S63 structure should be expected to be difficult for OMEGA to reproduce accurately.

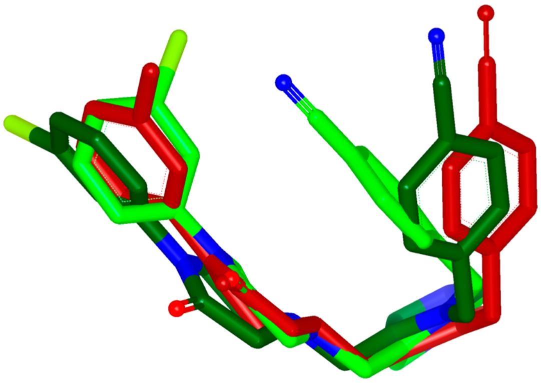


Figure 10. Comparison of the experimental conformation of the 1S63 ligand 778 (light green) with the best overlay as provided by RMSD (dark green, 1.70 Å) and the best overlay as provided by TC (red, 1.22).

As can be seen from Figure 10, the reproduction of the crystal structure is poorest for the 4-cyanobenzyl group. The experimental structure exhibits a 66.5° angle for the torsion governing the disposition of this group, and, while this correct angle is closely matched in the OMEGA torsion library, a conformer bearing this torsion is not output with the default settings. We found that simply increasing the *maxconfs* parameter to 400 solved this failure; a conformation close to the experimental was found (TC = 1.38, RMSD = 1.03 Å, RMSTanimoto = 0.906) that was 214th in the energy ranked list, 5.4 kcal/mol above the lowest energy found. Therefore, although OMEGA does have the capability of generating an appropriate conformation, the pruning algorithm removes it from the default output as it is too high in energy to be reached within the maximum number of conformations allowed for a molecule of this flexibility (200). We conclude that the low-energy landscape of 778 is densely populated by geometrically diverse conformers, and the folded conformation that 778 exhibits in the 1S63 structure imparts a moderate degree of strain to the conformation. The poor reproduction from OMEGA arises solely from the choice of the maximum number of conformations permitted in the output ensemble; increasing *maxconfs* to explore higher energy conformers produced a good match to the experimental structure. The contrast with the results reported by Chen and Foloppe[22] is interesting – in their analysis improved reproduction of a set of four folded ligands is improved by removing bias toward folded conformations in the conformation generators studied, while in this single example simply allowing for more conformations to be generated improves reproduction. Unfortunately there is no obvious property of AAX that distinguishes it from other ligands of similar flexibility that are well reproduced. Therefore the required change in the parameters to produce an acceptable reproduction could not be known *a priori*.

To further investigate the conformational properties of the 778 ligand, we tested OMEGA against its other instances in the PDB. In all ligand 778 occurs in eight different instances in three different complexes in the PDB: 1S63 (one instance), 1S64 (six instances), and 3Q7A (one instance). As can be seen from Table 7, the different conformers are relatively similar to one another, with the maximum RMSD between any pair of instances being 1.49 Å. However, the 778 instance from the 1S63 structure is the most difficult of the eight instances for OMEGA to reproduce.

Analysis of the failure to reproduce the deposited model for the 1T32 ligand, OHH, found in the Supporting Information, produced very similar results. The deposited ligand model shows relatively high strain against MMFF94 (due to the presence of a cis amide bond), requiring 1600 conformations to be generated before finding one that closely matches the deposited model (again no change in the energy window was required). The 1T32 instance of the OHH ligand is also rather more difficult for OMEGA to reproduce than the only other known instance of OHH, in the 1T31 structure.

The genuine failures, unrelated to the metrics used to assess failure, seen so far fell into two classes:

1. The low-energy conformational space for the ligand, under the energy window cutoff, is larger than expected given its flexibility, and this space is inadequately sampled (778 in 1S63 and OHH in 1T32).

2. An implementation issue (AAX in 1Y6B) related to control of OMEGA's memory requirements.

We imagined that the last failure in the set, the ligand from the 1EC0 structure (ligand code BED) would likely be due to the first of these two reasons, as the BED ligand is a large, flexible molecule (with 15 rotatable bonds). BED is a member of the well-studied dihydroxyethylene class of HIV-1 protease inhibitor.[26] When the best matching conformer from OMEGA was compared to the experimental structure, much of the divergence between them arose from discrepancies in the disposition of the pendant 3-fluorobenzyl groups, as seen in Figure 11.

The torsion that governs the disposition of this group in the X-ray structure is 68°, while the closest matching conformer output from OMEGA is 180°. The OMEGA torsion library contains settings for this torsion at 60° and 80°, so closely matching conformers could be produced but are not in the output. We assumed that, as before, the conformational landscape of BED under the default energy landscape was simply inadequately sampled. However, preliminary studies showed that even enumerating all conformers for this molecule with energies under 25 kcal/mol above the lowest energy conformer found did not produce a satisfactorily close match to the crystallographic conformation. Examination of the energetics of this ligand according to MMFF94 found that the deposited conformation seemed curiously strained (over 12 kcal/mol from its local minimum). While the goal of this paper is not to discuss the issue of strain in ligands bound to protein structures, we found this observation provocative. Whether this large level of strain is due to a deficiency in MMFF94's treatment of this molecule's energetics or some other factor was not, to this point, well understood. Preliminary investigations with *ab initio* and DFT quantum mechanical methods also showed a large amount of strain in the 1EC0 ligand structure, up to approximately 30 kcal/mol from an approximation to the global minimum (P. Hawkins, in preparation). Given the apparently high strain energy in the deposited conformation and that OMEGA's goal is to produce low energy conformations, it should come as no surprise that reproduction of the conformation of the 1EC0 ligand proves difficult. The BED ligand is found only once in the PDB, so we had no direct comparator to use to determine if there was some systematic error in the BED ligand conformation in 1EC0 or whether this high energy is actually a genuine property of the structure. However, in the original data set, before graph-based similarity filtering, there was a ligand of very similar structure to BED—the BEG ligand in structure 1D4I. This ligand is reproduced very easily by OMEGA (RMSD = 1.12 Å, TC = 1.84, RMSTanimoto = 0.89), and its structure is almost identical to BED (see Figure 12). We therefore reasoned that the change from a hydroxyethylene central motif in BEG to the dihydroxyethylene motif in BED (the only significant difference between the two molecules) had caused some perturbation of the structure that made reproduction by OMEGA difficult.

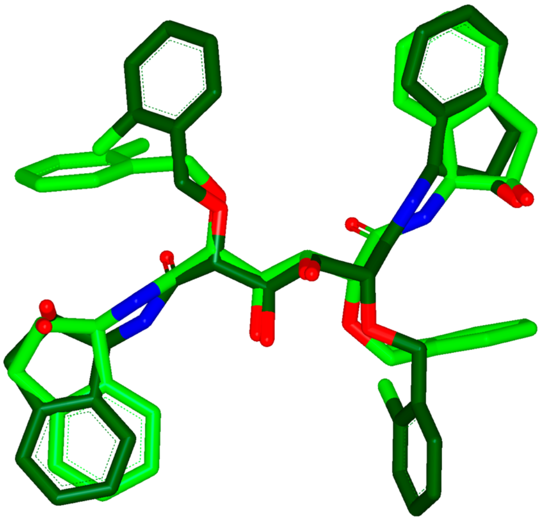


Figure 11. Comparison of the experimental conformation of the 1EC0 ligand BED (light green) with the best overlay as provided by TC (dark green, TC = 1.08).

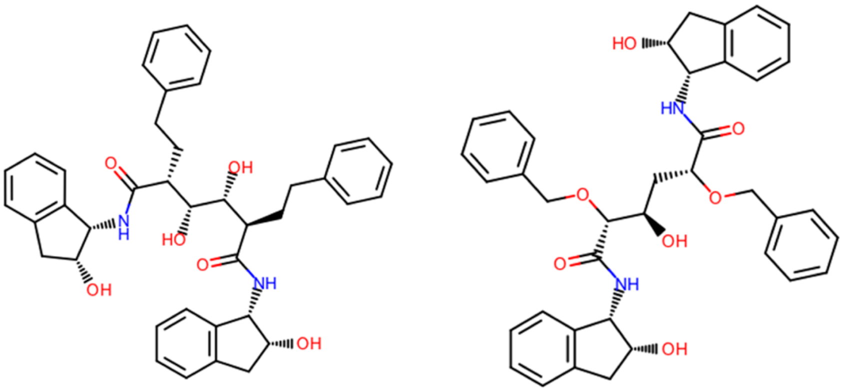


Figure 12. Comparison of the 1EC0 ligand, BED (left), with the 1D4I ligand, BEG (right).

Closer inspection of the dihydroxyethylene motif in the BED structure revealed two significant discrepancies between the experimental conformation and those calculated by OMEGA:

1. The central dihydroxyethylene O−C−C−O motif in BED has a torsion angle of 48°, while the OMEGA conformations have the same torsion at 68°.

2. The bond angles at entry and exit to this motif in the experimental structure are significantly discrepant from the angle of 109.5° found in the OMEGA conformations; the experimental bond angles are 101° and 115° (the level of asymmetry alone is striking).

This large asymmetry of the bond angles entering and exiting the central motif of the ligand, their divergence both from theory and the distribution found for this motif in the CSD (data not shown) is clearly due to an error in the deposited model. The origin of the divergence in the torsion angle for the central motif was less clear. To gain further insight into the solid state conformational preferences of the 1,2-disubstituted dihydroxyethylene motif, we interrogated the CSD for a frequency distribution for this torsion when contained in organic molecules with well-solved, reliable structures. We then compared this distribution to a torsion scan of this bond (in the model compound *cis*-2,3-dihydroxybutane) at two different levels of theory; MMFF94 and MP2/6-31G**. The results are shown in Figure 13.

The energetic analyses from both DFT and MMFF94 are pleasingly consistent with one another and show remarkably good agreement with the population distribution from the CSD. The overwhelmingly preferred angle for this torsion is 170−180° (82% of the CSD structures), in agreement with the intuition that the oxygen atoms, both bearing a significant negative partial charge, would prefer to be disposed as far apart from one another as possible. The next most preferred angle is 60−75° (12% of the CSD structures), as found in the best matching conformations for the 1EC0 ligand from OMEGA, with an energy penalty of around 3 kcal/mol above the 180° conformation. The 48° conformation found in the PDB structures mentioned above is calculated to have a strain energy of around 5.1 kcal/mol above the 180° conformation by MMFF94; this conformation is never found in the CSD.

We then interrogated the druglike subset of the PDB that we had already assembled for the torsion distribution of the dihydroxyethylene group and found a striking disparity between the PDB distribution and the results above. This discrepancy between the conformational distributions of this motif in the CSD and the PDB is illustrated in Figure 14.

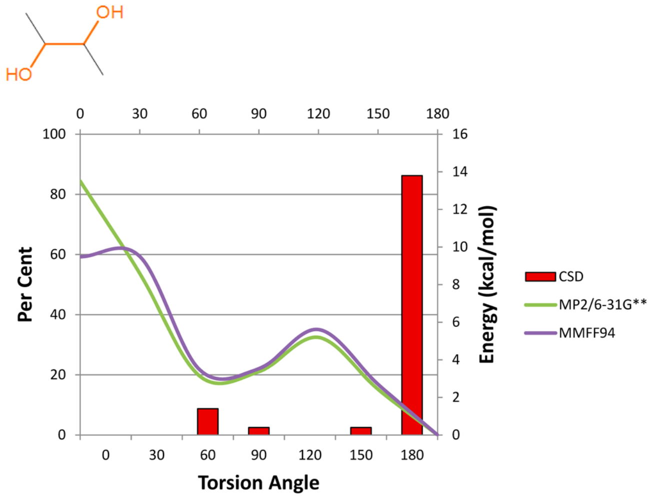


Figure 13. Energy profile and CSD conformational distribution for the 1,2-disubstituted dihydroxyethylene motif. The vertical axis at left shows the distribution frequency of the given torsion angle in the CSD; the vertical axis at right shows the energy above the lowest energy found (which was set to zero).

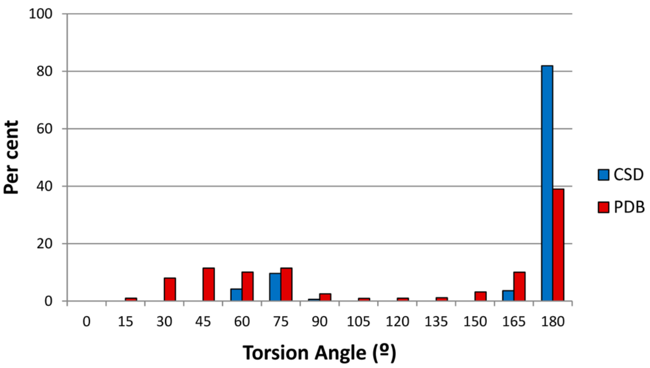


Figure 14. Conformational distribution for the 1,2-disubstituted dihydroxyethylene motif in the PDB (red) and the CSD (blue).

While the majority of the PDB cases and the majority of the CSD cases have the same torsion angle (180°), the predominance of this form in the PDB is much lower than in the CSD. Angles not populated at all in the CSD have significant populations in the PDB, including the 45° found in BED (and other HIV-1 protease ligands, vide infra). As mentioned by Brameld et al.,[12] the large discrepancy between torsion distributions from the PDB and the CSD can arise from a number of sources but probably implies a softer energy potential for the torsion in question. The energy profile for the OCCO torsion in Figure 13 is indeed soft between 60° and 120° but is rather harder between 60° and 0°, implying that achieving the experimental 48° torsion would require a significant expenditure of the binding energy of the protein–ligand complex. In the case of dihydroxyethylene HIV-1 protease inhibitors like BED, the requirement for productive hydrogen bond formation between the active site Asp residues of HIV-1 protease (D25 and D125) and the OH groups of this motif may help to enforce an otherwise unfavorable conformation. However, the degree of deviation from observed torsion angles in the CSD was disturbing.

To gain a better understanding of the discrepancies between the PDB structures containing this motif and those in the CSD and the properties predicted by theory, we searched the PDB for other HIV-1 protease-ligand structures where the ligand contained the dihydroxyethylene motif. We found 12 structures (each with one ligand instance) for which electron density has been deposited. We analyzed all these ligand conformations to determine if patterns in bond and torsion angles were present. Of the 12 ligand instances, all but one shared a high degree of similarity in the geometry of the central motif—different bond angles at entry and exit and an approximately 45° torsion for the O−C−C−O group, as seen in BED. The lone different structure is 1WBM, containing the BLL ligand (the two ligands are shown in Figure 15).

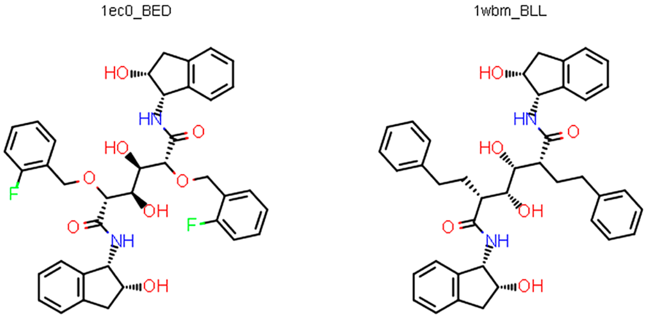


Figure 15. Comparison of the structures of the 1EC0 ligand, BED (left), with the 1WBM ligand, BLL (right).

While the molecular graphs of these two molecules are almost identical, their 3D structures are somewhat different. The BLL structure exhibits bond angles for entry and exit to the central motif that are both equal at 109.8°, consistent with our expectations of a sp^3 hybridized carbon atom, while the central O−C−C−O torsion is 60°. Both observations are much more consistent with theory, our computations and the frequency analysis of the CSD. The effect of these difference in the bond angles and torsion of the O−C−C−O motif on the conformations of the BLL and BEG ligands is shown in Figure 16. It is obvious from the alignment that this motif is significantly different in the two structures.

We then turned to a close examination of the experimental electron density for the ligand in these two structures. When judged by numerical metrics of local model quality (like RSCC and RSR[9]) BED is a better fit to its density than BLL is to its density. Likewise both ligands seem by visual inspection to be a good fit for their electron density, as represented in the $2F_o$-F_c maps. However, the F_o-F_c difference maps for the ligands clearly show the existence of a blob of unfilled density adjacent to one of the dihydroxyethylene carbon atoms in the 1EC0/BED structure, while the 1WBM/BLL structure does not have this unfilled density (see Figure 17). The existence of unfilled density in a structure indicates that the deposited coordinates are insufficient models for the electron density—the unfilled density should be occupied by one or more heavy atoms for better matching with the diffraction data but is not.

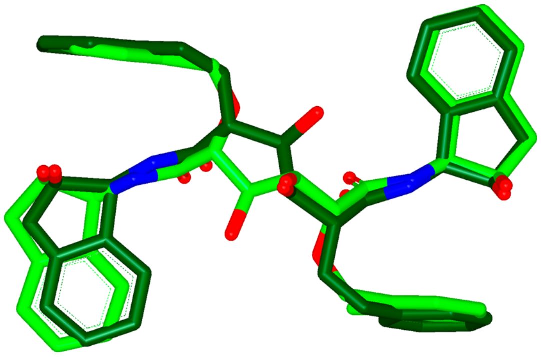


Figure 16. Comparison of the conformations of the 1EC0 ligand, BED (light green), with the 1WBM ligand, BLL (dark green).

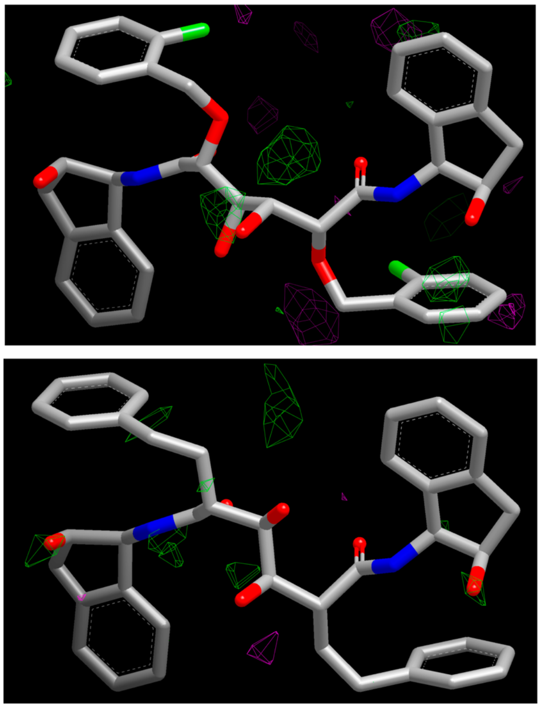


Figure 17. F_o-F_c difference maps contoured at 3 sigma for 1EC0/BED (upper) and 1WBM/BLL (lower). Unfilled electron density is shown in green blobs.

That the F_o-F_c difference map for the 1WBM structure shows no significant unfilled density adjacent to the ligand indicates that the BLL ligand in this structure is in some respects a better model of the electron density than the BED ligand in the 1EC0 structure is. Comparative analysis of the other 10 structures of HIV-1 protease complexed with this class of inhibitor showed that all 10 possessed the blob of unfilled density adjacent to the ligand in the F_o-F_c difference map. It seems likely, therefore, that the dihydroxyethylene linker atoms in these inhibitors have been consistently misfit to the electron density, giving rise to the incorrect geometry observed in 11 of the 12 structures of this class. Accordingly, our finding that OMEGA failed to reproduce the deposited structure for the BED ligand is due to the deposited structure of BED, and most of its congeneric ligands, being incorrect. The reason for this consistent error in the majority of structures of this ligand type is not known. The BLL ligand, however, is probably a much better model of ligands of this type than the other examples in the PDB, as shown by its much better consistency with CSD data and theory. Perhaps not coincidentally, of the 12 molecules examined here, BLL is the easiest for OMEGA to reproduce well (RMSD = 1.66 Å, RMSTanimoto = 0.91, TC = 1.13).

This rather exhaustive analysis of a single failure case illustrates the unfortunate difficulties with using PDB structures to validate conformer generators. In spite of our extensive efforts to ensure that only the best quality ligand models were used for validation,[9] the BED ligand did end up in our PDB validation set, indicating that even greater care needs to be taken in assembling ligand sets from the PDB. It is now clear that even use of local quality metrics like RSCC and RSR is insufficient to identify poorly fit ligands, and further tests of structural quality must be applied before a ligand can be included in a validation set such as the one used here. A qualitatively similar finding on the difficulty of accurately reproducing misfit ligand structures was reported by Bostrom[27] on a different set of molecules, but in that case the molecules were not first preselected for model quality, as was done here.

Overall, a simple change in OMEGA's defaults (increasing *maxconfs* to 800 for molecules with 8 or more rotatable bonds) to allow greater searching of low-energy conformational space for flexible molecules resolved 50% of our original failures. Almost 50% of the remaining failures, when judged by visual inspection, were due to the nature of TC as an overlay and scoring metric. Of the genuine failures, one highlighted an algorithmic issue in OMEGA due to the memory demands of processing large, flexible molecules, two showed that *a priori* estimation of the size of the low energy conformational landscape of even relatively small molecules is difficult, while another demonstrated the eternally vexing issue of strain in ligands in the PDB, especially when the ligands are not reasonable physical models of the electron density. With these data in hand, we turned to an examination of the failure cases in the CSD data set.

Failures among CSD Molecules. The CSD is a fruitful source of challenging small molecule conformations to reproduce, and it avoids some of the problems inherent in using small molecule structures from the PDB, particularly the structural pathologies highlighted above for the HIV-1 protease ligands. In our previous publication, we used 483 druglike molecules from the CSD as a parallel validation set for OMEGA. On closer examination of the set using the torsion fingerprinting approach outlined in the PDB ligand section, we found two molecules that were close analogues of another compound in the set, and so the analogue with the smallest rotatable bond count was removed to give a final set of 481 molecules. Conformers were generated for this set using OMEGA's defaults; these were compared to the experimental conformation, using the three metrics discussed previously. The aggregate results are shown in Table 8, and the complete results are shown in Figure 18.

We do not supply variance-weighted means for the CSD data, as the experimental error in the atom positions for these models is very low.

Table 8. Mean, Standard Deviation (StdDev), and Median for TC, RMSD, and RMSTanimoto for 481 Ligands from the CSD

	TC	RMSD (Å)	RMSTanimoto
mean	1.72	0.51	0.97
StdDev	0.23	0.34	0.05
median	1.79	0.42	0.98

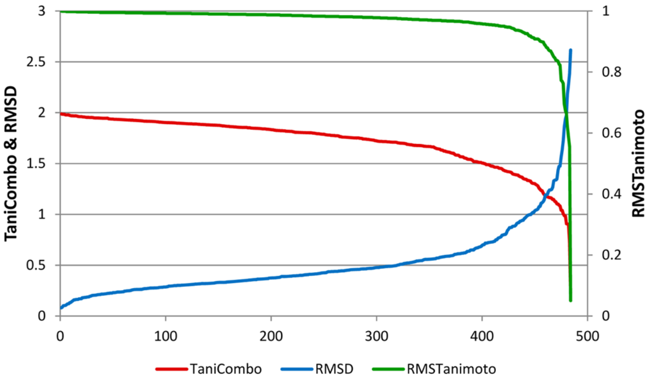


Figure 18. Distribution of TC (red), RMSD (blue) (vertical scale at left), and RMSTanimoto (green, vertical scale at right) between the closest conformer from OMEGA and the experimental conformation for 481 molecules from the CSD.

It can immediately be seen from Figure 18 that there are very few (3/481) failures in the CSD set when the standard cutoff of 2Å RMSD is applied. We might therefore be justified in declaring OMEGA an excellent tool for reproducing conformations of druglike molecules in the CSD. However, inspection of the RMSTanimoto and TC plots reveals that, once again, these metrics find failures not found by RMSD; RMSTanimoto finds five failures, three of which were not found by RMSD, while TC finds three failures, one of which was not found by RMSD. The five ligands that fail at least one of the three criteria are found in Table 9, along with their rotatable bond count. Plots illustrating the relationship between RMSD and RMSTanimoto, the ratio of RMSD to heavy atom count to RMSTanimoto, and the relationship between RMSTanimoto and heavy atom count, for the CSD ligands are contained in the Supporting Information.

Table 9. RMSD, TC, RMSTanimoto, and Rotatable Bond Count for the Five Ligands from the CSD Data Set That Cannot Be Successfully Reproduced As Judged by One or More of These Metrics[a]

CSD code	TC	RMSD (Å)	RMSTanimoto	rotors
DCTXAN	**0.91**	1.94	**0.76**	5
FECPUY	1.01	1.86	**0.56**	3
NADYIA	**0.72**	**2.63**	**0.59**	7
SIHDIW	**0.90**	**2.28**	**0.66**	6
SURREC	1.10	1.64	**0.77**	6

[a] Failures highlighted in bold.

Unlike the situation with the PDB data set, where the great majority of poorly reproduced molecules have large numbers of rotatable bonds, two of the CSD failures have five or fewer (see also the depictions of these molecules in Figure 19). Clearly, the problem with these molecules is not simply an insufficient sampling of a large low-energy conformational space. Very

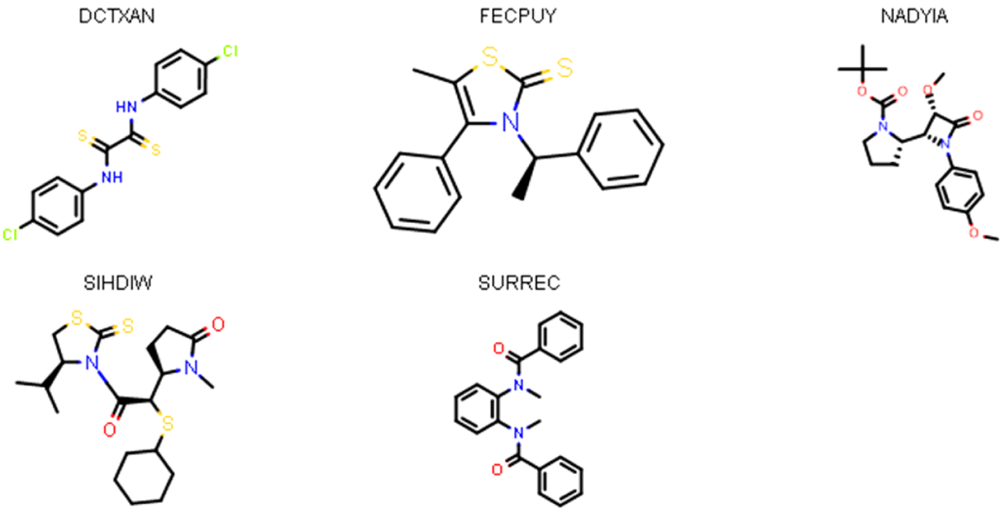


Figure 19. 2D structures of the five ligands from the CSD set unsuccessfully reproduced with default OMEGA settings.

simple structures like DCTXAN, FECPUY, and SURREC should be straightforward to reproduce, and indeed these structures are reproduced with an RMSD < 2Å. However, RMSTanimoto rescoring of the RMSD overlays indicated that these relatively small molecules were not well reproduced, and visual inspection of the overlays confirmed that the low RMSTanimoto scores did in fact indicate poor reproductions. The small size of molecules like DCTXAN and FECPUY reveals an inherent weakness in RMSD; molecules with a low heavy atom count simply have a limited ability to show high RMSD once overlain.

However, as we had hoped, the fact that RMSTanimoto scales over a defined range allowed it to identify failures of reproduction for small molecules like DCTXAN, FECPUY, and SURREC that were not identified by RMSD; for DCTXAN RMSD = 1.94 Å, while RMSTanimoto = 0.76. Inspection of the best overlay of any conformer from OMEGA to the crystallographic conformation of DCTXAN is shown in Figure 20, confirmed that the reproduction was poor. This poor quality of the OMEGA reproduction of DCTXAN is mostly due to incorrectly setting the central S–C–C–S torsion of the thiooxalamide motif, as shown in Figure 20. This motif is planar in the crystal structure (torsion is 180°), while the best OMEGA solution shows a torsion of 92°. The great majority of organic molecules with this motif in the CSD show a torsion of 170–180° (around 85%), with the rest showing the more difficult to reproduce torsion of around 90° (data not shown).

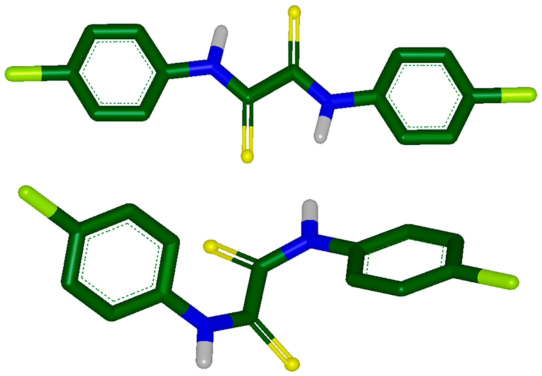


Figure 20. The best overlay by RMSD between the experimental structure of DCTXAN (upper) and the closest conformation from OMEGA (lower, RMSD = 1.94 Å).

The planar nature of the central thiooxalamide motif in the crystal is not energetically favorable; according to MMFF94, the experimental conformation is 3.5 kcal/mol higher in energy than its closest local minimum *in vacuo,* 9.1 kcal/mol higher in solution. solution. This is most probably due to the close 1,4 N–H to S contacts in the crystal structure (2.27 Å, less than the sum of the vdW radii). The most computationally stable conformation both in vacuo and in solution is the orthogonal arrangement shown by the OMEGA conformation in Figure 20. Planarity of the thiooxalamide motif may be enforced in the solid state by crystal packing forces (this assertion is supported by the large number of close contacts in the crystal lattice between this motif and neighboring molecules). It has also been suggested that crystal field effects may be the cause of planarity.[28] There exists another conformation of DCTXAN in the crystal lattice having the same essentially planar conformation of the central motif but differing slightly in its N-phenyl torsion. The two variant conformations are reproduced with almost the same accuracy by OMEGA. As such, we hypothesized that the difficulty in reproduction of the DCTXAN structure lies in its high strain due to the planar disposition of its central motif. An analysis of the failure to reproduce the conformation of FECPUY (see the Supporting Information) arrived at the same conclusion – the experimental conformation of FECPUY is difficult for OMEGA to reproduce due to high strain.

Just as we found with the PDB-derived set, overlays driven by TC are often straightforward to analyze by inspection as well as to identify the torsion(s) responsible for a poor alignment. This is illustrated in Figure 21 for the SURREC molecule, where the overlay of the best matching conformer from OMEGA by TC is compared to the crystallographic conformation. It can clearly be seen that the OMEGA conformation possesses one trans and one cis N-methyl amide bond, while the crystallographic conformation has two cis amides, as expected for N-methylated aromatic anilides.[29] This is again a small molecule that we would not have expected to be difficult to reproduce (the OMEGA torsion library contains entries for both cis and trans amide bonds). The experimental conformation of SURREC, like that of DCTXAN, exhibits significant strain against MMFF94 (5.5 kcal/mol from its local minimum). The

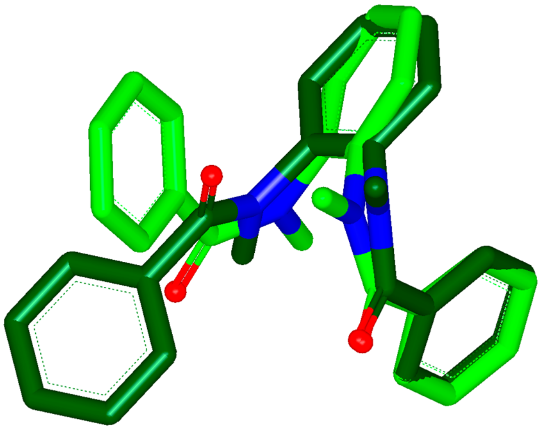


Figure 21. The best overlay by TC between the experimental structure of SURREC (light green) and the closest conformation from OMEGA (dark green, TC = 1.10).

consistent observation of strain in the conformations of these rather small molecules led us to suspect that some or all of the failures in the CSD set may be due to the CSD structures being strained (at least according to MMFF94). Examination of the asymmetric units of the failures revealed that close contacts to other molecules in the crystal lattice exist in 4/5 of the failures. This suggested to us that, in these four cases at least, the experimental conformation is being influenced by the crystal environment and that this environment imposes a higher energy conformation on the molecule. In Figure 22, we compare a simple estimation of strain energy (energy above the nearest local minimum for MMFF94) for our entire CSD set and for just the failures.

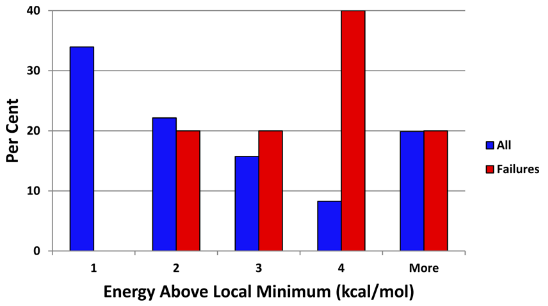


Figure 22. The distribution of local strain energy using MMFF94 in the gas phase for the total CSD set (blue, 481 molecules) and the failure set (red, five molecules).

Figure 22 clearly shows that the failures are more locally strained than the validation set as a whole. While the small size of the failure set does not allow us to draw any statistically firm conclusions, it seemed likely that part of the reason for the failures may be high strain in the crystallographic conformations. We therefore elected to alter the energy cutoff or *ewindow* used in the OMEGA calculation to 20 kcal/mol, to allow more strained conformations to be produced. With this one change in the default parameter set, OMEGA succeeded in reproducing all the crystallographic conformations in the validation set. This straightforward resolution of the failures lends further support to our hypothesis that, at least according to MMFF94, the failure conformations are high in strain. The aggregate statistics for reproduction using any of the three metrics used were almost identical to the statistics for the defaults and were not separable at a 95% confidence interval (data not shown). This is entirely expected, as the failures were a very small proportion of the entire data set.

Given that we have already shown that our CSD-derived validation set is an imperfect sample of the druglike molecules in the CSD as a whole, we attempted to expand our coverage of difficult torsions in the CSD by finding analogues of the molecules that failed to be well reproduced with OMEGA's defaults. In this way, we hoped to further explore the solid-state conformational preferences of related molecules and perhaps identify some other informative failures (or discover if we had inadvertently selected a particularly difficult structure to reproduce). Unlike the PDB set, where exactly the same ligand can occur many times, either in the same asymmetric unit cell or in different structures, in the CSD a given molecule almost always occurs only once. Accordingly, rather than look for the same molecules that failed to be well reproduced, we performed substructure searches to find analogues of the five molecules in the failure set. These substructure searches deliberately omitted cyclic versions of the failure molecules, as cyclization imposes new constraints upon the conformation, making comparisons between the acyclic and cyclic versions essentially uninformative. We then examined OMEGA's performance in reproducing these analogue structures. In Table 10, we provide the number of hits found for each substructure search and the number of OMEGA failures in the new set of molecules found by the substructure search.

Table 10. Number of Hits from a Substructure Search for Each of the Failure Molecules and the Number of Those Hits That OMEGA Cannot Successfully Reproduce with Default Settings[a]

	analogues in CSD	OMEGA failures
DCTXAN	13	2
FECPUY	0	-
NADYIA	8	1
SIHDIW	1	0
SURREC	10[a]	2[a]

[a] In three cases, the hit molecule was found in two different entries in the CSD.

For the SURREC molecule, the substructure search found three structures that were each found in two different conformations in the CSD (SURREC itself, TOHVIW, and TORTIE). In each case, OMEGA successfully reproduced one of the versions and failed on the other. Overall, OMEGA was successful in reproducing almost all the conformations of the analogues using its default settings (see Table 10) and, as before, increasing the *ewindow* to 20 kcal/mol resolved all the failures. As such, our expanded search of analogues of difficult structures confirmed our original findings that a simple increase in the *ewindow* is a robust solution to reproducing conformations from the CSD. This expanded set of analogues also confirmed our original finding that the molecules that are difficult to reproduce tend to be higher in local strain than the validation set as a whole (data not shown).

■ CONCLUSION

We have utilized two previously selected data sets for identifying failure cases in OMEGA's ability to reproduce crystallographic conformations. These validation data sets have been carefully examined to understand the relationships between the properties governing molecular flexibility (rotatable bond count and number of flexible rings) in these data sets and the parent sets (the druglike subsets of the PDB and the CSD). We found that the space of torsions is well explored by our data sets, while the space of flexible rings is much less so. We are therefore confident that our validation sets have produced reliable predictions for OMEGA's performance on a wide variety of molecules lacking flexible rings, though our ability to predict OMEGA's performance on molecules containing flexible rings is much less reliable. In future work, we will expand our validation sets by identifying ligands with well-solved models that contain a variety of flexible rings to more thoroughly evaluate OMEGA's ability to correctly reproduce the conformations of these rings.

As a complement to the reproduction measures used in our previous publication, RMSD and TC, we have introduced a new metric, RMSTanimoto. RMSTanimoto was conceived as a way to overcome some of the well-known deficiencies in RMSD, particularly its inherent size dependence. Since it is both bounded and metric, RMSTanimoto offers the possibility of a much more reliable and transferrable measure of reproduction success across molecules of varying size. The use of this set of three different metrics was applied to the problem of identifying failures of reproduction in the CSD- and PDB-derived sets. Pleasingly, the three metrics found significantly nonoverlapping sets of failures, RMSTanimoto being particularly effective for identifying failures on very small molecules. Understanding the reasons for failures was often greatly helped by the TC-optimized overlays, where maximization of functional group overlay is sought, as these often clearly showed the areas of maximum deviation between the crystallographic and computed conformers (information that was often not obvious from the RMSD-optimized overlays).

Close analysis of the PDB failures revealed that most were due to undersampling the large low-energy conformational space accessible to molecules with eight or more rotatable bonds. Simple adjustment of the maximum number of conformations generated for these more flexible ligands removed most of the failures. The more informative of the residual failures revealed some indication of high strain/overestimation of strain in these structures by the MMFF94 force-field (1T32), as well as an algorithmic problem. A final case indicated that difficulty in reproduction lies in a poor deposited model for the ligand (and most of its congeners). The reason for the poor deposited models of these closely related molecules is unclear. In general we found, using the variance weighted mean, that OMEGA is better able to accurately reproduce PDB structures with lower experimental error (as measured by DPI). Analysis of the failures from the CSD set showed that flexibility was not an issue in this case but rather that strain induced in some of the experimental structures by close contacts in the crystal was causing failures. This apparent strain can be accounted for by increasing the energy window permitted for these molecules. This simple change to the defaults resolved all the failures, leading us to suggest that these difficult conformations exhibit unusually high levels of strain against the MMFF94 force-field.

On the assumption that the reasons the molecules failed are worth studying in depth, we expanded the coverage of the failure cases by finding the same molecule present in other structures (PDB) or close substructural analogues (CSD). These expanded sets confirmed the solutions identified for the original failures and, for the PDB set, showed that significant conformational diversity can exist in different deposited models for the same molecule. Unexpectedly, in each case where multiple instances exist, our data set contained the most difficult to reproduce instance of the different variants of the same ligand.

At the beginning of this work we posed three questions:

1. Are OMEGA's defaults satisfactory for conformer generation for small molecules?

2. Is the OMEGA knowledge base (the torsion library and the fragment library) adequate to describe the conformational sampling required for reasonable reproduction of crystallographic structures?

3. Is the MMFF94 force-field suitable to describe small molecule conformational energetics?

With the data in hand, we can answer the first question with a qualified yes. For more flexible molecules (those with eight or more rotatable bonds), a single change to the defaults produced very good performance, eliminating all failures. In answer to the second question, we can say that, given that the data sets we used cover a good fraction of the available torsion space we wish to examine, the OMEGA torsion library is indeed adequate for the problem at hand. The issue of the adequacy of the fragment library and its method of generation will be reserved for a future publication since, while OMEGA's performance is very good on our data sets, the fraction of relevant flexible rings examined is too small to draw reliable conclusions. The response to the third question is much more ambiguous. It is worth recalling that at no point in OMEGA's conformer generation process are the conformers optimized against MMFF94 (or any other function); as such, we would expect that in some cases tiny alterations in geometry could relieve steric clashes and considerably reduce the conformational strain observed in a given conformer. With that caveat in mind, one possible conclusion to be drawn from both the PDB and CSD analyses is that the MMFF94 force-field used in OMEGA is prone to overestimate strain in some structures, necessitating a seemingly large energy window to adequately reproduce all the experimental structures under investigation. Another is that some well-solved solid-state structures from both sources do in fact possess significant torsional strain for reasons that are not at the moment completely understood.

The clear indication for future work in this area is a close investigation of the issue of strain in crystallographic structures. An interesting start in this area using quantum mechanical approaches is the work of Sitzmann[30] on strain in ligands from the PDB. The broader issue of the adequacy with which force-fields in general, and specifically MMFF94, describe torsional strain in crystallographic structures is still, unfortunately, unresolved. This area of investigation seems likely to be fruitful for some years to come.

The data sets themselves are available upon request from the authors.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The names of the 200 PDB structures from which the ligands used in this study were extracted, along with their ligand codes. The names of the 481 CSD molecules. A Python script illustrating the calculation of confidence intervals by bootstrapping. Additional analyses of the reasons for failure of molecules from the PDB and CSD set not discussed explicitly in the text. Plots illustrating the relationship between RMSTanimoto and heavy atom count and between RMSTanimoto and the ratio of RMSD to heavy atom count. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Phone: 505-473-7385 ext. 65. E-mail: phawkins@eyesopen.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Dr. Ken Brameld is thanked for providing the molecules from the CSD. Dr. Krisztina Boda and Mr. Brian Cole are thanked for their invaluable assistance on the intricacies of Python to one of the authors (P.C.D.H.). Dr. Robert Tolbert is thanked for his implementation of the RMSTanimoto metric. We thank Dr. Gregory Warren for useful discussions and insights on the practice and theory of X-ray crystallography, Dr. Stanislaw Wlodek for assistance with the quantum chemical calculations, and Dr. Benjamin Ellingson for insights into the inner workings of OMEGA. Ms. Annie Lux is thanked for proofreading the manuscript. We are grateful to the reviewers for useful suggestions.

REFERENCES

(1) Leach, A. R.; Prout, K. Automated conformational analysis: directed conformational search using the A* algorithm. *J. Comput. Chem.* **1990**, *11*, 1193−1205.

(2) Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P.; Miller, M. D. Flexibases: A way to enhance the use of molecular docking methods. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 565−582.

(3) Smellie, A.; Stanton, R.; Henne, R.; Teig, S. Conformational analysis by intersection: CONAN. *J. Comput. Chem.* **2003**, *24*, 10−20.

(4) Bohme-Leite, T.; Gomes, D.; Miteva, M. A.; Chomilier, J.; Villoutreix, B. O.; Tuffery, P. Frog: A Free Online drug 3D conformation generator. *Nucleic Acids Res.* **2007**, *35*, W568−W572.

(5) Vainio, M. J.; Johnson, M. S. Generating conformer ensembles using a multiobjective genetic algorithm. *J. Chem. Inf. Model.* **2007**, *47*, 2462−2474.

(6) 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.

(7) Gippert, G. P.; Wright, P. E.; Case, D. A. Distributed torsion angle search in high dimensions: a systematic approach to NMR structure determination. *J. Biomol. NMR* **1998**, *11*, 241−263.

(8) *Confort*. Tripos Inc. St. Louis, MO, 2012. http://tripos.com/index.php?family=modules,SimplePage,,,&page=Confort (accessed September 18, 2012).

(9) 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 the Cambridge structural database. *J. Chem. Inf. Model.* **2010**, *50*, 572−584.

(10) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. H.; Bourne, P. E. The Protein Databank. *Nucleic Acids Res.* **2000**, *28*, 235−247 http://www.rcsb.org. Accessed 20 December 2010..

(11) CSD: *The Cambridge Structural Database*. Version 5.1. CCDC 12 Union Road, Cambridge, CB2 1EZ, UK. http://www.ccdc.cam.ac.uk/products/csd.

(12) Brameld, K. A.; Kuhn, B.; Reuter, D. C.; Stahl, M. Small molecule conformational preferences derived from crystal structure data. A medicinal chemistry focused analysis. *J. Chem. Inf. Model.* **2008**, *48*, 1−8.

(13) Agrafiotis, A. K.; Gibbs, A. C.; Zhu, F.; Izrailev, S.; Martin, E. Conformational sampling of experimental molecules: A comparative study. *J. Chem. Inf. Model.* **2007**, *47*, 1067−1075.

(14) Chen, I.-J.; Foloppe, N. Conformational sampling of druglike molecules with MOE and Catalyst; Implications for pharmacophore modelling and virtual screening. *J. Chem. Inf. Model.* **2008**, *48*, 1773−1780.

(15) Cole, J. C.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. Comparing protein-ligand docking programs is difficult. *Proteins: Struct., Funct., Bioinf.* **2005**, *60*, 325−334.

(16) Kirchmair, J.; Laggner, C.; Wolber, G.; Langer, T. Comparative analysis of protein-bound ligand conformations with repsect to Catalyst's conformational space subsampling algorithms. *J. Chem. Inf. Model.* **2005**, *45*, 422−430.

(17) Halgren, T. A. The Merck Molecular force field I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−499 . In OMEGA the electrostatic term is turned off by default, see ref 9 and citations therein.

(18) *OpenEye toolkits*. http://www.eyesopen.com (accessed April 14, 2012).

(19) Warren, G. L.; Warren, S.; Do, T. Essential considerations for using protein−ligand structures in drug discovery. *Drug Discovery Today*, **2012**, ASAP. http://dx.doi.org/10.1016/j.drudis.2012.06.011.

(20) http://www.msg.chem.iastate.edu/GAMESS/GAMESS.html (accessed December 2011).

(21) Baber, J. C.; Thompson, D. C.; Cross, J. B.; Humblet, C. GARD: A generally applicable replacement for RMSD. *J. Chem. Inf. Model.* **2009**, *49*, 1889−1900.

(22) *LigandExpo*. http://ligand-expo.rcsb.org (accessed January 12, 2011).

(23) Chen, I.-J.; Foloppe, N. Drug-like bioactive structures and conformational coverage with the LigPrep/ConfGen Suite: Comparison to programs MOE and Catalyst. *J. Chem. Inf. Model.* **2010**, *50*, 822−839.

(24) McGregor, M. J.; Pallai, P. V. Clustering of large databases of compounds: Using MDL "keys" as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.

(25) Blow, D. M. Rearrangement of Cruickshank's formulae for the diffraction component precision index. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *D58*, 792−796.

(26) Lindberg, J.; Pyring, D.; Lowgren, S; Hallberg, A.; Unge, T. Symmetric fluoro-substituted diol-based HIV protease inhibitors. Ortho-fluorinated and meta-fluorinated P1/P1′-benzyloxy side groups significantly improve the antiviral activity and preserve binding affinity. *Eur. J. Biochem.* **2004**, *271*, 4594−4605.

(27) Bostrom, J. Reproducing the conformations of protein-bound ligands: A critical evaluation of several popular conformational searching tools. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 1137−1153.

(28) Shimanouchi, H.; Sasada, Y. Structure of p-p'-Dichlorodithiooxanilde. *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **1979**, *B35*, 1928.

(29) Azumaya, I.; Okamoto, I.; Kagechika, H. Total asymmetric transformation of an N-Methylbenzamide. *Tetrahedron* **1999**, *55*, 11237−11246.

(30) Sitzmann, M.; Weidlich, I. E.; Fillipov, I. V.; Cachau, R. E.; Nicklaus, M. C. PDB ligand conformational energies calculated quantum-mechanically. *J. Chem. Inf. Model.* **2012**, *52*, 739−756.