Journal of Chemical Information and Modeling

Article

pubs.acs.org/jcim

Terms of Use

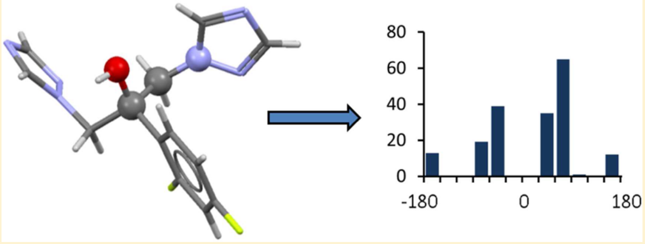
Knowledge-Based Libraries for Predicting the Geometric Preferences of Druglike Molecules

Robin Taylor,* Jason Cole, Oliver Korb, and Patrick McCabe

Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, United Kingdom

S Supporting Information

ABSTRACT: We describe the automated generation of libraries for predicting the geometric preferences of druglike molecules. The libraries contain distributions of molecular dimensions based on crystal structures in the Cambridge Structural Database (CSD). Searching of the libraries is performed in cascade fashion to identify the most relevant distributions in cases where precise structural features are poorly represented by existing crystal structures. The libraries are fully comprehensive for bond lengths, valence angles, and rotamers and produce templates for the large majority of unfused and fused rings. Geometry distributions for rotamers and rings take into account any atom chirality that may be present. Library validation has been performed on a set of druglike molecules whose structures were published after the latest CSD entry contributing to the libraries. Hence, the validation gives a true indication of prediction accuracy.



INTRODUCTION

The ability to predict molecular geometries is important in many areas of chemistry, none more so than structure-based drug design. Interest is primarily focused on the preferred values of torsion angles around rotatable bonds, as these have the predominant influence on overall molecular shape.[1−6] The conformations of flexible rings, particularly macrocycles, are also of significance.[7] Accurate modeling of valence angle preferences has been shown to be important for the conformational analysis of some molecules.[8] Bond lengths have little influence on molecular shape, but their preferred values are clearly of widespread interest, judging by the many thousands of citations that standard bond-length compilations have received.[9,10]

Methods for predicting molecular dimensions may be classified as theoretical or knowledge-based. The former comprise force-field and quantum-mechanical calculations, while the latter usually rely on crystal-structure data contained in databases, most notably the Cambridge Structural Database (CSD).[11] Knowledge-based methods have become progressively more important as the amount of data has increased. Pioneering work was done by Klebe and Mietzner, who devised a program for generating low-energy molecular conformations by using a library of CSD-derived torsion-angle distributions for 216 different types of rotatable bonds.[12] They presented evidence supporting the relevance of CSD-based torsion distributions to the conformational preferences of protein-bound ligands, and additional such evidence was published recently.[13] Torsional distributions based on protein−ligand crystal structures tend to have broader peaks and more outliers than those from small-molecule data, but this may largely be ascribed to the lower experimental precision of the protein structures. It has been argued that packing forces in small-molecule crystal structures rarely cause systematic conformational biases.[14]

In 2004, rapid and automated access to a huge number of CSD-based distributions of bond lengths, valence angles, and acyclic torsion angles was provided in the Mogul system.[15] Subsequently, a library of flexible-ring conformations was added.[16] Mogul is released each year, so the libraries keep pace with the rapidly growing number of published crystal structures. Distributions retrieved from Mogul have been used for many purposes, including (a) determining target bond lengths and angles for use as restraints in the refinement of small-molecule[17] and protein−ligand[18−20] crystal structures, (b) facilitating the solution and refinement of crystal structures from powder diffraction data,[21−23] (c) rationalizing structure−activity relationships,[24] (d) assisting conformational sampling for crystal structure prediction,[25] (e) validating theoretically calculated molecular geometries,[26] and (f) validating ligand geometries in the Protein Data Bank (PDB).[27,28]

We recently embarked on a project to develop a molecular geometry optimizer and a high-throughput conformer generator, both based on geometric data from small-molecule crystal structures. Mogul was an obvious starting point for the knowledge base required by our algorithms, but we found it expedient to customize the system for our particular needs. For example, we wished to improve search speeds, produce torsion distributions that correctly represent the influence of any chirality that might be present, extend the ring library to cover fused rings, improve handling of symmetry, and precluster ring

Received: May 12, 2014

distributions so that only distinct conformational minima are retained. Another important aim was to establish the accuracy with which both the standard version of Mogul and the new customized version predict the geometric preferences of druglike molecules. This will enable objective comparisons to be made between Mogul and other knowledge-based geometry libraries.[2,29,30] The results of our work are reported below. For brevity, "Mogul" is used to refer to the standard version of the system, and "CV" is used to indicate the customized version described herein.

■ METHODS

We begin with a brief description of Mogul. Full details may be found elsewhere.[15,16] Then we describe the changes made to produce the CV.

Summary of Mogul. Mogul contains four data libraries, one each for bond lengths, valence angles, torsion angles, and unfused, unbridged rings. Taking torsion angles as an example, building the library involves finding every acyclic torsion-angle fragment (W−X−Y−Z, where X−Y is acyclic) in the CSD. Each is described by a set of keys that collectively characterize the chemical nature of the fragment and its immediate intramolecular environment. For example, there are keys denoting the atomic numbers of the fragment atoms, keys capturing the types of bonds that each atom forms, and so on. Keys are first evaluated as character strings (e.g., the key "1.1.2" represents the bond types of an atom forming two single bonds and one double bond) and then hashed to integers. The fragments are divided into subsets, or "distributions", with fragments that have the same key values being assigned to the same distribution. The observed torsion angles of the fragments are stored in the distribution, together with information that allows each fragment to be retrieved and viewed in the Mogul graphical interface. A tree is constructed, with each level indexed on one of the keys. The leaves of the tree point to the distributions. Both the tree and the distributions are stored as binary files.

To perform a search, the key values of a query fragment are evaluated and used to traverse the tree. If a leaf exists with key values identical to those of the query, the distribution to which it points matches the query exactly at the level of substructural precision encoded by the key set. Retrieval of the distribution therefore gives rapid access to the observed geometries of the query fragment in CSD structures. It is possible that an exactly matching leaf does not exist or does exist but corresponds to a distribution containing too few observations (determined by a user-defined minimum acceptable value, MINOBS). In this case, the exact search has failed and a "generalized search" is performed. This involves back-tracking one level up the tree from the point of failure and collecting all of the distributions below the node thus reached. A similarity calculation is performed on the key values of each of these distributions to determine whether they are sufficiently similar to the key values of the query. Distributions passing this test are retained and pooled. If the pooled distribution still contains too few observations, more back-tracking is allowed until the required number of observations is achieved or a tree level is reached beyond which further generalization is not considered meaningful (in which case the search fails). Tree levels close to the leaves are indexed on the least chemically important keys, so back-tracking generalizes the least significant chemical features first.

Improving Search and Retrieval Times. Several changes were made in order to provide a tool that could be applied to the high-throughput generation of conformers. The CV libraries were limited to organic fragments (no metal within one bond of any atom in the fragment). Only molecular dimensions were stored in distributions, the details enabling fragment display being omitted. This means that the fragments contributing to a CV distribution cannot be viewed, but this is irrelevant in the contexts in which the CV will be used. Distributions were reduced in size by restricting them to a maximum of 500 observations for bond lengths, 250 for valence angles, and 1000 for torsion angles (more precisely, "rotamers"; see below). Random sampling was used where necessary to reduce distribution sizes to these values. Mogul allows distributions to contain up to 10 000 observations, but we show later that this limit is unnecessarily high. A maximum of 500 observations is allowed in a ring distribution in both Mogul and the CV.

Finally, the similarity-based generalization method was replaced because it can sometimes be slow when many distributions have to be pooled. We implemented a simpler approach, "cascade searching", which is always fast. For each fragment type, we constructed not one library but several. The first library is based on an extensive set of keys that describes fragment substructural environments to a relatively high degree of precision. Each subsequent library uses a less extensive key set than the preceding one, corresponding to progressively cruder representations of fragment environments. At search time, an exact search is performed on the first library. If it fails to find a distribution with ≥MINOBS observations, an exact search is made on the second library, and so on, until the specified MINOBS is satisfied or no further libraries remain. Only in the latter case does the search fail. In all other cases, the retrieved distribution is taken from the most chemically precise library possible given the constraint imposed by MINOBS. The number of failures can be minimized by ensuring that the last library in the sequence has a very small key set, corresponding to a coarse-grained fragment classification scheme. Although several exact searches may need to be performed in a cascade search, each one is fast.

Bond-Length and Valence-Angle Libraries. The one bond-length library in Mogul was replaced by four libraries (BOND1 to BOND4) for searching in cascade fashion. The fragment properties captured by the key sets used in these libraries are summarized in Table 1. Similarly, four bond-angle libraries (Table 2) were built. The key sets used in BOND2 and ANGLE2 correspond to those used in Mogul.

Rotamer Libraries. The conformation around the rotatable bond X−Y in the fragment $R_A(R_B)X-Y(R_C)R_D$ can be defined by any of the torsion angles $R_A-X-Y-R_C$, $R_A-X-Y-R_D$, $R_B-X-Y-R_C$, and $R_B-X-Y-R_D$. Each of these is considered a separate torsion-angle fragment in Mogul. Therefore, if X−Y is in a CSD crystal structure, it contributes to four distributions in the Mogul torsion-angle library. If it is in a query molecule, four distributions are retrieved. Furthermore, the keys used in the Mogul torsion library capture more information about the atoms defining the torsion angle than the other atoms bonded to X and Y: for example, there is more information about R_A and R_C than about R_B and R_D for the torsion-angle fragment $R_A-X-Y-R_C$. As a consequence, the four separate distributions are based on different crystallographic observations, though some overlap is likely. This would create extra work for a conformer-generation algorithm using the library because the

Table 1. Properties of Bond-Length Fragment X−Y Captured by Key Sets Used in the Four Bond-Length Libraries

property	whether included in keys of BOND library			
	1	2	3	4
X−Y bond type	yes	yes	yes	yes
atomic numbers of X and Y	yes	yes	yes	yes
bond types of all bonds formed by X and Y	yes	yes	yes	no
hydrogen counts of X and Y	yes	yes	yes	no
size of the smallest ring containing X−Y[a]	yes	yes	yes	no
atomic numbers of all atoms bonded to X and Y	yes	yes	yes	no
coordination numbers of all atoms bonded to X and Y	yes	yes	no	no
hydrogen counts of all atoms bonded to X and Y	yes	yes	no	no
bond types of all bonds formed by atoms bonded to X and Y	yes	yes	no	no
atomic numbers, coordination numbers, and hydrogen counts of all atoms bonded to atoms bonded to X and Y	yes	no	no	no

[a] Set to 0 if no ring or ring size ≥ 9.

probability of any hypothesized geometry around X−Y would be a function of all four distributions.

We therefore defined a new type of fragment, the rotamer. In a rotamer library, each rotatable bond in the CSD contributes to just one distribution, and only one distribution per rotatable bond is retrieved in a search. The key sets used capture an equal amount of information about all atoms bonded to the atoms forming the rotatable bond. Table 3 summaries the keys used to generate the seven rotamer libraries that were built. The IUPAC Gold Book definition of a rotamer is “one of a set of conformers arising from restricted rotation about one single bond”.[31] In practice, the term is often further restricted to acyclic bonds only. However, we included in our rotamer libraries all single, double, and aromatic bonds, whether acyclic or cyclic. This broad coverage enables modeling of the small twists that may occur around strained double and aromatic bonds. As many of the rotamer fragments in our libraries are comparatively rigid, we avoid the potentially misleading term “rotatable bond” when discussing these fragments below. Instead, we use the term “central bond”. The atoms forming the central bond are “central atoms”, and other atoms to which they are bonded are “connected atoms”.

In rotamer libraries, it is still necessary to use a torsion angle to define the geometry about the central bond. We call this the “reference torsion”, and it must be chosen consistently to ensure that the same reference torsion is selected for all rotamer fragments in a given distribution. The choice must therefore be based solely on the information captured in the keys, which is the only information guaranteed to be constant for all fragments in the distribution. As an example, suppose that the keys for the connected atoms in $R_A(R_B)X{-}Y(R_C)R_D$ capture only atomic number (ATNO) and the type of bond that the atom forms to the attached central atom (BT). The algorithm for selecting the reference torsion uses this information to rank the connected atoms in such a way that they have identical ranks only if they have identical key values. Considering the connected atoms bonded to X, a suitable ranking procedure would be the following:

```
if ATNO($R_A$) > ATNO($R_B$) then RANK($R_A$) = 1 and RANK($R_B$) = 2
else if ATNO($R_B$) > ATNO($R_A$) then RANK($R_B$) = 1 and RANK($R_A$) = 2
else if BT($R_A$) > BT($R_B$) then RANK($R_A$) = 1 and RANK($R_B$) = 2
else if BT($R_B$) > BT($R_A$) then RANK($R_B$) = 1 and RANK($R_A$) = 2
else RANK($R_B$) = 1 and RANK($R_A$) = 1
```

where bond types are ranked as aromatic > triple > double > single. The atom with the best rank is chosen for the reference torsion. In the event of a tie, it is permissible to choose arbitrarily from the tied atoms unless the central atom to which they are bonded is prochiral, in which case it is necessary to use stereochemistry as a tiebreaker (see below). All of the connected-atom properties captured by the keys must be included in the ranking procedure, and a tie is declared only if they are all equal.

The order in which the connected-atom properties are tested, and whether larger property values give rise to better or worse ranks, can be chosen arbitrarily. If the choice is changed, leading to a different reference torsion, the resulting distribution will be almost unaltered except that it will be shifted along its axis. This is the case because the difference between the torsion angles of two alternative choices of

Table 2. Properties of Bond-Angle Fragment X−Y−Z Captured by Key Sets Used in the Four Valence-Angle Libraries

property	whether included in keys of ANGLE library			
	1	2	3	4
X−Y and Y−Z bond types	yes	yes	yes	yes
atomic numbers of X, Y, and Z	yes	yes	yes	yes
coordination numbers of X, Y, and Z	yes	yes	yes	yes
bond types of all bonds formed by X, Y, and Z	yes	yes	yes	no
hydrogen counts of X, Y, and Z	yes	yes	yes	no
size of the smallest ring containing X−Y−Z[a]	yes	yes	yes	yes
size of the smallest ring containing Y[a]	yes	yes	yes	yes
atomic numbers of all atoms bonded to X, Y, and Z	yes	yes	yes	no
coordination numbers of all atoms bonded to X, Y, and Z	yes	yes	no	no
hydrogen counts of all atoms bonded to X, Y, and Z	yes	yes	no	no
bond types of all bonds formed by atoms bonded to X, Y, and Z	yes	yes	no	no
atomic numbers, coordination numbers, and hydrogen counts of all atoms bonded to atoms bonded to X, Y, and Z	yes	no	no	no

[a] Set to 0 if no ring or ring size ≥ 9.

Table 3. Properties of Rotamer Fragment X−Y Captured by Key Sets Used in the Seven Rotamer Libraries

property	whether included in keys of ROTAMER library						
	1	2	3	4	5	6	7
size of the smallest ring containing X−Y[a]	yes	yes	yes	yes	yes	yes	yes
bond type of X−Y	yes	yes	yes	yes	yes	yes	yes
coordination numbers and hydrogen counts of X and Y	yes	yes	yes	yes	yes	yes	yes
bond types of all bonds formed by X and Y[b]	yes	yes	yes	yes	yes	yes	yes
atomic numbers of X and Y	yes	yes	yes	yes	yes	yes	no
size of the smallest ring containing X and the smallest ring containing Y[c]	yes	yes	yes	yes	yes	yes	no
relative handedness of X and Y[d]	yes	yes	yes	yes	yes	no	no
atomic numbers of all atoms bonded to X and Y	yes	yes	yes	yes	yes	no	no
coordination numbers of all atoms bonded to X and Y	yes	yes	yes	yes	no	no	no
hydrogen counts of all atoms bonded to X and Y	yes	yes	yes	no	no	no	no
bond types of all bonds formed by atoms bonded to X and Y[b]	yes	yes	no	no	no	no	no
atomic numbers, coordination numbers, and hydrogen counts of all atoms bonded to atoms bonded to X and Y	yes	no	no	no	no	no	no

[a] Set to 0 if no ring or ring size ≥ 25 (for libraries 3−7), ≥ 9 (for libraries 1 and 2). [b] Bond-type keys also discriminate between cis and trans double bonds. [c] Set to zero if no ring or ring size ≥ 9. [d] Set to 1 if X and Y have opposite handedness (one left, one right; see the text), otherwise set to 0.

reference torsion will be approximately constant from one structure to the next. It will vary slightly because of changes in the valence angles at the two central atoms, but this will be unimportant compared with the variance of the rotamer distribution. However, we do try to avoid selecting hydrogens as reference atoms because in X-ray structures they usually have appreciably higher positional uncertainties than non-hydrogen atoms. This is done by linking higher atomic numbers to better ranks.

When the central bond is cyclic, an extra rule is invoked requiring that both connected atoms defining the reference torsion belong to the ring that contains the central bond (or the smallest ring if the central bond is in more than one). For central bonds in two rings of equal size, this could potentially result in different reference torsions being chosen depending on which end of the reference torsion is chosen first. Therefore, the central atoms are also ranked. If they have different ranks, the selection of atoms for the reference torsion starts with those that are bonded to the central atom of higher rank. A final rule, for double bonds, ensures that a cis reference torsion is chosen in preference to a trans one, provided that it can be achieved using non-hydrogen atoms.

Rotamer Chirality. The torsion library in Mogul does not take account of chirality, with all distributions assumed to be symmetric about 0°. Therefore, each distribution is only presented in the range 0−180°. We removed this limitation in the rotamer libraries so that all of the distributions span a full 360° range and are asymmetric about 0° if either central atom is chiral. Whether a central atom is considered chiral depends on the key set being used. Specifically, even if it is chiral in the molecule as a whole, a central atom is deemed achiral if it is bonded to at least two connected atoms that give rise to identical key values (or, equivalently, are assigned equal ranks by the algorithm used to select the reference torsion). This is done because such atoms are indistinguishable at the level of substructural precision captured by the keys. Therefore, a central atom is chiral in two circumstances: (a) when it is four-coordinate and the three connected atoms to which it is bonded all have different ranks or (b) when it is three-coordinate and nonplanar (on the basis of a user-specified tolerance) and the two connected atoms to which it is bonded have different ranks.

If a chiral central atom, X, is detected, its handedness is determined. With Y defined as the other central atom, R_1 and R_2 defined as the connected atoms bonded to X that have the highest and second-highest ranks, respectively, and **a**, **b**, and **c** defined as the vectors from X to Y, X to R_1, and X to R_2, respectively, the vector product $\mathbf{v} = \mathbf{a} \times \mathbf{b}$ is computed, and the handedness of X deemed to be right (left) if the scalar product $\mathbf{v} \cdot \mathbf{c}$ is positive (negative). The overall handedness of a chiral rotamer fragment is then defined as the handedness of its highest-ranked chiral central atom; if the central atoms are of equal rank, either of them can be chosen arbitrarily. During library building, if a left-handed rotamer fragment is encountered, the torsion angle stored for that fragment in the distribution is reversed in sign. In the final library, therefore, the torsion angles of distributions pertaining to chiral rotamers all relate to right-handed fragments. At search time, if the query fragment is left-handed, the signs of all torsion angles in its retrieved distribution are reversed. Rotamer fragments that are diastereomeric are assigned to different distributions by the use of a "relative handedness" key (see Table 3, footnote *d*).

When a nonplanar central atom is bonded to exactly two connected atoms that have identical ranks (i.e., the central atom is prochiral at the level of structural precision captured by the keys), choosing one of them arbitrarily for the reference torsion would lead to meaningless distributions. Instead, the handedness of each of the identically ranked connected atoms is determined, and the right-handed one is used for the reference torsion.

Figure 1 shows three examples of the treatment of rotamer chirality. The top molecule is taken from the CSD entry ISACAG, and we focus on the conformation around the N−C(sp^3) bond. The ROTAMER1 library keys detect the fact that the C(sp^3) atom is bonded to two different types of carbon (methyl and phenyl), and therefore, the atom is treated as chiral. As a consequence, the retrieved distribution is asymmetric about 0° (top histogram, blue bars). The reference torsion is actually C(methyl)−C(sp^3)−N−C, but the distribution in the figure has been transformed to the H−C(sp^3)−N−C torsion to assist a later comparison in this paper. The histogram shown as red bars is the symmetrical distribution obtained if the chirality is ignored (normalized to the same area as the original histogram), and it is obviously very different. The middle molecule in Figure 1 is L-alanyl-L-alanine (CSD reference ALAALA), and the distribution shown next to it (retrieved from the ROTAMER2 library) is of the (O═)C−N−C−C(═O) torsion angle. This is equivalent to the ϕ angle

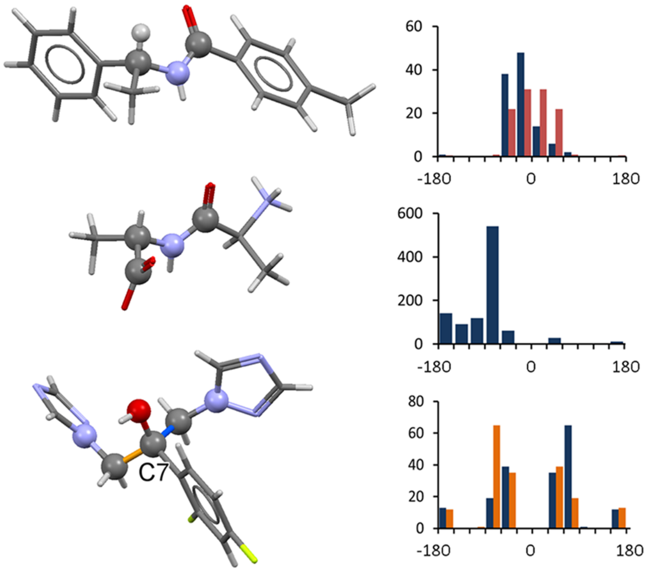


Figure 1. CSD molecules used to illustrate rotamer chirality. Atoms shown as spheres indicate the torsion angles discussed in the text. Top: geometric preferences of the C−N−C−H torsion angle in ISACAG. The blue histogram shows the distribution when chirality is taken into account, and the red histogram shows the distribution made symmetrical about 0°. Middle: distribution of the (O=)C−N−C−C(=O) torsion angle in ALAALA. Bottom: distributions of the two N−CH_2−C−O torsion angles in MEWTAL, one shown in blue, the other in orange.

of a Ramachandran plot, and the expected pronounced asymmetry is clearly seen.

Finally, the bottom molecule is the cytochrome P450 inhibitor fluconazole (CSD code MEWTAL). This is particularly interesting because the carbon atom bearing the hydroxyl group (C7) is achiral, being bonded to two identical triazolylmethyl groups. However, each of the N−CH_2−C−O rotamers is considered chiral because the ROTAMER1 library keys detect the chemical difference between the sp^2 and sp^3 carbons bonded to C7. This correctly results in an asymmetric torsion distribution. Moreover, the perceived handedness of C7 is different depending on which of the two N−CH_2−C−O torsions is considered, resulting in mirror-image distributions. Thus, the distribution shown in blue bars in the bottom histogram in Figure 1 is retrieved for the rotamer involving the central bond shown in blue, while that shown in orange is retrieved for the other N−CH_2−C−O moiety. According to both distributions, the relevant nitrogen atom has a small but distinct preference to be gauche to the phenyl carbon and anti to the methylene group, which would correspond to a positive N−CH_2−C−O torsion angle for one (the blue histogram) and a negative angle for the other. This conformational preference was confirmed by searches of the CSD using the ConQuest[32] program. However, in MEWTAL, only one of the rotamer fragments (the one with the blue central bond) adopts the preferred geometry gauche to the phenyl group, the other being anti.

Rotamer Symmetry. A further improvement over Mogul is that rotamer symmetry is taken into account. A four-coordinate central atom is considered threefold-symmetric if all of the connected atoms to which it is bonded are assigned the same rank by the ranking procedure used to select reference torsions. A three-coordinate central atom is considered twofold-symmetric if it is planar (within a user-specified tolerance) and bonded to connected atoms of equal rank. If at least one of the central atoms of a rotamer fragment is threefold-symmetric, every torsion angle τ in the distribution is used to generate two additional values, $\tau + 120°$ and $\tau - 120°$. If at least one of the central atoms is twofold-symmetric, each τ value generates a further value at $\tau + 180°$. Finally, if the fragment is achiral (i.e., neither central atom has a handedness), mirror symmetry about 0° is imposed by generating a value at $-\tau$ for every value τ. The resulting distribution is then consistent with the symmetry of the rotamer fragment. During cascade searching, values generated by symmetry are ignored when the number of observations in a distribution is counted to determine whether it is less than MINOBS.

Ring Libraries. Of the four ring libraries that were built (Table 4), the first three include both ring-atom positions and those of any non-hydrogen atoms attached to the ring, while the fourth (RING4) contains only ring-atom positions. Whereas the ring library in Mogul contains only rings of size ≥5, the CV libraries also include three- and four-membered rings. Although the former must be planar, the (ring-atom)−(substituent-atom) vectors are of interest. We filtered out aromatic and other delocalized rings in which all of the ring and substituent atoms are likely to be exactly or very nearly coplanar. This reduces the library sizes and search times without loss of any significant information. Specifically, rings were excluded if they (a) contained ≤6 atoms; (b) contained no atoms other than three-coordinate carbon or nitrogen or two-coordinate nitrogen, oxygen, or sulfur; and (c) were likely to be fully delocalized (each ring atom forming at least one unsaturated bond or both of its neighbors in the ring forming unsaturated bonds, e.g., the nitrogen of pyrrole and the oxygen of oxadiazole).

Table 4. Ring-Fragment Properties Captured by Key Sets Used in the Four Ring Libraries

property	whether included in keys of RING library			
	1	2	3	4
ring size	yes	yes	yes	yes
ring-atom types[a]	yes	yes	yes	yes
ring-atom coordination numbers	yes	yes	yes	yes
ring bond types	yes	yes	yes	yes
substituent orientation	yes	yes	yes	no
substituent size[b]	yes	yes	yes	no

[a] Classified using the Sybyl atom-typing scheme,[33] except for three-coordinate nitrogen, which is classified as "non-conjugated" (bonded only to saturated neighbors) or "possibly conjugated" (anything else). [b] For a substituent atom A bonded to a ring atom R, with X representing any non-hydrogen atom, substituent size is classified as "very large" (R−AX_nH_m, where $n + m = 3$, $n \geq 2$), "large" (R−AX_2), "small" (any other non-hydrogen substituent), or "hydrogen" in RING1; as "large" (R−AX_nH_m, where $n \geq 2$, m = any value), "small" (any other non-hydrogen substituent), or "hydrogen" in RING2; and as "non-hydrogen" or "hydrogen" in RING3.

The key set of each library captures information about ring size, ring-atom atomic numbers and coordination numbers, ring bond types, and (except for the RING4 library) the size and orientation of the substituent(s). The classification used in RING2 is the closest to that used in Mogul, although it is not identical because Mogul does not use a bond-type key. This creates occasional misclassifications. For example, the eight-membered rings in the CSD entries ABEFOD and YALHAV are assigned by Mogul to the same distribution, despite the fact that one has an intra-annular double bond and the other an exocyclic double bond (Figure 2). The use of the ring-bond-type key causes them to be assigned to different distributions in the CV.

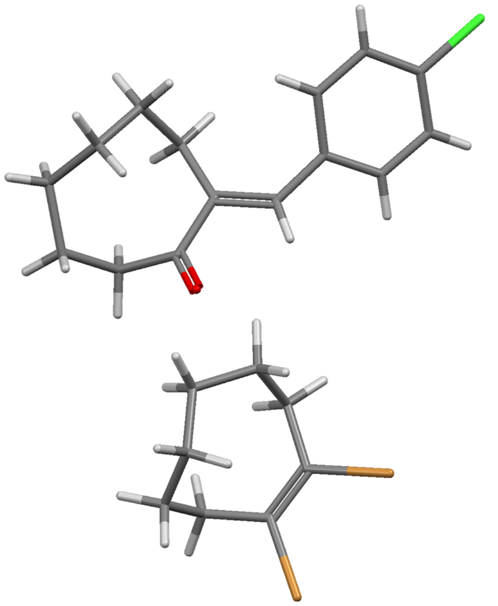


Figure 2. The eight-membered rings in CSD entries ABEFOD and YALHAV are assigned to the same distribution in Mogul but different distributions in the CV.

Substituent size is classified with different degrees of precision in the various libraries (see Table 4, footnote *b*). The orientation(s) of the substituent(s) on each ring atom is (are) evaluated with respect to the plane defined by the two adjacent ring bonds and classified as either "above", "below", or, for substituents on sp^2-hybridized ring atoms, "in plane". The direction corresponding to "above" is chosen arbitrarily but is consistent for all atoms in the ring. Substituents on three-coordinate, nonplanar nitrogen ring atoms are classified as having "uncertain" orientation to reflect the fact that the nitrogen may able to invert.

Ring Chirality. Whether a ring atom is deemed to be chiral depends on the keys being used. A four-coordinate ring atom is considered achiral if it is bonded to substituents belonging to the same size classification. If it is bonded to differently sized substituents, it is deemed chiral as long as the ring itself is asymmetric, on the basis of only those properties of the ring atoms and bonds that are captured in the key set. Nonplanar three-coordinate atoms such as sulfone sulfur are deemed chiral, but not three-coordinate nitrogen, which is assumed to be capable of inverting. All ring atoms in the RING4 library are considered achiral because its key set contains no substituent information. A ring is considered chiral if it contains any chiral atoms. A standard chirality is defined for any distribution of chiral rings. During library building, all rings in the distribution are set to the same chirality, inversion being applied if necessary. At search time, all rings in the retrieved distribution are inverted if the query ring has the nonstandard chirality.

Ring Symmetry and Clustering. A distribution in the Mogul ring library contains all of the crystallographically independent observations from the CSD of a particular type of ring. However, many of these are likely to be geometrically similar. To avoid having to do cluster analysis on the fly, we wanted the distributions in the CV to contain only geometrically distinct conformations, including any relevant ones generated by symmetry. This was achieved by applying a two-stage process to each distribution during library building. In the first stage, each ring in the distribution is considered in turn. The ring atoms are labeled, each with a unique label. All topologically equivalent ways of assigning the labels are then generated, using the properties captured in the library key set to determine atom equivalence. This effectively generates all of the symmetry-equivalent conformations. For example, the ring at the top left in Figure 3 generates three symmetry-equivalent ones, as explained in the figure caption. The collection of symmetry-equivalent rings is then subjected to cluster sampling. This is a technique for clustering a data set and selecting one item from each cluster, biasing the selection toward items near cluster centers.[34] The algorithm requires as input the dissimilarity of each pair of items. For a given pair of symmetry-related ring conformations A and B, we calculate dissimilarity as the least-squares root-mean-square deviation (RMSD) of the pair when each atom in A is mapped onto the identically labeled atom in B.

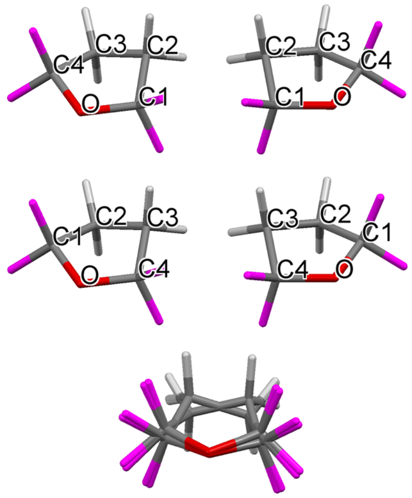


Figure 3. Example of symmetry expansion of ring geometries. The ring at top left was taken directly from the CSD. The magenta atoms represent substituents falling into the same size classification under the key scheme in use for the library being built; white atoms are hydrogen. The other three rings shown with labels are symmetry-generated conformations. The ones on the right are generated from those on the left by inversion, which is allowed because the ring is achiral at the level of substituent classification being used. Each ring at the top produces the ring below it by swapping of the labels of equivalent ring atoms. The superposition at the bottom is produced by least-squares overlay of each of the three symmetry generated rings onto the original, always pairing atoms with identical labels. The four rings fall into two geometrically distinct groups. One representative from each group would be selected by cluster sampling.

In the second stage, all of the ring geometries selected by the cluster samplings in the first stage are pooled. If the ring has *x*-fold topological symmetry and the original distribution contains *M* crystallographically independent ring geometries, the pooled collection will contain between M and *xM* geometries. Cluster sampling is applied to the pooled collection using the same dissimilarity measure as in the first stage. The representative ring geometries thus selected, one per cluster, constitute the final distribution to be stored in the library. Also stored is the "weight" of each geometry, which is proportional to the size of the cluster that the stored geometry represents. Larger weights therefore indicate more popular geometries. The second stage usually reduces the number of retained geometries to a small number, often three or less, representing the distinct conformations that the ring can adopt, including symmetry-generated forms if they are sufficiently geometrically distinct. The RING4 library typically has larger distributions than the other libraries because it does not take substituents into account, which tends to increase the number of topologically equivalent ways of labeling the ring atoms. The number of observations in a ring distribution is deemed to be the number of crystallographically independent rings that were in the distribution before clustering, not the number of ring geometries after clustering.

Fused Rings. The ring library in Mogul contains only simple rings (rings that are neither fused nor bridged). In contrast, the libraries in the CV also contain fragments derived from fused (but not bridged) rings. Each component ring of a fused-ring system constitutes a separate fragment; for example, a steroid comprises four fragments, each belonging to a separate distribution. The bond key of each fragment of a fused-ring system captures the position(s) of the fusion bond(s) and the size(s) of the fused rings.

Ring Templates. When performing conformational analysis on a molecule containing a flexible ring, it may be helpful to identify the most likely conformations of the ring and build a model ("template") of each. For an unfused ring, the ring geometries in a distribution retrieved from the CV are directly usable as templates, as the distribution has been preclustered, taking account of any symmetry. As each geometry in a distribution has a weight proportional to how often it occurs in the CSD, higher-weight templates can be sampled preferentially. Templates retrieved from three of the ring libraries (RING1, RING2, and RING3) include substituent positions, which therefore define the (ring-atom)−(substituent-atom) vectors to be used when splicing the template into the query molecule. This vector information is not available in the RING4 library, so substituents are added to any template from this library so as to reproduce (ring-atom)−(ring-atom)−(substituent-atom) valence angles in the query molecule. Hydrogen atoms are added to ring templates in calculated positions using mean bond lengths and angles determined from neutron diffraction data.

The construction of templates for fused ring systems can usually be achieved, provided that templates are available for the component rings that make up the fused system. For example, consider the two component-ring templates shown at the top of Figure 4. A rigid transformation is applied to one of them using the rotation−translation operator obtained by least-squares fitting of the atoms that are common to both templates. There will usually be at least six of these: the two atoms of the fusion bond and the two adjacent atoms in each ring (the number may be less if either or both of the rings is three-membered). After the transformation is applied (Figure 4, middle), the position of each of the common atoms is set to the average of its positions in the two components, resulting in a template for the fused-ring system (Figure 4, bottom). The averaging introduces distortions of the local bond lengths and angles, but they are usually small, except when the conformations of the two component-ring templates are mutually incompatible. In this case, the RMSD from the least-squares fit of the common atoms will be high. Therefore, we reject the fused-ring template if the RMSD exceeds 0.3 Å.

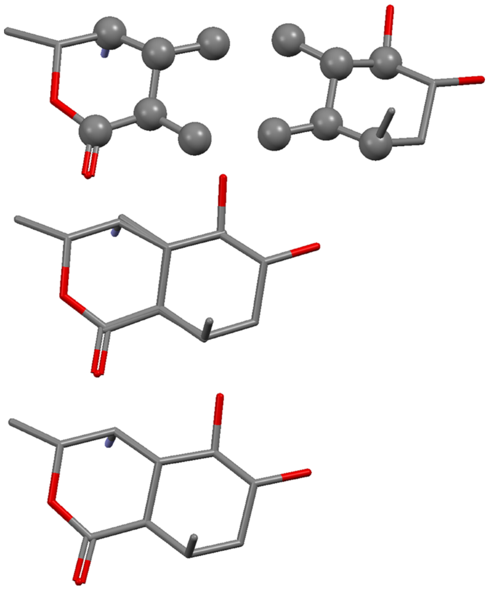


Figure 4. Construction of a template for a fused-ring system. Top: separate templates for the component rings, with atoms in common shown as spheres. Middle: templates have been superimposed by least-squares fitting of the atoms in common. Bottom: final template produced by taking average positions for the atoms in common.

If there are two or more templates for either or both of the component rings, all combinations are tried, with each successful combination resulting in a new template for the fused-ring system. Templates for fused-ring systems containing three or more component rings can be built in a stepwise fashion. The algorithm also works for spiro-fused systems. Building fused-ring templates from the templates of the component rings affords the possibility of generating templates for a fused-ring system that is not present in the CSD, provided that each of the component ring fragments is present.

Library Optimization. Experiments were performed to determine (a) the optimum MINOBS values for cascade searching, (b) whether it is better to build libraries from the complete CSD or a subset of high-quality entries, and (c) whether it is necessary to use all of the libraries listed in Tables 1−4 or whether some can be dropped because they add little value. The results (available in the Supporting Information) guided us in our choice of parameters and libraries for the validation described below.

■ VALIDATION

Validation Data. For the validation, the CV libraries were built from CSD entries published before 2011. Bond-length libraries were filtered (*R* factor ≤ 5%; only molecules with ≥15 non-hydrogen atoms to exclude ill-determined solvates). All others were unfiltered. The ROTAMER2 and RING1 libraries were dropped because the optimization results suggested that they add little or no value. MINOBS was set to the values suggested by the optimization: 5 for bond lengths and angles, 20 for rotamers, and 5 for rings. The average value of a bond length or angle distribution was measured by the interquartile mean, which is based only on observations between the lower and upper quartiles. The optimization suggested that this was necessary to remove the effects of any gross outliers that might be present. Only ring templates with ranks of ≤5 were used because it is rare for a template of worse rank to give the best match to an observed ring geometry.

Tests were performed on a set of 3636 druglike molecules (the "validation set") taken from CSD structures published in or after 2011 with an R factor of ≤5% and no disorder (see the Supporting Information for a list of the CSD reference codes and the definition of "druglike"). The validation results are therefore a true measure of predictive ability, as the crystal structures of all of the validation-set molecules were published after the latest CSD entry contributing to the libraries.

Success Criteria. With Δ defined as the absolute difference between an observed bond length or valence angle and the interquartile mean of the corresponding CV distribution, the success criterion for the bond-length and valence-angle libraries was

$$\mathrm{RMSD}(\Delta) = \sqrt{\frac{\sum_i \Delta_i^{\,2}}{n}}$$

where the summation is over the bonds or angles included in the validation, excluding those for which no distribution was successfully retrieved, and n is the number of such bonds or angles.

A different success criterion was required for rotamer distributions because they tend to be multimodal, making statistics based on sample means unsuitable. The distribution retrieved for any given rotamer was represented as a histogram using twelve 30° bins collectively spanning a full 360° range (specifically, −195 to +165°, so that −180° and 0° fall in bin centers rather than on bin boundaries). The "occupied bin" (i.e., the bin to which the observed value of the rotamer reference torsion angle belonged) was found. BINRANK was then defined as the number of bins in the histogram whose counts were equal to or greater than that of the occupied bin, including the occupied bin itself. BINRANK must therefore fall in the range 1 to 12. The success measure was the mean value of BINRANK over the rotamers included in the validation, termed μ(BINRANK). Any rotamer for which no distribution could be retrieved (this was very rare) was assigned a default uniform distribution comprising 12 bins with equal counts. Such rotamers were included in the calculation of μ(BINRANK).

For ring libraries, the templates retrieved for a given ring or fused-ring system were ranked on the basis of their weights, a rank of 1 indicating the template with the highest weight. Each template was overlaid on the observed ring (non-hydrogen substituents included) by least-squares, and the one giving the lowest RMSD was accepted. For rings with topological symmetry, only one atom pairing was used for the RMSD calculation, since any necessary symmetry-generated geometries were present as templates in the distribution (see Ring Libraries). The success of a test was judged by the mean RMSD of the accepted templates, termed μ(RMSD).

In addition to the primary success criteria defined above, the median and 95th percentile of Δ, BINRANK, and ring RMSD were also calculated, along with the mean rank of the ring template giving the best RMSD.

Validation Results. Cascade searches of the libraries described above were performed for (a) all bond lengths and angles in the validation set except those involving hydrogen; (b) all acyclic and cyclic rotamers in which the central bond is single, excluding hydrogen rotors such as R−CH_3; (c) the subset of those acyclic rotamers that have no symmetry (e.g., excluding rotamers such as R−CF_3); and (d) all simple rings and fused-ring systems, excluding those expected to be planar (determined as described in Ring Libraries). The results are summarized in Tables 5−7 (the column headed "n" shows the numbers of fragments on which the results are based). Also given in the tables are the percentages of fragments for which the CV failed to produce a result because no distribution containing ≥MINOBS observations could be retrieved (% failure). Figure 5 shows the distributions of Δ for bond lengths and angles. Figure 6 shows the distributions of BINRANK for acyclic rotamers and for acyclic rotamers that have no symmetry. Figure 7 shows the distributions of the RMSD between the observed ring and the best retrieved template for simple rings and fused-ring systems.

Table 5. Summary of Validation Results for Bond Lengths and Valence Angles

fragment type	n	% failure	RMSD(Δ)	median Δ	95th percentile of Δ
bond lengths	89083	0	0.0118 Å[a]	0.0058 Å[b]	0.0224 Å
valence angles	125583	0.05	1.58°[c]	0.63°[d]	3.44°

95% confidence intervals: [a]0.0116−0.0120 Å; [b]0.0058−0.0059 Å; [c]1.56−1.59°; [d]0.63−0.64°.

Discussion. RMSD(Δ) for bond lengths is 0.012 Å; 50% of the bond lengths are estimated within 0.006 Å and 95% within 0.022 Å. However, there are some extreme outliers in the Δ distribution. The largest is 0.258 Å for C16−C17 in CSD entry YEBVIM. We manually inspected the 20 most extreme Δ values, together with other large values chosen at random. The 17 worst outliers are plainly due to experimental errors in the validation set. Even in structures with R factors of ≤5%, badly misplaced atoms occasionally occur, and we believe that many of the Δ values above 0.05 Å are due to validation-set errors. The use of Mogul to identify such errors is perhaps an answer to the view that geometries in small-molecule crystal structures are not easy to validate automatically.[35] The largest Δ that could be ascribed to a poor prediction was 0.203 Å for the C3−C4 bond in the pyrazole ring of CSD structure BEGJUU. This result was based on the lowest-quality bond-length library, BOND4. Such bad predictions are rare.

RMSD(Δ) for valence angles is 1.6°, and the median and 95th percentile are 0.6° and 3.4°, respectively. Again, the tail of the distribution contains some extreme outliers. The largest is 24.7° for C8−C7−C8N in LAZJIH, where the experimentally observed valence angle does not seem credible. Several more of the worst Δ values can safely be ascribed to experimental errors in the validation set. However, a significant number are due to problems with the CV distributions. Often, this is because the distribution is based largely on structures that are in some way systematically different from the query. An example is S1−C7−N1 in AYUGOR (observed 110.4°; CV interquartile mean 125.2°). The problem here is that the CV distribution mainly contains angles from molecules such as GODZAA, where the S−C═N angle is very wide to allow a S═O···H intramolecular hydrogen bond to form (Figure 8). In contrast, the query molecule contains a different type of hydrogen bond, C═N···H, formation of which requires the S−C═N angle to tighten. In essence, the valence angle is highly dependent on the

Table 6. Summary of Validation Results for Rotamers

rotamer type	n	% failure	μ(BINRANK)	median BINRANK	95th percentile of BINRANK
acyclic single bonds	17454	0	2.87[a]	2[b]	8
nonsymmetric, acyclic single bonds	15018	0	2.63[c]	2[d]	8
cyclic single bonds	14966	0.2	1.63[e]	1[f]	4

95% confidence intervals: [a]2.83−2.90; [b]2−2; [c]2.59−2.67; [d]2−2; [e]1.61−1.64; [f]1−1.

Table 7. Summary of Validation Results for Rings

ring type	n	% failure	μ(RMSD) (Å)	median RMSD (Å)	95th percentile of RMSD (Å)	mean rank
simple rings	880	3.0	0.12[a]	0.09[b]	0.30	1.5
fused-ring systems	841	21.0	0.16[c]	0.11[d]	0.43	1.4

95% confidence intervals: [a]0.114−0.127 Å; [b]0.086−0.097 Å; [c]0.151−0.173 Å; [d]0.103−0.118 Å

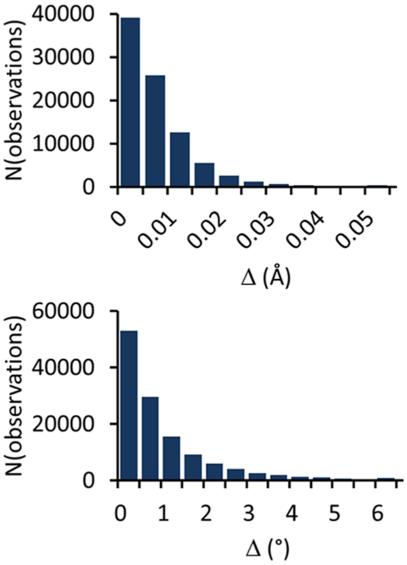


Figure 5. Top: distribution of absolute differences between observed bond lengths and the interquartile mean of the CV distribution (the bin on the extreme right shows the number with Δ > 0.05 Å). Bottom: corresponding distribution for valence angles (the bin on the extreme right shows the number with Δ > 6°). Separate histograms of the upper tails of these distributions are available in the Supporting Information.

molecular conformation and could only be accurately predicted by a bivariate (torsion angle, valence angle) distribution.

Among acyclic rotamers in which the central bond is single, 39.5% have BINRANK = 1, which means that the bin of the 12-bin histogram occupied by the observed rotamer torsion angle is the one with the highest count. However, the true situation is better than this because some of these rotamers are symmetric. Consider, for example, the rotamer R−CF_3. Because it is threefold-symmetric, every value τ in the CV distribution also generates values at τ + 120 and τ − 120°. Therefore, the best possible result for this system is that the count of the occupied bin is one of the three equal-highest in the histogram, corresponding to BINRANK = 3. But the three bins are equivalent, so in conformer generation it would not matter which was sampled. The second row of Table 6 and the bottom histogram of Figure 6 give the BINRANK results for the subset of rotamers without symmetry: 45.8% of the rotamers have BINRANK = 1, 67.6% have BINRANK ≤ 2, and 76.9% have BINRANK ≤ 3.

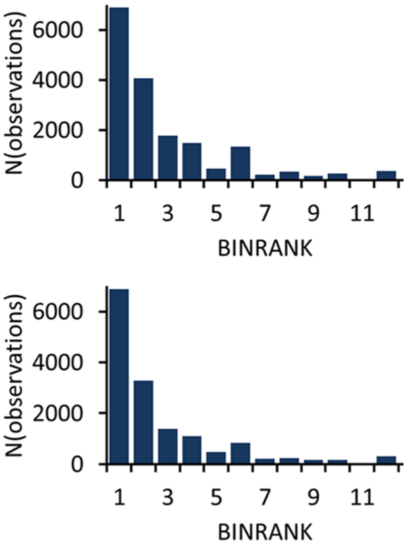


Figure 6. Distributions of BINRANK, the number of bins in a 12-bin histogram with counts greater than or equal to that of the bin containing the observed rotamer reference torsion angle. Top: all acyclic rotamers; bottom: nonsymmetric acyclic rotamers.

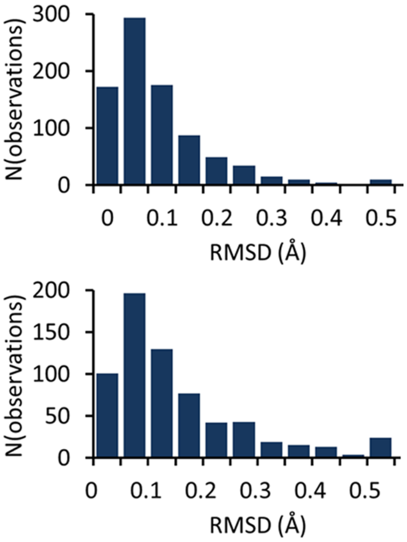


Figure 7. Distributions of RMSD (Å) between the best template and the observed ring. Top: simple rings; bottom: fused ring systems. The bins on the extreme right show the numbers with RMSD > 0.5 Å.

On the other hand, 7.2% of the nonsymmetric rotamers have BINRANK > 6, corresponding to the less-probable half of the CV distribution, and 2.0% have BINRANK = 12, corresponding to the bin with the lowest (or equal-lowest) count. A disproportionate number of the latter group involve distribu-

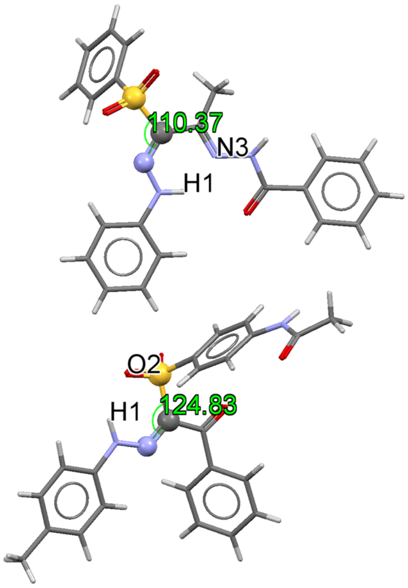


Figure 8. CSD structures AYUGOR (top) and GODZAA (bottom). Atoms involved in the S−C═N valence angles are shown as spheres, and atoms involved in intramolecular hydrogen bonds are labeled.

tions that were retrieved from one of the lower-quality rotamer libraries, implying that they may be poorly predicted because of a lack of experimental observations on rotamers that are very similar to the query. Rotamers that were poorly predicted even when the distribution was retrieved from the highest-quality ROTAMER1 library might be better predicted by a library in which the substructural environment is captured even more precisely. However, only 2.3% of the nonsymmetric rotamers had distributions retrieved from ROTAMER1 and BINRANK > 6, so the possible gains from using more elaborate keys are limited. Moreover, manual inspection of examples suggests that many rotamers with large BINRANK values are genuine outliers caused by unusual intramolecular or crystal-field environments (e.g., high intramolecular steric strain).

Predictions of simple ring geometries have μ(RMSD) = 0.12 Å, with the 95th percentile being 0.30 Å. The average rank of the best template is 1.5, and the failure rate is 3.0%. The mean and 95th percentile RMSD for fused-ring systems are 0.16 and 0.43 Å, respectively. However, the failure rate is over 20%, and we cannot produce templates for bridged-ring systems. Moreover, many of the ring systems in the validation set probably represent easy cases, where there is little flexibility. The success rates would be much lower if we focused on highly flexible rings such as large aliphatic macrocycles. Illustrative examples of retrieved ring templates are given and discussed in the Supporting Information.

Occasionally, ring templates are retrieved with poor geometries as a result of unresolved disorder in the CSD structures. For example, a template with a chemically unreasonable geometry is retrieved for the tetrahydroaminoacridinium system in the ligand of PDB entry 1odc. This ring system contains a partly saturated ring that has a puckered geometry in the PDB structure, as would be expected on chemical grounds. However, the highest-ranked template retrieved for the ring system from the CV is almost flat. It is based on several structures in the CSD in which the authors have not resolved disorder, resulting in an averaged structure with an ostensibly flat geometry (e.g., HAGWUJ; Figure 9). We have established that the problem can be solved by checking the valence angles of ring templates using the valence-angle libraries and eliminating any with at least one unreasonable angle.

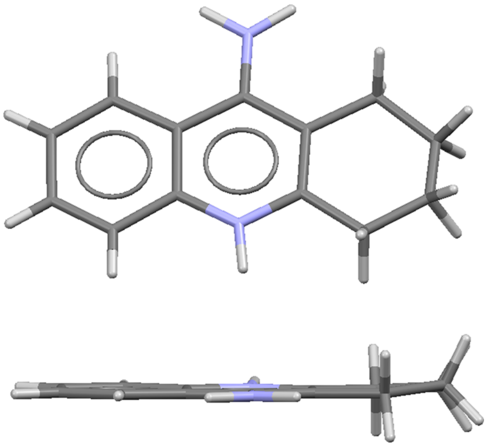


Figure 9. Two views of the tetrahydroaminoacridinium molecule in CSD entry HAGWUJ, showing the apparently flat geometry of the partly saturated ring.

COMPARISON BETWEEN MOGUL AND THE CV

Mogul and the CV were compared to determine (a) the relative effectiveness of the generalization and cascade-search methods of retrieving extra hits when an exact search fails to retrieve a distribution with ≥MINOBS observations, (b) whether the correct handling of rotamer chirality in the CV is advantageous, and (c) relative search speeds. The comparison was based on a set of 1269 druglike molecules (the "comparison set") added to the CSD after version 5.34 was built in 2012 (CSD reference codes available in the Supporting Information). Mogul searches for bond lengths, valence angles, and torsion angles were conducted using libraries built from CSD version 5.34, allowing generalization where necessary and using all program defaults. In particular, MINOBS was 15 for bond lengths and angles and 40 for torsions. For each rotamer fragment, the torsion angle distribution with the largest number of observations was used as the basis for prediction. The CV searches were performed on libraries built from CSD entries published before 2011. Cascade searches were used with the libraries BOND2, BOND3, BOND4, ANGLE2, ANGLE3, ANGLE4, ROTAMER2, ROTAMER4, and ROTAMER6. MINOBS parameters were set to the Mogul default values.

Comparison Results. Results are summarized in Table 8. For bond lengths and valence angles, Mogul and the CV yield very similar RMSD(Δ) values and failure rates. For rotamers, the CV gives a slightly better μ(BINRANK) than Mogul on the full comparison set, but the difference is not statistically significant. The last line of Table 8 gives results for rotamers in the 384 comparison-set molecules that crystallize in Sohnke space groups (i.e., space groups that can accommodate enantiomerically pure compounds). The superior prediction accuracy of the CV is now statistically significant (probability < 0.005, two-tailed unpaired t test). The Sohnke subset almost certainly contains a higher proportion of rotamers involving chiral atoms than the full comparison set. Thus, the result suggests that correct handling of chirality improves the results. The extent to which the improvement has practical relevance was investigated by identifying the rotamer fragments in the

Table 8. Results of the Comparison between Mogul and the CV

statistic	Mogul result	CV result
RMSD(Δ), bond lengths (Å)	0.013	0.014
% failure, bond lengths	0.1	0
RMSD(Δ), valence angles (deg)	1.6	1.7
% failure, valence angles	1.1	0.1
μ(BINRANK), rotamers[a]	3.01	2.95
% failure, rotamers	4.1	0.5
μ(BINRANK), rotamers in Sohnke groups[a]	3.12	2.87

[a]Based only on rotamers for which a distribution with at least MINOBS observations was found [in contrast to Table 6, where rotamers with fewer than MINOBS observations were assigned default uniform distributions and included in the μ(BINRANK) calculation; see Success Criteria].

comparison set for which at least one of the central atoms was considered chiral in the CV library from which the rotamer distribution was retrieved. There are 985 such fragments (on average, therefore, about 0.8 per comparison-set molecule). Of these, more than half (606) have distributions that are highly asymmetric, which means that the highest bar of the 12-bin histogram is more than twice the height of the bar that would be its mirror image if chirality were not taken into account (i.e., if the distribution were made symmetric about 0°). For this subset, μ(BINRANK) is 2.77 for the rotamer libraries and appreciably higher (4.14) for the standard Mogul torsion library. The number of highly asymmetric distributions amounts to about 10% of the acyclic rotamer fragments in the comparison set, suggesting that correct treatment of chirality is likely to be relevant to a significant proportion of molecules.

Search Times. The CV is typically 3 to 7 times faster than Mogul on the basis of elapsed times. For the CV, the average search and retrieval times per molecule are about 0.03 s for bond lengths, 0.06 s for valence angles, 0.07 s for rotamers, and 0.03 s for rings (averaged over all molecules, including those with no rings). These are elapsed times using a 2.3 GHz Intel i7-3610QM processor. However, the times must be interpreted with caution because they are very dependent on disk caching. Search and retrieval was always slower for the first few hundred molecules on the first test run. After that, popular distributions evidently became cached, and the searches accelerated sharply. In short, the method is more efficient if applied to large numbers of molecules (hundreds or more). The average times quoted above relate to tests performed after an initial test run was done to achieve caching. For comparison, the average time per molecule for a bond-length search on the first 25 molecules of the optimization set when no previous run had been done was 0.15 s.

AUTOMATED AND MANUAL APPROACHES TO LIBRARY GENERATION

The libraries described above were created in an automated fashion. This contrasts with the work of some other groups, who compiled libraries of torsion distributions obtained by searching for substructures defined manually by chemical experts.[12,29,30] The question arises as to whether one approach is superior to the other. To gain some insight, we compared distributions retrieved from our libraries with results reported by Schärfer et al.[30] (henceforth SCHÄRFER), who developed the most recent of the hand-crafted torsion libraries.

For the most part, we found good agreement, as illustrated by the following: (a) We concur with SCHÄRFER's analysis of the reported conformations of PDB ligands 3tv7 and 1gwx. (b) The distributions we retrieved for aryl ethers were similar to those shown in SCHÄRFER's Figure 5. For example, distributions were retrieved from our ROTAMER1 library for two rotamer fragments in CSD entry AJUXUZ (Figure 10, top). They are shown in Figure 11 alongside the very similar distributions obtained by searching the CSD with ConQuest for the substructures defined by SCHÄRFER for these types of fragments. (c) The symmetrical distribution shown in red at the top right in our Figure 1 (see Methods) is reasonably similar to the distribution relating to the same type of torsion angle in SCHÄRFER's Figure 4 (although, of course, the actual distribution retrieved from our library is asymmetric about 0° because it respects chirality). (d) Good agreement was found for several aryl amides (SCHÄRFER's Figure 6).

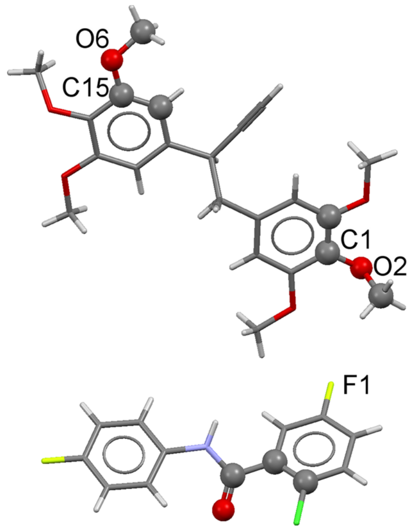


Figure 10. CSD molecules AJUXUZ (top) and ENUKAA (bottom). Atoms shown as spheres indicate torsion angles discussed in the text.

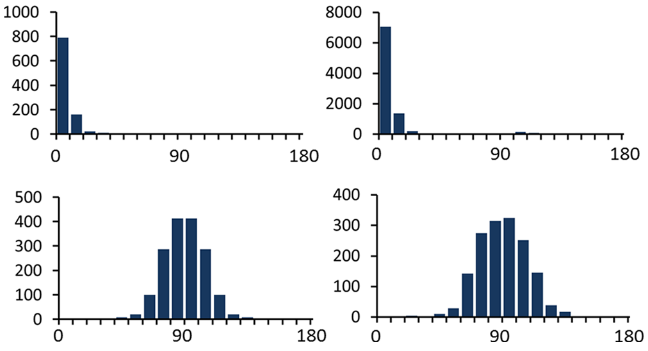


Figure 11. Distributions of torsion angles around C15−O6 (top) and C1−O2 (bottom) of AJUXUZ based on the ROTAMER1 library (left) and a ConQuest search for the substructures defined for these bonds in ref 30 (right).

However, there is an interesting and instructive discrepancy for a rotamer fragment in CSD entry ENUKAA (Figure 10, bottom). When MINOBS is set to 25, the distribution we obtain for this fragment (Figure 12, top) is retrieved from the

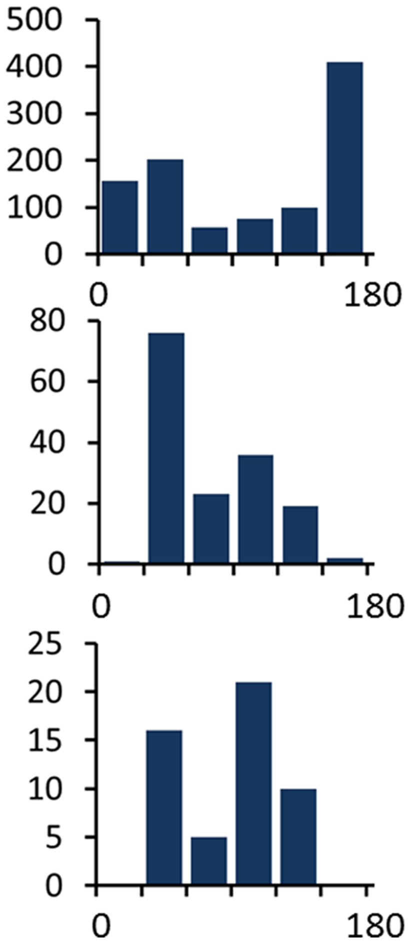


Figure 12. Distributions of the (Cl)C−C−C═O torsion angle in ENUKAA. Top: from the ROTAMER2 library. Middle: from a ConQuest search for the substructure used in ref 30. Bottom: from the ROTAMER1 library after replacement of F1 by a hydrogen atom.

ROTAMER2 library. It shows very poor agreement with the distribution determined manually by searching for the substructure defined for this type of fragment by SCHÄRFER (Figure 12, middle; also see their Figure 6, row a). If, however, the structure is changed by altering the fluorine substituent (F1 in Figure 10) to hydrogen, a distribution with >25 observations is retrieved from ROTAMER1 (Figure 12, bottom). This distribution is in much better (albeit not perfect) agreement with SCHÄRFER's. The explanation is as follows. The ROTAMER2 keys capture the fact that there is a substituent ortho to the amide group but do not capture its identity (i.e., chlorine). The keys for ROTAMER1 do capture the identity of this substituent but also recognize that there is a meta substituent on the phenyl ring (the fluorine). Because of the requirement for a meta substituent, there are insufficient fragments in the ROTAMER1 distribution. The ROTAMER2 distribution is perforce used, which is dominated by fragments containing ortho substituents other than chlorine, many of which induce very different conformational preferences. When the fluorine is changed to hydrogen, a distribution with sufficient observations is obtained from ROTAMER1 that, being entirely based on fragments with ortho chloro substituents, is a much better model for the query. When devising their search substructure for this fragment, SCHÄRFER realized that the nature of the ortho substituent (Cl) is likely to have an important influence on conformation but the presence of a meta substituent will have little or no effect. Therefore, they included the chlorine in the substructure, made no specification about meta substitution, and got good results. This highlights the advantage that can be obtained from a substructure manually defined by experienced chemists.

However, there are arguments in favor of the automated approach. Table 9 gives the numbers of distributions in our rotamer libraries. Even the library with the least extensive key set, ROTAMER7, contains almost 2000 distributions with at least 20 observations (the value of MINOBS that gave the best results in our library optimization). The ROTAMER1 library contains over 40 000 distributions with ≥20 observations, of which over 1500 contain 500 observations or more. It seems unlikely that even the most diligent research group will manually create a library that contains comparable numbers of distributions (e.g., SCHÄRFER's library contains about 500). Moreover, we have clear evidence that subdivision into such large numbers of rotamer types is, on balance, efficacious because the values of μ(BINRANK) systematically improve in the order ROTAMER7, ROTAMER6, ..., ROTAMER1 (Table 9, final column). A further advantage of the automated method is that novel substructures appearing in the CSD will have their geometric preferences captured in the form of new distributions without the need for human intervention.

Table 9. Numbers of Distributions in Rotamer Libraries

	number of distributions with n_{obs} ≥				
library	20	50	100	500	μ(BINRANK)[a]
ROTAMER1	44175	17089	8467	1668	2.66
ROTAMER2	26000	12303	6971	1785	2.81
ROTAMER3	24513	11819	6685	1730	2.85
ROTAMER4	16329	8592	5179	1501	3.15
ROTAMER5	11110	6241	3903	1315	3.22
ROTAMER6	4330	2668	1874	762	3.50
ROTAMER7	1757	1259	952	498	3.56

[a] Based on the validation set using libraries built from pre-2011 CSD entries.

Ultimately, we believe that the best approach is likely to be an automated system providing a robust, comprehensive, and "future-proof" foundation for knowledge-based geometry prediction, supplemented by a smaller, hand-crafted library to deal with fragment types that are not well predicted by the automated system. The latter can be populated on an ad hoc basis as problem fragments are encountered during the practical use of the automated system. We have written the infrastructure that will enable this approach to be implemented. Of course, there will remain problematic molecules. For example, Figure 13 shows two CSD molecules that have unusual torsion angles about one or more rotatable bonds. Their conformations are stabilized by attractive intramolecular interactions: a hydrogen bond in FUGQON and hydrophobic stacking interactions in RAMPEA. The groups involved in the interactions are separated by many chemical bonds, so their effect on conformation is likely to remain unpredictable by either the automated or manual approaches.

■ CONCLUSIONS

Data libraries for predicting the geometric preferences of druglike molecules have been built. They contain distributions of molecular dimensions derived from small-molecule crystal structures and were developed primarily for use in knowledge-

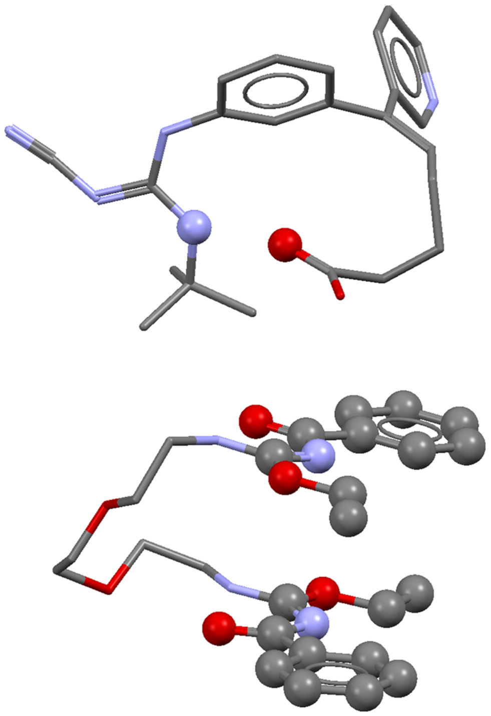


Figure 13. CSD molecules FUGQON (top) and RAMPEA (bottom); hydrogen atoms have been omitted for clarity. Atoms involved in the intramolecular hydrogen bond in FIGQON are shown as spheres. Unusual torsion angles in the linking chain of RAMPEA are probably due to the attractive stacking interaction between the groups shown as spheres.

based conformer generation and geometry optimization. Other possible uses include protein−ligand geometry validation and generation of restraints for structure refinement. The libraries have been validated on 3636 druglike molecules retrieved from CSD entries, all of which were published later than the latest entry contributing to the libraries. Interquartile means of bond length and valence angle distributions from the libraries predict observed lengths and angles in the validation set with RMSDs of 0.012 Å and 1.6°, respectively, and median deviations of 0.006 Å and 0.6°. Torsion angle distributions retrieved for rotatable bonds (one distribution per bond) were represented as histograms using twelve 30° bins. The observed torsion angle belonged in the bin with the highest count in almost 50% of cases and in one of the three highest in about 75%. Stereochemistry is taken into account when building and searching the libraries, so if either of the atoms forming a rotatable bond is chiral, the torsion distribution returned for the bond is asymmetric about 0°. This improves the prediction accuracy for chiral molecules. Ring geometries are predicted with mean RMSDs of 0.12 Å (unfused rings) and 0.16 Å (fused-ring systems). The libraries are fully comprehensive for bond lengths, valence angles, and rotamers and produce templates for the large majority of unfused rings and about 80% of fused-ring systems. However, no templates are stored for bridged rings, and the prediction of macrocycle conformations undoubtedly requires more work. The inclusion of cyclic bonds in the rotamer libraries affords the possibility of approaching this problem by knowledge-based conformational sampling with ring-closure constraints.

We appreciate that some might question our decision to include in the validation set only molecules from small-molecule crystal structures. However, we feel that inclusion of ligands from protein−ligand crystal structures is problematic because it is difficult to determine whether "poor predictions" are due to inappropriate or inaccurate library distributions or experimental ligand geometries with significant errors.[36] Another possibility would be to validate against geometries optimized by quantum-mechanical methods. However, the results would be dependent on the theoretical approximation used, and calculations applying to an in vacuo situation would not necessarily be good guides to condensed-phase conformational preferences.

The libraries and program refinements described herein constitute an extensively customized version of the standard Mogul system that has been distributed for several years by the Cambridge Crystallographic Data Centre. The two versions are complementary. Standard Mogul covers the complete range of chemistry in the CSD, including organometallic systems, and is more transparent than the customized version. Its users are able to inspect distributions and the structures underlying them, exclude outliers, and apply a variety of filters on the fly. In the case of a distribution created by generalized searching, users can control which of the component types of structures contributing to the distribution are included. The customized version does not offer these interactive features but (a) is appreciably faster than standard Mogul, (b) produces only one distribution per rotatable bond (simplifying conformer generation), (c) takes full account of rotamer and ring stereochemistry and symmetry, and (d) provides prebuilt templates, ranked on probability, for both simple and fused rings (standard Mogul can take several seconds to cluster a ring distribution on the fly and does not provide coverage of fused rings). It is particularly suitable for use with algorithms such as conformer generation that may be applied in an automated fashion to millions of molecules. The Mogul libraries developed in this work and a conformer generator based on them have been released to a number of users for beta testing.

A reviewer noted that one use of the rotamer distributions would be to establish rules for typical structural conditions under which a given rotamer would adopt one dominant conformation to the exclusion of all others. Such rules could help correct ill-conceived counts of rotatable bonds when assessing molecular flexibility. We entirely agree. The O═C−O−C torsion angle of acyclic esters, for example, is highly constrained, almost never deviating from 0° by more than 10°, yet it is commonly counted as a "rotatable bond".

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Text files containing lists of the CSD reference codes of the entries from which the optimization, validation, and comparison sets were taken; examples of ring templates; details of the optimization experiments; and histograms of the upper tails of the bond-length and valence-angle Δ distributions from the validation. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: 0044-1923-775972. E-mail: robin@justmagnolia.co.uk.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Colin Groom is thanked for critical reading of the manuscript.

■ ABBREVIATIONS USED

CSD, Cambridge Structural Database; MINOBS, minimum number of observations required for a distribution to be considered usable; RMSD, root-mean-square deviation; BINRANK, rank of the histogram bin containing the observed rotamer torsion angle.

■ REFERENCES

(1) 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.

(2) Dorfman, R. J.; Smith, K. M.; Masek, B. B.; Clark, R. D. A knowledge-based approach to generating diverse but energetically representative ensembles of ligand conformers. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 681−691.

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

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

(5) O'Boyle, N. M.; Vandermeersch, T.; Flynn, C. J.; Maguire, A. R.; Hutchison, G. R. Confab—Systematic generation of diverse low-energy conformers. *J. Cheminf.* **2011**, *3*, 8.

(6) Ebejer, J.-P.; Morris, G. M.; Deane, C. M. Freely Available Conformer Generation Methods: How Good Are They? *J. Chem. Inf. Model.* **2012**, *52*, 1146−1158.

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

(8) Taylor, R. Short Nonbonded Contact Distances in Organic Molecules and Their Use as Atom-Clash Criteria in Conformer Validation and Searching. *J. Chem. Inf. Model.* **2011**, *51*, 897−908.

(9) Allen, F. H.; Kennard, O.; Watson, D. G.; Brammer, L.; Orpen, A. G.; Taylor, R. Tables of bond lengths determined by X-ray and neutron diffraction. Part 1. Bond lengths in organic compounds. *J. Chem. Soc., Perkin Trans. 2* **1987**, S1−S19.

(10) Orpen, A. G.; Brammer, L.; Allen, F. H.; Kennard, O.; Watson, D. G.; Taylor, R. Tables of bond lengths determined by X-ray and neutron diffraction. Part 2. Organometallic compounds and co-ordination complexes of the d- and f-block metals. *J. Chem. Soc., Dalton Trans.* **1989**, S1−S83.

(11) Allen, F. H. The Cambridge Structural Database: A quarter of a million crystal structures and rising. *Acta Crystallogr., Sect. B* **2002**, *58*, 380−388.

(12) Klebe, G.; Mietzner, T. A fast and efficient method to generate biologically relevant conformations. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 583−606.

(13) 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*, 1−24.

(14) Cruz-Cabeza, A. J.; Liebeschuetz, J. W.; Allen, F. H. Systematic conformational bias in small-molecule crystal structures is rare and explicable. *CrystEngComm* **2012**, *14*, 6797−6811.

(15) Bruno, I. J.; Cole, J. C.; Kessler, M.; Luo, J.; Motherwell, W. D. S.; Purkis, L. H.; Smith, B. R.; Taylor, R. Retrieval of Crystallographically-Derived Molecular Geometry Information. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2133−2144.

(16) Cottrell, S. J.; Olsson, T. S. G.; Taylor, R.; Cole, J. C.; Liebeschuetz, J. W. Validating and Understanding Ring Conformations Using Small Molecule Crystallographic Data. *J. Chem. Inf. Model.* **2012**, *52*, 956−962.

(17) Watkin, D. Structure refinement: Some background theory and practical strategies. *J. Appl. Crystallogr.* **2008**, *41*, 491−522.

(18) Smart, O. S.; Womack, T. O.; Flensburg, C.; Keller, P.; Paciorek, W.; Sharff, A.; Vonrhein, C.; Bricogne, G. Exploiting structure similarity in refinement: Automated NCS and target-structure restraints in BUSTER. *Acta Crystallogr., Sect. D* **2012**, *68*, 368−380.

(19) Debreczeni, J. É.; Emsley, P. Handling ligands with Coot. *Acta Crystallogr., Sect. D* **2012**, *68*, 425−430.

(20) Klei, H. E.; Moriarty, N. W.; Echols, N.; Terwilliger, T. C.; Baldwin, E. T.; Pokross, M.; Posy, S.; Adams, P. D. Ligand placement based on prior structures: The guided ligand-replacement method. *Acta Crystallogr., Sect. D* **2014**, *70*, 134−143.

(21) David, W. I. F.; Shankland, K.; van de Streek, J.; Pidcock, E.; Motherwell, W. D. S.; Cole, J. C. DASH: A program for crystal structure determination from powder diffraction data. *J. Appl. Crystallogr.* **2006**, *39*, 910−915.

(22) Ferreira, F. F.; Trindade, A. C.; Antonio, S. G.; Paiva-Santos, C. d. O. Crystal structure of propylthiouracil determined using high-resolution synchrotron X-ray powder diffraction. *CrystEngComm* **2011**, *13*, 5474−5479.

(23) Vallcorba, O.; Rius, J.; Frontera, C.; Miravitlles, C. TALP: A multisolution direct-space strategy for solving molecular crystals from powder diffraction data based on restrained least-squares. *J. Appl. Crystallogr.* **2012**, *45*, 1270−1277.

(24) Lepailleur, A.; Bureau, R.; Paillet-Loilier, M.; Fabis, F.; Saettel, N.; Lemaître, S.; Dauphin, A.; Lesnard, A.; Lancelot, J.-C.; Rault, S. Molecular Modeling Studies Focused on 5-HT$_7$ versus 5-HT$_{1A}$ Selectivity. Discovery of Novel Phenylpyrrole Derivatives with High Affinity for 5-HT$_7$ Receptors. *J. Chem. Inf. Model.* **2005**, *45*, 1075−1081.

(25) Vasileiadis, M.; Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C. The polymorphs of ROY: Application of a systematic crystal structure prediction technique. *Acta Crystallogr., Sect. B* **2012**, *68*, 677−685.

(26) Karamertzanis, P. G.; Price, S. L. Challenges of Crystal Structure Prediction of Diastereomeric Salt Pairs. *J. Phys. Chem. B* **2005**, *109*, 17134−17150.

(27) Liebeschuetz, J.; Hennemann, J.; Olsson, T.; Groom, C. R. The good, the bad and the twisted: A survey of ligand geometry in protein crystal structures. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 169−183.

(28) Gore, S.; Velankar, S.; Kleywegt, G. J. Implementing an X-ray validation pipeline for the Protein Data Bank. *Acta Crystallogr., Sect. D* **2012**, *68*, 478−483.

(29) Sadowski, J.; Boström, J. MIMUMBA Revisited: Torsion Angle Rules for Conformer Generation Derived from X-ray Structures. *J. Chem. Inf. Model.* **2006**, *46*, 2305−2309.

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

(31) International Union of Pure and Applied Chemistry. IUPAC Gold Book. http://goldbook.iupac.org/ (accessed April 20, 2014).

(32) Bruno, I. J.; Cole, J. C.; Edgington, P. R.; Kessler, M.; Macrae, C. F.; McCabe, P.; Pearson, J.; Taylor, R. New Software for Searching the Cambridge Structural Database and Visualizing Crystal Structures. *Acta Crystallogr., Sect. B* **2002**, *58*, 389−397.

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

(34) Taylor, R. Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 59−67.

(35) Spek, A. L. Structure validation in chemical crystallography. *Acta Crystallogr., Sect. D* **2009**, *65*, 148−155.

(36) Hawkins, P. C. D.; Nicholls, A. Conformer Generation with OMEGA: Learning from the Data Set and the Analysis of Failures. *J. Chem. Inf. Model.* **2012**, *52*, 2919−2936.