Journal of
CHEMICAL INFORMATION
AND MODELING

ARTICLE

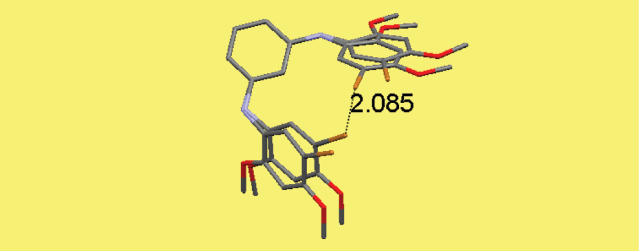
pubs.acs.org/jcim

Short Nonbonded Contact Distances in Organic Molecules and Their Use as Atom-Clash Criteria in Conformer Validation and Searching

Robin Taylor*

Taylor Cheminformatics Software, 54 Sherfield Avenue, Rickmansworth, Hertfordshire WD3 1NL, U.K.

Ⓢ *Supporting Information*



ABSTRACT: Short, intramolecular nonbonded contact distances from a large sample of organic molecules retrieved from the Cambridge Structural Database (CSD) have been analyzed. With the exception of the element pairs N···S, O···P, O···S, and S···S, the first percentiles of X···Y distance distributions (X, Y = C, Br, Cl, F, N, O, P, S) are well estimated by Σvdw −0.5 Å, where Σvdw is the sum of the Bondi van der Waals radii. The 0.1th percentiles are typically a further 0.1 Å shorter. Some 99% of well-refined organic molecules from the CSD have no nonbonded contacts shorter than the 0.1th percentile and no more than two contacts shorter than the first percentile. This can be used as the basis of an atom-clash test for validating less precise crystal-structure data, such as the geometries of protein ligands. In principle, the same test can be used in molecular modeling to identify and filter out unacceptable conformations generated in a conformational search. This is complicated by the fact that conformer generation is usually performed on molecular models with standard bond angles that are not relaxed during the search. In consequence, conformations often appear to contain untenable nonbonded contacts, which would, however, be removed by bond-angle relaxation. This is particularly likely for molecules containing conjugating substituents bonded to adjacent atoms of an aromatic ring, or on the same side of a double bond. Other molecules particularly likely to be affected are those containing rings or other bulky groups separated by a single-atom linkage, and those with the capacity to form intramolecular hydrogen bonds. The problem is greatly ameliorated by the fact that there are many ways to approximate a true conformation, leading to an increased probability that at least one of the approximations will satisfy atom-clash criteria.

INTRODUCTION

Algorithms for generating low-energy conformers of organic molecules are needed for many purposes, such as pharmacophore elucidation and searching,[1] protein−ligand docking,[2] and crystal-structure prediction.[3] New conformer-generation methods have been published regularly over the past few years,[4−12] illustrating that interest in the area continues. The problem is difficult because of its size: a molecule with several rotatable bonds has a huge number of theoretically possible conformations. Further, the computer time available for exploring them is limited because large databases of molecules may need to be processed. In consequence, various time-saving devices are invariably employed.

One such device is to keep bond lengths and bond angles fixed at their input values. Similarly, double bonds and aromatic rings may be kept precisely planar, and three-coordinate nitrogen atoms may be denied the out-of-plane movement needed to alter their degree of pyramidality. "Hydrogen rotors" (groups such as $-CH_3$, $-NH_3^+$, and −OH) are often fixed (not spun during the search). These approximations are collectively referred to herein as the "rigid rotor" (RR) approximation.

The RR approximation can be unsafe because of the extremely steep gradient of the atom−atom nonbonded potential energy curve at short distances. In a fully relaxed search, the strain energy introduced by a torsion rotation that brings two nonbonded atoms into close proximity can often be relieved by distortions to local internal coordinates such as bond angles. The steepness of the nonbonded energy curve means that even minor distortions can often result in enormous energy decreases. Crystal structures in the Cambridge Structural Database (CSD[13]) frequently show such distortions. For example, Figure 1 shows an experimentally observed structure taken from the CSD, together with the same structure with bond angles set to standard (CORINA[14]) values. In the latter, the highlighted Br···Br contact is 1.97 Å, corresponding to an atom−atom repulsion of 783.3 kcal/mol (based on the Tripos force field[15]). In the CSD structure, the two C−N−C bond angles have opened out by several degrees from the standard value of 120°, resulting in a greatly increased Br···Br separation of 4.02 Å.

Various steps can be taken to reduce the effects of the RR approximation, though none is a complete solution. CORINA makes use of some substructure-specific bond angles. For example, the C−O−C angle assigned by CORINA to RCH_2-O-CH_2R' is around 114°, but a wider angle of about 118° is used for the more sterically crowded $RC_6H_4-O-C_6H_5$.

Received: November 28, 2010
Published: March 18, 2011

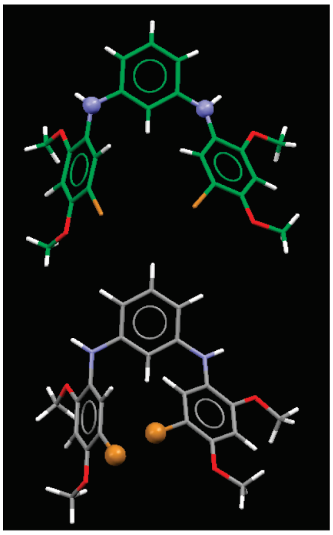


Figure 1. CSD structure TAZMIR (top, carbons in green) and the same structure with bond lengths and angles set to values assigned by CORINA (bottom). The short contact between the bromine atoms shown as spheres in the bottom structure is relieved by widening of the bond angles at the nitrogen atoms shown as spheres in the top structure.

Another common tactic is to soften the nonbonded repulsive potential, either by reducing the atom radii used in the potential functions to (typically) about 60−90% of their normal values,[16] or by altering the functional form of the potential.

An even more important approximation in conformer generation is the in vacuo assumption. Although the modeler's interest is almost always confined to the conformational preferences of molecules in a condensed phase, conformer generators usually treat the molecule in vacuo. While this massively simplifies the calculation, it introduces systematic error. Diller and Merz[17] noted that the conformations of 65 protein ligands taken from the Protein Data Bank (PDB[18]) tended to have larger exposed polar and apolar surface areas than randomly generated low-energy conformations. The latter were produced by minimizing (in vacuo) the energies of random starting points using a force field comprising van der Waals (vdw) and torsional terms. Other authors also concluded that active conformations tend to be less compact than randomly generated in vacuo conformations.[19] The accepted explanation is that extended geometries allow ligands to maximize both their hydrophobic and polar interactions with the protein. Conversely, a molecule in vacuo tends to fold up so as to maximize atom−atom dispersion interactions. Several well-known conformer generators correct (and occasionally perhaps overcorrect[20]) for this effect. For example, ConfGen penalizes conformers containing atom pairs that are close together in space but separated by several chemical bonds.[11] Another tactic is to discourage folded conformations by ignoring attractive vdw and electrostatic terms.[6]

The latter approach recognizes that the conventional (exchange repulsion plus dispersion) vdw energy is not necessarily helpful in conformer generation, as the attractive component may unduly favor folded geometries. On the other hand, the modified repulsion-only function is required to avoid conformations with atom clashes. Given that this is now the only objective of the vdw calculation, an alternative approach is possible. Rather than using force-field methodology, it might be preferable to determine how closely nonbonded atoms are observed to approach in a large sample of high-quality crystal structures. This information can then be used directly to assess whether a generated conformer has unduly short nonbonded contacts. A very large amount of crystal-structure data is available (the CSD now contains over 1/2 million structures), so it may be possible to estimate closest-approach distances with better accuracy than they are predicted by force fields. Further, such "atom-clash" criteria can also be useful for validating low-precision crystal-structure geometries, notably those of ligands in the PDB.[6] On the other hand, they will be as vulnerable to the effects of the RR approximation as vdw calculations, so the extent to which they can be used for conformer generation on molecules with fixed, idealized bond angles is uncertain.

The present study attempts to clarify this issue. First, a set of atom-clash criteria is derived from a large subset of CSD structures. Values that do not correlate well with standard vdw radii are discussed. CSD molecules that violate the criteria (that is, molecules with an unusually large number of unusually short contacts) are examined. A pilot experiment is performed, in which the atom-clash criteria are used to validate a small number of ligands from the PDB. The atom-clash criteria are then applied to a set of molecules built with CORINA and the consequences of the RR approximation quantified. Molecules particularly vulnerable to the approximation (which, therefore, represent difficult cases for conformer generation) are identified and classified.

■ METHODS

The sets of molecules on which the study is based are summarized in Table 1 and were assembled as described below (explanation of the Gridded CORINA Set is deferred to Results and Discussion).

Large CSD Subset for Determination of Atom-Clash Criteria. Atom-clash criteria were determined from a set of 108 015 molecules (the Large CSD Subset) selected from the CSD (version 5.31). Each contained at least 20 atoms, all with three-dimensional (3D) coordinates, of which at least 10 were non-hydrogen. The molecules were taken from crystal structures satisfying the following criteria: R factor ≤ 8%; fully matched (connectivity inferred from atom coordinates matches chemical diagram in CSD); no unresolved errors noted by CSD editors. Structures marked as being disordered were not automatically excluded because, very often, disorder in CSD entries is confined to a small solvate molecule which will be excluded by the check on number of atoms. For those disordered molecules which are not filtered out, CSD editorial procedures ensure that the major component is used. If a structure appears more than once in the CSD, because it has been determined in different polymorphic forms or by different authors, only one example was used (the first when the structures are sorted alphanumerically by CSD reference code). Molecules containing elements other than H, C, N, O, F, P, S, Cl, Br, or with bridging hydrogen atoms, polymeric bonds, or bonds of unknown type, were rejected. Each molecule was required to have at least one rotatable (single, acyclic) bond, excluding hydrogen rotors and bonds for which no valid torsion angle can be found (that is, bonds to linear or near-linear groups such as R−CN).

The resulting subset contains a small proportion of duplicates, since chemically identical molecules may appear in more than one crystal structure, or more than once in the asymmetric unit of the same crystal structure. However, all molecules in the subset are

Table 1. Sets of Molecules Used in this Study

name	description and purpose
Large CSD Subset	organic molecules from the CSD used to determine atom-clash criteria
CSD Neutron Subset	molecules from CSD neutron-diffraction structures used to determine atom-clash criteria involving hydrogen
PDB Ligand Subset	small set of ligands from PDB protein−ligand complexes
Small CSD Subset	organic molecules from the CSD, each one paired with a molecule in each of the sets below
Original CORINA Set	molecules built with CORINA, each one identical in chemical structure to one of the members of the Small CSD Subset, all molecules in their original CORINA-generated conformations
Driven CORINA Set	molecules from the above set, each driven to the conformation observed in the matching molecule of the Small CSD Subset
Gridded CORINA Set	molecules from the above set, each driven to a conformation close to that observed in the matching molecule of the Small CSD Subset, but with torsion angles constrained to lie on a 30° grid

crystallographically independent observations. Hydrogen atom positions were "normalized", which involved setting X−H bond distances to standard values (1.083, 1.009, 0.983, 1.338 Å for X = C, N, O, other) by moving the hydrogen atom along the observed X−H bond direction. This corrects for the tendency of X−H bond lengths measured by X-ray diffraction to be artificially short. The standard values are those recommended in the CSD and were derived from neutron-diffraction measurements.[21]

CSD Neutron Subset for Determination of Atom-Clash Criteria for Hydrogen. Atom-clash criteria involving hydrogen were also determined from a set of 360 molecules taken from structures in the CSD (version 5.31) determined by neutron diffraction (the CSD Neutron Subset). This technique locates hydrogen atoms with much higher precision than is achievable by X-ray diffraction. Selection criteria were as above except that smaller molecules were allowed (at least 10 atoms), molecules were allowed to contain any element, and molecules with no rotatable bonds other than hydrogen rotors were accepted. This relaxation of the selection criteria was enforced by the paucity of neutron data. Hydrogen normalization is, of course, unnecessary for neutron structures and was not performed.

PDB Ligand Subset. Detailed validation of PDB ligand conformations is not the aim of this study. However, a small set of 83 PDB ligands was used to allow a quick assessment of the potential value of the atom-clash criteria for this purpose. The ligands had originally been collated for a different purpose (validation of a pharmacophore elucidation program) and were taken from 10 different protein−ligand series in the Astex Non-Native Set[22] (a collection of PDB structures used for testing docking programs). All structures in this set have resolution of <2.5 Å. Each ligand chosen has at least one rotatable bond.

Molecule Sets for RR Investigation. In order to investigate the RR approximation, pairs of molecules were required, with one member of each pair having crystallographically observed bond lengths, angles, etc. and the other having an identical conformation but with standard (CORINA) bond lengths and angles. This was achieved as follows.

Preliminary CSD Search. A set of 99 564 molecules was selected from the CSD (version 5.30). All the selection criteria used for the Large CSD Subset were applied. In addition, structures whose chemical name contains the substring "fuller" were omitted as CORINA is very slow to process fullerenes. Molecules containing acetylene linkages were eliminated to avoid having to use dummy atoms for torsion rotations.

Generation of CORINA Molecules. The CSD molecules were written out in Tripos mol2 format and read into CORINA (version 3.4) using version 3.2 of the graphical interface CORINA.direct. Three-dimensional structures were regenerated for as many of the molecules as possible using the following CORINA.direct options: keep any atom names given in input file; write formal charges into partial charge column; write added hydrogen atoms; do not write not converted molecules; calculate stereochemistry from input 3D coordinates.

CORINA structures were generated for 97 888 molecules (a 98.3% conversion rate). In 796 of these, CORINA had added one or more hydrogen atoms that were not present in the CSD molecule. This was either because of errors in CSD structure assignments or because the CSD structure was not interpreted correctly by CORINA (for example, hydrogen added to R_3S^+). In many cases, the difficulty probably stems from inconsistencies in the structure conventions used by the two software systems. The discrepant structures were eliminated.

Matching of CORINA and CSD Conformations. Each surviving CORINA molecule was driven to the conformation of the CSD molecule from which it was derived by suitable rotations around the rotatable bonds (excepting hydrogen rotors, each of which was left in its CORINA-generated orientation). The result was the "driven-CORINA molecule": the nearest approximation to the experimentally observed geometry that could be achieved in a conformational search using a CORINA-generated starting point and the RR approximation. In general, it was not possible to exactly reproduce all the torsion angles of the CSD molecule, only some of them. This is because the bond angles and (where relevant) nitrogen pyramidalities of the CSD and CORINA molecules were not identical. Given this, in a rotatable bond such as $R_1(R_2)(R_3)C-C(R_4)(R_5)R_6$, it is possible, when generating the driven-CORINA molecule, to set only one of the torsion angles (for example, $R_1-C-C-R_4$) to its CSD value. The others are then fixed and, because of the bond angle differences, may deviate slightly from the CSD torsion angles. For each rotatable bond, the torsion angle that was exactly reproduced (the "reference torsion") was chosen arbitrarily, except that reference torsions involving hydrogen atoms were avoided.

Analysis of Large rmsd's. The root-mean-square deviation (rmsd) of each driven-CORINA molecule from its "parent" CSD molecule was calculated by least-squares fitting all non-hydrogen atoms. For molecules with topologically equivalent atoms (such as the two oxygen atoms of $R_1-SO_2-R_2$), it was necessary to generate all possible pairings of the driven-CORINA and CSD molecule graphs and take the one that gave the lowest rmsd. Molecule pairs for which there were >1000 possible ways of pairing the atoms were rejected to save time. Of the remaining 89 438 molecule pairs, 8714 had rmsd values of >1 Å. Visual inspection of some examples suggested that most of the large

Table 2. First Percentiles (d_1) and 0.1th Percentiles ($d_{0.1}$) of Nonbonded Contact Distance Distributions in CSD Molecules[a,b]

element pair			no. observations		first percentile, d_1		0.1th percentile, $d_{0.1}$	
		Σvdw	1,*n* (*n* > 4)	1,4	1,*n* (*n* > 4)	1,4	1,*n* (*n* > 4)	1,4
C	C	3.40	85 731	366 582	2.95	2.79	2.83	2.75
C	Br	3.55	733	925	3.13	3.06		
C	Cl	3.45	2 595	1 955	3.00	2.95	2.88	2.84
C	F	3.17	7 930	16 369	2.66	2.63	2.53	2.56
C	N	3.25	14 891	81 562	2.80	2.66	2.66	2.60
C	O	3.22	95 347	230 825	2.66	2.61	2.54	2.53
C	P	3.50	1 289	6 855	2.97	2.95	2.87	2.85
C	S	3.50	4 627	15 603	2.96	2.92	2.85	2.82
Br	O	3.37	211	294	2.89	2.89		
Cl	Cl	3.50	104		3.00			
Cl	N	3.30	322	482	2.92	2.80		
Cl	O	3.27	636	1 210	2.80	2.76		2.73
Cl	S	3.55		115		3.06		
F	F	2.94	3 799	2 147	2.44	2.44	2.22	2.28
F	N	3.02	854	1 839	2.58	2.54		2.47
F	O	2.99	1 698	4 362	2.57	2.50	2.43	2.41
F	P	3.27	266		2.65			
F	S	3.27	293	383	2.71	2.73		
N	N	3.10	1 649	6 924	2.54	2.57	2.44	2.50
N	O	3.07	12 323	17 866	2.51	2.54	2.47	2.49
N	P	3.35		247		2.65		
N	S	3.35	412	1 546	2.50	2.70		2.63
O	O	3.04	12 831	9 891	2.42	2.54	2.37	2.46
O	P	3.32	277	852	2.56	2.71		
O	S	3.32	1 888	3 073	2.32	2.70	2.05	2.63
S	S	3.60	199	269	2.55	2.84		
C	CH	2.79	341 321	358 250	2.36	2.41	2.22	2.21
			493	*1 253*	*2.41*	*2.47*		*2.36*
C	XH	2.79	22 490	38 645	2.20	2.35	2.04	2.22
			100	*366*	*2.08*	*2.30*		
Br	CH	2.94	1 122	174	2.48	2.69	2.33	
Cl	CH	2.84	2 928	519	2.33	2.42	2.24	
Cl	XH	2.84	394		2.23			
F	CH	2.56	5 544	468	2.11	2.24	1.98	
F	XH	2.56	615		1.87			
N	CH	2.64	42 257	13 959	2.21	2.35	2.08	2.16
N	XH	2.64	9 159	4 674	1.60	2.26	1.54	2.15
O	CH	2.61	183 852	26 970	2.13	2.30	1.97	2.06
			325	*133*	*2.19*	*2.30*		
O	XH	2.61	17 985	3 658	1.49	2.19	1.41	2.01
			145		*1.23*			
P	CH	2.89	4 013	3 493	2.30	2.52	2.18	2.44
P	XH	2.89	201	208	2.35	2.45		
S	CH	2.89	10 115	3 285	2.36	2.50	2.23	2.30
S	XH	2.89	801	1 048	1.95	2.51		2.09
CH	CH	2.18	87 862	1 164	1.69	1.29	1.45	1.04
			157		*1.86*			
CH	XH	2.18	10 820	178	1.61	1.59	1.36	
XH	XH	2.18	1 150		0.87		0.49	

[a] All contacts less than sum of van der Waals radii[23,24] and separated by at least one rotatable bond were included in the distributions. [b] Σvdw = sum of van der Waals radii; distances in angstroms. Values in italics refer to the element pair on the row above and are based solely on structures determined by neutron diffraction.

rmsd's were due to the presence of rings (often macrocyclic). While ring conformations generated by CORINA should be energetically accessible, and those in the CSD obviously are, they will not necessarily be the same. CORINA has an option to generate multiple ring conformations, but the effort involved in finding the best match for each CSD molecule then becomes much more significant. Therefore, all molecules containing any ring other than a three- or four-membered ring or a planar (largest ring torsion of <5°) five- or six-membered ring were eliminated.

Of the 40 443 pairs of CSD and driven-CORINA molecules that now remained, only 457 had rmsd values of >1 Å while the large majority (37 880) had rmsd values of <0.5 Å. A brief visual inspection was made of some of the molecule pairs with large rmsd's. One of the largest was 4.47 Å for CSD entry FEJGUW, and was due to CORINA generating a C═C double bond with the stereochemistry opposite that seen in the CSD. Other large rmsd's were due to a variety of causes, including the following: (a) some remaining ring-conformation discrepancies; (b) occasional CORINA errors (for example, a bent allene group for CSD entry AMEROZ); (c) several fairly exotic structures on which CORINA struggles (for example, DENHOT, which contains a $C{=}N^{+}{=}C$ group); (d) some structures where the CSD geometry is suspect (for example, one of the carbonyl carbons in MEZMOT is very nonplanar); and (e) occasional cases (such as WOVZUC) where the structure-matching algorithm failed to find the pairing corresponding to the best rmsd because the symmetry of the chemical representation in the CSD is lower than that of the 3D structure.

All structures with rmsd values of >1 Å were rejected, leaving a final sample of 39 986 driven-CORINA molecules (the Driven CORINA Set; the corresponding CSD molecules constitute the Small CSD Subset, and the 39 986 CORINA molecules in their original conformations constitute the Original CORINA Set). One hundred pairs of driven-CORINA and CSD molecules with rmsd's in the range 0.5−1.0 Å were chosen at random, least squares overlaid, and the overlays were inspected visually to determine whether geometric differences could be entirely ascribed to the RR approximation. Nine of these pairs had rmsd's between 0.9 and 1.0 Å, of which three exhibited none of the problems discussed in the previous paragraph, leaving only the RR approximation to cause geometric differences. Corresponding figures for the other 0.1 Å ranges in the sample were the following: 0.8−0.9 Å, 7 of 10 with differences due solely to RR approximation; 0.7−0.8 Å, 12 of 17; 0.6−0.7 Å, 16 of 20; 0.5−0.6 Å, 41 of 44. In another random sample of 100 pairs with rmsd's in the range 0−0.5 Å, all differences could be ascribed to the RR approximation. Assuming these ratios apply across all structures in the Driven CORINA Set, it was then possible to estimate that, for about 99% of driven-CORINA molecules, geometric differences from their CSD analogues are consequences only of the RR approximation. A reviewer is thanked for suggesting this analysis.

van der Waals Radii. van der Waals radii were taken from Bondi[23] (C, 1.70; Br, 1.85; Cl, 1.75; F, 1.47; N, 1.55; O, 1.52; P, 1.80; and S, 1.80 Å) except for that of hydrogen (1.09 Å), which was taken from Rowland and Taylor.[24]

RESULTS AND DISCUSSION

Low Percentiles of Intramolecular Nonbonded Contact Distance Distributions. For each element pair X,Y (where X and Y can be C, Br, Cl, F, N, O, P, S, H), the objective is to determine the shortest intramolecular nonbonded X···Y contact that can be expected to occur in an organic molecule. The minimum of the X···Y distance distribution calculated from the Large CSD Subset is unlikely to be a robust measure since there is a good chance it will be an artifact of experimental error. A better measure is a low percentile of the distribution.[6] In this work, the first percentile and the 0.1th percentile were used, which will be referred to as d_1 and $d_{0.1}$, respectively. In calculating the X···Y nonbonded distance distribution, 1,3 contacts were excluded since they merely reflect bond-angle preferences. Also excluded were distances greater than the sum of the vdw radii (Σvdw) of X and Y, since the object of the exercise is to focus on atom pairs that have been pushed closer together than this "ideal" value (in other words, are on that part of the nonbonded energy−distance curve that has negative gradient). Contacts between atoms in rigid fragments were rejected, since their distances are fixed by bond length, bond angle, and, possibly, ring-closure constraints, and will remain invariant in a conformational search. Separate distributions were calculated for 1,4 contacts and 1,n (n > 4).

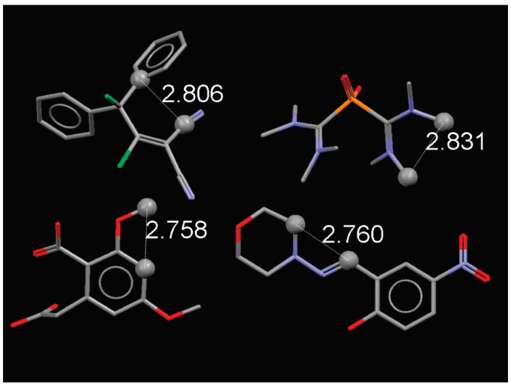


Figure 2. Some C···C contacts at about the 0.1th percentile distance (CSD entries CLCYPE, FAGBUK, DMXHPA01, FINNIA).

Table 2 lists the d_1 values of the nonbonded distance distributions calculated from the Large CSD Subset for all element pairs, excluding those for which fewer than 99 observations were available; $d_{0.1}$ values are given for those distributions with at least 999 observations. Separate values are given for carbon-bound hydrogen, CH, and other hydrogens, XH. Where sufficient data were available, percentiles for distributions involving hydrogen were also derived from the CSD Neutron Subset, and these are given in italics in Table 2.

Visual inspection of CSD structures with contacts at d_1 or $d_{0.1}$ levels usually suggests a plausible physicochemical explanation for the extreme value. For example, Figure 2 shows some of the C···C contacts at or close to the $d_{0.1}$ distances. Although these contact distances are extremely unusual—by definition, there is only one chance in a thousand of finding a contact as short as the $d_{0.1}$ value—they appear credible. Often, they occur because of conjugation, which favors planar conformations that tend to bring nonbonded atoms into close proximity. Another common cause of abnormally short contacts is steric overcrowding; tertiary carbons featured in several of the examples that were examined. Of course, the most extreme outliers in the distribution are surely the result of experimental error (the minimum C···C distance is <2 Å).

Figure 3 shows the d_1 values plotted against Σvdw for the 1,n (n > 4) and 1,4 contacts of all element pairs except those

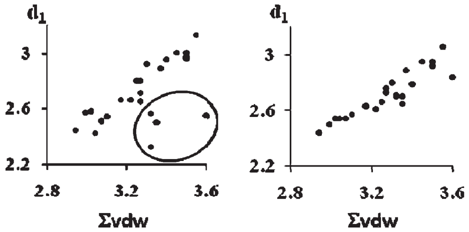


Figure 3. First percentile distances (d_1, see Table 2) plotted against sums of vdw radii (Σvdw) for 1,n ($n > 4$) contacts (left) and 1,4 contacts (right). All element pairs for which sufficient data are available are included except those involving hydrogen.

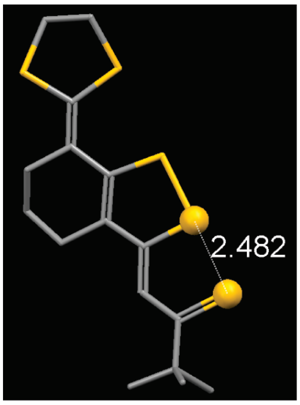


Figure 4. Exceptionally short S···S contact in CSD entry DITTBT10.

involving hydrogen. In the 1,n ($n > 4$) distributions, d_1 correlates reasonably well with Σvdw except for four gross outliers (circled in Figure 3), where the observed d_1 is much lower than would be expected from vdw radii. These outliers are the following element pairs: N,S; O,P; O,S; and S,S. The latter is the most discrepant point. (It is interesting that the d_1 values for N,N and O,O are not gross outliers, despite the fact that many short contacts between these element pairs will surely be intramolecular hydrogen bonds.) S,S is also an outlier on the 1,4 plot, though to a lesser extent. The tendency for sulfur to form exceptionally short contacts to nitrogen and oxygen is well established[25,26] and is a consequence of electrostatic attraction between the electronegative O or N and polarized (electropositive) S. Figure 4 shows an example of one of the exceptionally short S···S contacts. In this example, it would be possible to draw an alternative representation of the structure in which the two sulfur atoms involved in the short contact are formally bonded. Similar situations occur in other structures containing exceptionally short S···S contacts (for example, CSD entries LIKCOX, MPTVTP, MXTYCT). The few short O···P contacts that result in the low d_1 value for this element pair mostly occur in somewhat exotic molecules (for example, with negatively charged phosphorus).

Turning now to element pairs involving hydrogen, it is evident that, despite hydrogen normalization, some of the values in Table 2 are the products of experimental error. For example, the X-ray d_1 and $d_{0.1}$ values of the XH,XH 1,n ($n > 4$) distribution are 0.87 and 0.49 Å, both hopelessly unrealistic. Experimental errors are likely to be particularly important here as both members of the pair are hydrogen. More surprising is the exceptionally short neutron value (1.23 Å) for the first percentile of the O,XH distribution. This is caused by sample bias: several structures have been solved by neutron diffraction because they were suspected to contain very short, possibly symmetrical, O−H···O hydrogen bonds (for example, CSD entry AJOHEM). The small sample size exacerbates the consequences of this bias.

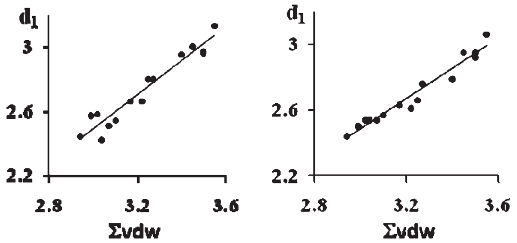


Figure 5. First percentile distances (d_1, see Table 2) regressed against sums of vdw radii (Σvdw) for 1,n ($n > 4$) contacts (left) and 1,4 contacts (right). Plots based on the following element pairs: C,C; C,Br; C,Cl; C,F; C,N; C,O; C,P; C,S; Cl,O; F,F; F,N; F,O; N,N; N,O; O,O.

Despite these problems, many of the heavy-atom···hydrogen distributions have credible d_1 and $d_{0.1}$ values. For example, the d_1 values for the 1,n ($n > 4$) distributions in which CH is paired with Br, C, F, Cl, and O are all about 0.4−0.5 Å shorter than the sum of vdw radii, which is similar to many of the distributions not involving hydrogen atoms. N,XH and O,XH have particularly short d_1 values relative to vdw radii, and although this may partly be due to experimental errors, it may also reflect intramolecular hydrogen bonding.

An obvious problem with using the percentile values in Table 2 as atom-clash criteria is that there are several missing values, corresponding to element pairs for which insufficient data were available. A predictive model is therefore necessary for these values. For this purpose, a subset of element pairs was taken, including all pairs not involving hydrogen, and excluding also (a) the four outliers discussed above (N,S; O,P; O,S; S,S) and (b) points based on <500 observations. For this subset, the d_1 versus Σvdw correlation is very good (Figure 5). Linear regression of d_1 on Σvdw gives r^2 values of 0.931 and 0.954 for the 1,n ($n > 4$) and 1,4 plots, respectively, with the gradients of the regression lines being 1.064 and 0.922, respectively. Predicted values from the regression line have rmsd's from the observed d_1 values of 0.058 and 0.040 Å, respectively. However, the simple relationship

$$d_1 = \sum \text{vdw} - 0.5 \qquad (1)$$

is almost as good (giving rmsd's of 0.059 and 0.055 Å) and has the intuitively appealing gradient of 1. Using analogous methodology, the relationship

$$d_{0.1} = \sum \text{vdw} - 0.6 \qquad (2)$$

gives predictions of 0.1th percentile values with rmsd's of 0.058 Å for the 1,n ($n > 4$) distributions and 0.049 Å for the 1,4 distributions.

Robustness of Percentile Estimates. The good correlations between Σvdw and the percentile values suggest that the latter have been estimated with good accuracy, at least for element pairs not involving hydrogen. However, in order to test the robustness of the percentile estimates, they were recalculated from a new CSD subset selected with the same criteria as before, except for a more stringent limit on the R factor (≤5%) and rejection of all disordered structures. Detailed results are available as Supporting Information. Most of the new estimates are within 0.05 Å of those discussed in the preceding section. Discrepancies in excess of

Table 3. Numbers of Atom-Clash Violations in Various Sets of Molecules

number of atom-clash violations		percentage of molecules in set with stated number of violations					
primary	secondary (ns)	CSD	PDB ligands	original CORINA	driven CORINA	driven CORINA[a]	gridded CORINA[a]
0	0	93.0	78.3	54.0	50.7	53.7	57.8
0	1	5.0	6.0	9.3	10.2	9.9	20.8
0	2	1.0	2.4	3.8	3.7	2.9	10.0
1	1	0.5	3.6	10.7	10.6	11.4	4.3
1	2	0.2	3.6	3.7	5.2	5.1	1.5
2	2	0.1	3.6	5.8	5.3	5.3	1.1
≤ns	3−5	0.3	2.4	9.8	10.9	9.5	3.6
≤ns	>5	<0.1	0.0	3.0	3.3	2.2	0.8

[a] Restricted to molecules with molecular weight ≤500 and ≤10 rotatable bonds.

0.1 Å are seen for only six distributions, four of which (discrepancies of −0.17, 0.31, −0.15, and −0.11 Å) involve hydrogen. The other two are the 1,*n* (*n* > 4) $d_{0.1}$ value for F,F and the 1,*n* (*n* > 4) d_1 value for O,P. In the former case, the value based on the new subset of structures is larger than the value in Table 2 (2.42 versus 2.22 Å) and is probably more accurate, since CF_3 groups are notoriously prone to disorder. In the latter case, the new value is much shorter than that in Table 2 (2.16 versus 2.56 Å) and does not seem credible; probably, it is less accurate because it is based on a very small sample size (there are only 142 observations from the *R* factor <5% set). This shows that restriction to more experimentally precise crystal structures is a double-edged sword: what is gained in the precision of individual distances can be lost by the concomitant reduction in sample sizes. One possible way around the dilemma is to use less extreme percentiles: for example, the fifth percentile. This will be less prone to extreme outliers and therefore more reliably estimated from small samples. However, a 1 in 20 chance is not particularly unlikely, and it may be too draconian to base atom-clash criteria on distances that have been observed many times in CSD structures.

Atom-Clash Criteria. The above results were used to set atom-clash criteria as follows:

1. For element pairs not involving hydrogen, 0.1th percentile ($d_{0.1}$) values were used as *primary clash criteria* and were taken from Table 2 if possible. If no value was available in Table 2, $d_{0.1}$ was set to d_1 − 0.1 unless the d_1 value was also missing from Table 2. In this situation, $d_{0.1}$ was set to Σvdw − 0.6.
2. For element pairs not involving hydrogen, first percentile (d_1) values were used as *secondary clash criteria* and taken from Table 2 if possible. If not available in Table 2, they were set to Σvdw − 0.5.
3. No atom-clash criteria were set for hydrogen, for the following reasons. First, they are difficult to determine reliably (see above). Second, they are likely to be of very limited use in validating X-ray geometries since, in many structures of relatively low precision (including almost all macromolecular structures), hydrogen atom positions are undetermined. Third, their use in conformational searching is complicated if, as is usual, hydrogen rotors are not allowed to spin.

Atom-Clash Violations in CSD Molecules. Since a crystallographically observed molecule may contain several nonbonded contacts shorter than Σvdw, the probability that the molecule will contain at least one contact shorter than d_1 (or $d_{0.1}$) is greater than 1 in 100 (or 1000). To quantify this, each molecule in the Large CSD Subset was examined to determine the number of contacts it contained that violated the primary or secondary clash criteria defined above. Results are summarized in the column of Table 3 headed "CSD". For example, the second figure in this column indicates that 5.0% of molecules in the Large CSD Subset contained one contact violating a secondary atom-clash criterion and no contacts violating primary criteria. Only contacts separated by at least one rotatable bond were considered in the calculation.

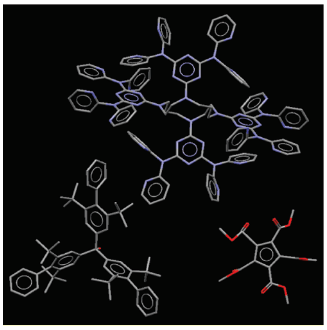


Figure 6. Highly overcrowded molecules from CSD structures EJUXIQ, MAYTOV, and VAFVIH.

The results show that 99.0% of the CSD molecules have no primary violations (nprimary = 0) and no more than two secondary violations (nsecondary ≤ 2). One or two secondary violations cannot be taken as an indication of experimental error. Any molecular conformation is due to a balance of attractive and repulsive interactions, so some of the latter should be expected. In particular, the planar geometries favored by conjugation often go hand in hand with short nonbonded contacts. The combined criterion (nprimary = 0, nsecondary ≤ 2) represents a 99% confidence limit for organic molecules in well-refined small-molecule crystal structures. Any molecule exceeding this limit represents an event with probability of <0.01, suggesting exceptional internal steric strain or experimental error. Visual inspection of molecules in the Large CSD Subset that have the largest number of violations suggested that most were genuine examples of severe steric overcrowding. Figure 6 shows three examples.

The Large CSD Subset was chosen from the more precisely determined structures in the CSD (*R* factor ≤ 8%, all hydrogen atom positions present). Tests on a subset selected from relatively imprecise structures (*R* factor ≥ 10%, hydrogen atoms

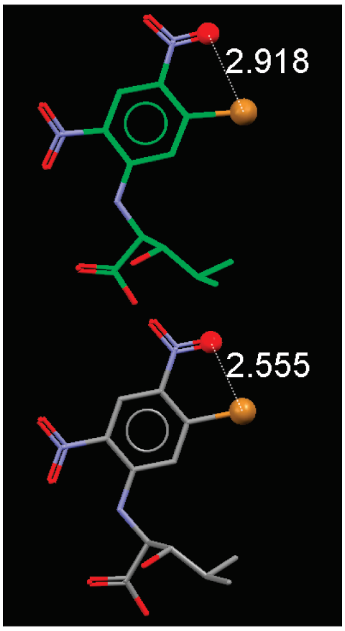


Figure 7. CSD geometry (top) and driven-CORINA geometry (bottom) for CSD entry GANVUM.

allowed to be missing) found that 6.0% of molecules exceeded the 99% confidence limit defined above. Since all other selection criteria were as for the Large CSD Subset, the higher violation rate can be ascribed to an increased number of experimental errors.

Atom-Clash Violations in PDB Ligands. The column in Table 3 headed "PDB ligands" gives the number of atom-clash violations in the PDB Ligand Subset. Some 13.2% of this small test set exceed the 99% confidence limit derived above (nprimary = 0, nsecondary ≤ 2). Either or both of two explanations may be advanced: (a) some of the PDB ligands are significantly in error; (b) binding to a protein can introduce significantly greater steric strain than packing in a small-molecule crystal structure. The strain energy of protein-bound ligands has been a subject of some controversy, with wildly different suggestions regarding how much strain can be tolerated.[19,27,28] However, a recent study[29] using higher-level approximations than earlier work suggests that ligands rarely if ever surpass 2 kcal/mol strain energy. On this basis, the first of the two explanations is more likely. It is worth remembering that the test set was taken from relatively well-refined PDB structures, so more atom-clash violations may be expected in a random sample of ligands.

Atom-Clash Violations in CORINA Molecules. Table 3 also gives the numbers of atom-clash violations in the Original CORINA Set and the Driven CORINA Set. It can be seen immediately that both contain many more atom clashes than are seen in the sets discussed above. It may plausibly be assumed that this is a direct consequence of the RR approximation. Driving the CORINA molecules to the conformations observed in the CSD makes the problem worse: 35.3% of the Driven CORINA Set exceed the 99% confidence limit (nprimary ≤ 0, nsecondary = 2), compared to 33.0% of the Original CORINA Set. The situation is improved a little by eliminating large molecules (molecular weight > 500, number of rotatable bonds > 10), a legitimate refinement if attention is focused on drug design. Results for the Driven CORINA Set after modification in this way are given in the penultimate column of Table 3, with the percentage of molecules exceeding the 99% confidence limit now being 33.5%.

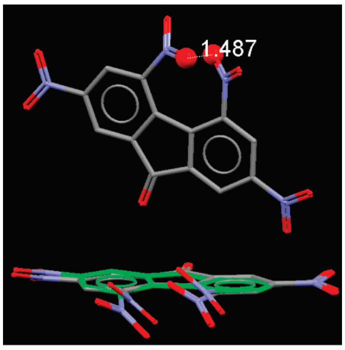


Figure 8. Driven-CORINA geometry of CSD entry FIJHAH (top) with oxygen atoms involved in close contact shown as spheres. At bottom, the driven-CORINA molecules is superimposed on the experimentally observed CSD molecule, the ring system of which (shown in green) twists slightly to relieve the contact.

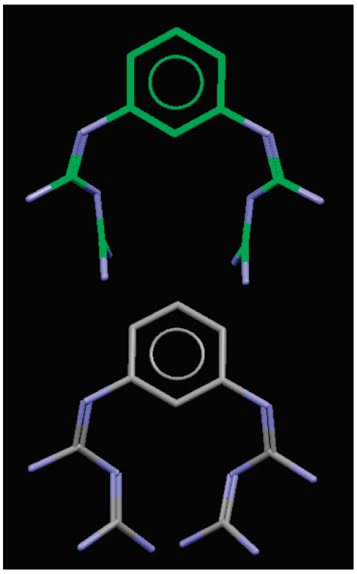


Figure 9. Driven-CORINA geometry (top) and CSD geometry (bottom) for CSD entry TESVAP.

Visual Inspection of Driven-CORINA Structures with Short Contacts. A few dozen driven-CORINA structures with particularly short contacts were visually inspected. Some examples illustrate typical problems that occur. CSD entry GANVUM (Figure 7) is a classic case. A Br···O contact that is short but not unrealistic (2.92 Å) in the CSD molecule becomes untenably short (2.56 Å) in the driven-CORINA molecule due to the use of standard bond angles (120°) in the latter (the relevant $(O_2)N-C-C$ and $C-C-Br$ angles in the CSD molecule are 124.5 and 125.4°, respectively). This simple situation, where unreasonably short contacts appear entirely due to the use of standard bond angles, was seen in the large majority of cases inspected. Frequently, as in the case just discussed, the close contacts occurred in conjugated systems. When groups exocyclic to an aromatic ring become coplanar with the ring to maximize electron delocalization, close contacts tend to be inevitable.

Figure 8 illustrates a different type of problem. An extremely short (1.49 Å) contact between the two oxygen atoms shown as spheres in the driven-CORINA molecule is relieved in the CSD molecule by a small twist in the geometry of the ring system,

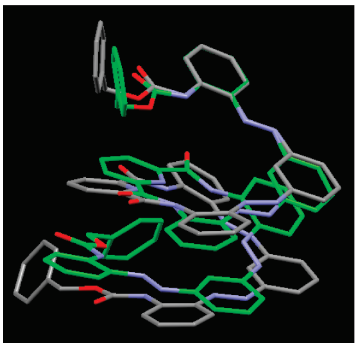


Figure 10. Driven-CORINA structure (carbon atoms in gray) for CSD entry VEBHOA overlaid on experimentally observed geometry (carbons in green) by least-squares fitting the ring at top right. Comparatively minor differences in internal coordinates accumulate along the chain, resulting in large deviations at the end of the chain.

which is perfectly planar in the CORINA model. Figure 9 shows a molecule which contains four C═N double bonds. CORINA has made all four almost planar and used 120° angles for the sp^2 atoms, which results in a very short N···N contact. In the CSD molecule, two of the C═N bonds are significantly nonplanar and key bond angles open out, relieving the contact.

A case of particular interest is CSD entry VEBHOA (Figure 10). Here, comparatively minor differences between the observed bond lengths and angles in the crystal structure and those assigned by CORINA occur at many points along a chain. In addition, there are small differences in the torsion angles around double bonds, which are not perfectly planar in either structure. The effect of these many minor discrepancies is cumulative, leading to a large rmsd (the molecule was actually excluded from the Driven CORINA Set as it failed the rmsd test).

Close Is Good Enough. In conformational searching, the effects of the RR approximation are ameliorated by a stroke of good fortune. It is rarely if ever necessary to reproduce the true conformation of the molecule under investigation with high accuracy; a reasonable approximation is usually good enough. Indeed, torsion angles in many conformational searches are only allowed to take certain discrete values (typically, those on a 30° grid), which virtually guarantees that, at best, there will be a few degrees of error in each torsion angle. If there are n rotatable bonds in a molecule and a 30° grid search is performed, there are 2^n ways of approximating the true conformation such that, in each approximation, all torsion angles are within 30° of the true value. Because there can be many ways in which the true conformer can be approximated, there is a correspondingly increased probability that at least one of them will have sufficiently few atom clashes to be found.

The matter was investigated as follows. The subset of the Driven CORINA Set comprising molecules with molecular weight of ≤500 and with ≤10 rotatable bonds was taken. For each molecule, all conformations were generated that satisfied the following conditions: (a) all torsion angles fell on a 30° grid (allowed values −180, −150, −90, ..., 150°); (b) all torsion angles fell within 30° of the values observed in the CSD; (c) the rmsd from the CSD conformation was <1 Å. Of these, the conformation with the lowest number of atom-clash violations (counting first on nprimary, then on nsecondary) was accepted. The result was a conformer for each of 34 951 molecules, which collectively comprised what will be termed the Gridded CORINA Set. The numbers of atom-clash violations in the Gridded CORINA Set are summarized in the last column of Table 3. Some 11.3% of the set exceed the atom-clash 99% confidence limit, substantially less than the corresponding figure for the Driven CORINA Set.

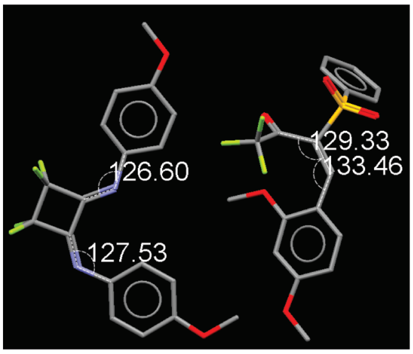


Figure 11. Problem case: molecules with substituents on the same side of a double bond (CSD entries QEHRUR, ASUWOA).

Finding the True Conformer as the Least Bad Answer. A further possibility is that a gridded-CORINA conformer with several atom clashes might still be found in a search if most of the other conformations that can be generated for the molecule are even worse. This might apply to sterically crowded molecules or to molecules in which the RR approximation results in unrealistically short nonbonded contacts irrespective of the conformation. The possibility was investigated on a sample of 1444 molecules taken from the 11.3% of the Gridded CORINA Set that exceed the atom-clash 99% confidence limit. For each molecule, a sample of conformations on a 30° grid was generated and the proportion that had fewer atom-clash violations than the gridded-CORINA conformer was computed. For molecules with four or fewer rotatable bonds, the entire conformational grid was examined. For more flexible molecules, random subsets of the grid were used, of sizes 50 000, 100 000, and 1 million for molecules with 5, 6−8, and 9−10 rotatable bonds, respectively.

Of the 1444 molecules, there were 230 (15.9%) for which the gridded-CORINA conformer had less severe atom-clash violations than at least 99.9% of the randomly generated conformations, suggesting that they might well be found in a search. (This is by no means guaranteed for a molecule with many rotatable bonds, but on the other hand, gridded-CORINA conformers scoring less well—say in the best 90%—might be found for molecules with relatively few rotatable bonds.)

Analysis of Problem Molecules. Molecules from the above experiment that performed especially poorly (that is, in which the gridded-CORINA conformer had more atom-clash violations than at least half of the random conformations) were subjected to a series of substructure searches using the program ConQuest.[30] These molecules represent cases where the RR approximation particularly disfavors the true conformation. The aim of the substructure searches was to identify common patterns among these worst cases. The most common situations were as follows.

Molecules with Substituents on the Same Side of C═C and C═N Double Bonds. Examples are CSD entries ASUWOA and QEHRUR (Figure 11). Bond angles in the X-ray structures open out to an extraordinary extent, sometimes exceeding 130°. The true conformation is often planar (to allow conjugation) and therefore tends to be particularly disfavored compared to others when bond angles are set to standard values. Similar problems can occur in aromatic or other delocalized rings with ortho substituents.

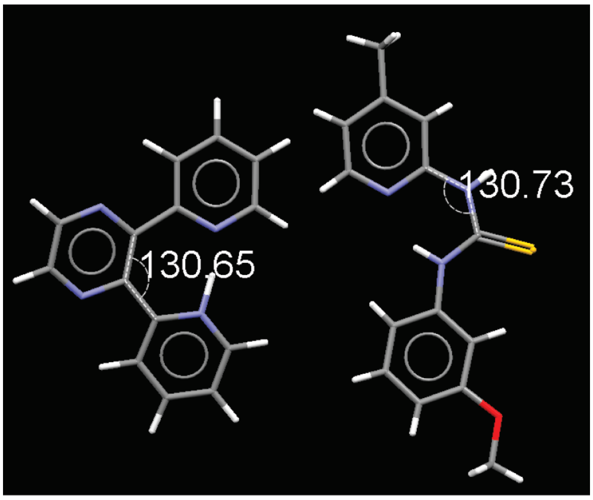


Figure 12. Problem case: molecules with intramolecular hydrogen bonds (CSD entries GOLCAL, BALFOK).

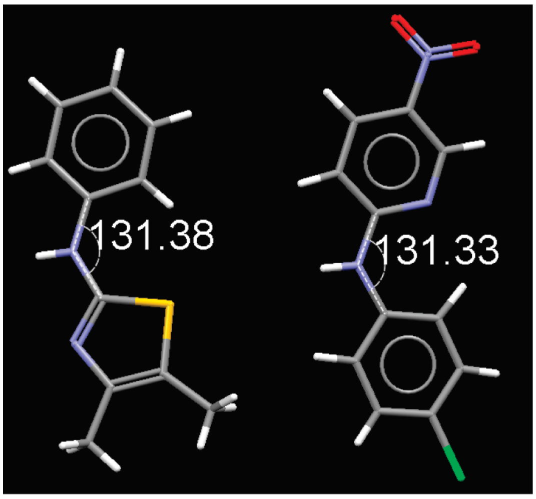


Figure 13. Problem case: molecules with single-atom linkages between two bulky groups, especially Ar—NH—Ar (CSD entries ADIHUR, FAKWAP).

Molecules with Intramolecular Hydrogen Bonds. Examples are CSD entries BALFOK and GOLCAL (Figure 12). In the X-ray structures, bond angles open out to allow the hydrogen bond to form. Since the structures are observed in a condensed phase, there is a very good chance that these intramolecular hydrogen bonds will occur in other condensed-phase situations, such as a protein binding site.

Molecules with Single-Atom Linkages between Two Bulky Groups, Especially Ar—NH—Ar. Examples are CSD entries ADIHUR and FAKWAP (Figure 13), where the angle at nitrogen opens out to >130°. In the right-hand example (FAKWAP), this facilitates a CH···N interaction that is probably attractive.

Amides and Ureas. The average C—N—C and N—C=O angles in the acyclic substructure C(aromatic)—NH—C(=O)— are 127.1 and 124.1°, based on a search of the CSD, version 5.31. Setting these angles to 120° is enough to cause atom-clash violations if the amide group is in a planar, conjugating position.

CONCLUSIONS

The atom-clash criteria determined in this work should be of use for identifying unlikely molecular conformations in X-ray structures, including those of protein-bound ligands. The applicability of CSD conformational data to protein ligand conformations was reviewed in detail recently by Brameld et al.,[31] who concluded that, in general, "small molecule crystal conformations are comparable to the protein-bound conformations and therefore are directly relevant to structure-based drug design".

Application of the clash criteria to conformer generation is complicated by the RR approximation. This is not, of course, a problem restricted to clash criteria; it exists whatever method is used to filter out unwanted conformations, whether it be knowledge-based heuristics or force-field calculations. Nor are the problems restricted to molecular models generated by CORINA. The complications arise from disallowing geometry relaxation during searching, and any method of generating molecular models is likely to suffer at least as much as CORINA (probably more, given that CORINA is recognized as one of the best programs of its type). Happily, the effects of the RR approximation are greatly reduced by the large number of ways in which an approximation to the true conformation can usually be made, and the fact that any one of these (provided it can be found!) is generally good enough for the purpose. Indeed, if this were not the case, conformer generation with fixed bond angles might not be a viable proposition.

The results obtained in this work suggest that, when CORINA geometries are used, the true conformation of about 11% of molecules might be missed if the atom-clash 99% confidence limit is applied as a rigorous filter (other 3D molecule builders will, of course, give different results, but CORINA is of particular interest because it is so widely used). In practice, however, it is unlikely to be applied as such and the outcome will be better than this. First, it is safe to assume that, in any conformational search, the best conformers will be kept even if they contain several atom clashes. On this basis, a proportion of the missed conformations will be found as the "least bad option". Second, this study has only dealt with the nonbonded-repulsion component of internal energy. In any practical conformer-generation algorithm it is also necessary to include a torsional term (which, like the atom-clash criteria, could also be knowledge-based[32]) to take into account conjugation effects. Many of the true conformers in the "missing 11%" are highly conjugated and will be strongly favored by the torsional term.

A number of possibilities exist for refining the atom-clash criteria reported in this work. They were based on a sample that included some very exotic and some very overcrowded molecules. Therefore, restriction of the data set to druglike molecules might be worthwhile for those whose interest lies in that area. The use of the very short clash criteria for S,S and O,P is questionable when the interest is solely focused on druglike molecules. Arguably, it might have been better to avoid using as clash criteria those percentile values in Table 2 that are based on relatively small numbers of observations (for example, the Cl,Cl element pair); the vdw-based expressions (1) and (2) could have been used instead. Since short S···N and S···O contacts may be expected only for certain types of sulfur atoms—those in electron-withdrawing environments—there is a strong argument for deriving separate clash values for polarized sulfur and nonpolarized sulfur. Finally, four categories of molecules that are particularly vulnerable to the RR approximation have been identified, and it should be possible to apply more relaxed clash criteria in these specific situations. Alternatively, bond angles for known, problem groups could be opened out. It might be possible to treat sterically crowded conjugated systems as special 3D substructures in conformer generation, rather than treating them as free rotors.

A last resort is to make all the clash criteria more relaxed when they are applied in a fixed bond angle situation (if all criteria are reduced by 0.1 Å, the 11% "failure rate" referred to above drops to 6.4%). This is the equivalent of making vdw potentials softer when conformer filtering is based on force-field calculations. The danger is that it will let through conformations that are genuinely strained as well as those that have unduly short nonbonded contacts solely as a result of the RR approximation.

Finally, how important is the RR approximation in practice? A reviewer argues that it is becoming less widely used: many programs now employ a "quick and dirty" relaxation step as an integral part of the algorithm. The point is acknowledged, but there are still many conformational-search algorithms that keep bond angles fixed, possibly because relaxation adds an extra computational burden to what is already a demanding problem. These include not only stand-alone conformer generators but also algorithms that perform on-the-fly conformational searching as a means to an end, such as protein–ligand docking. (Docking, in fact, may be particularly vulnerable to the RR approximation. The more exactly we need to reproduce a true conformation—for example, if docking into a tight, rigid binding site—the more likely it is that the RR approximation will be detrimental.) Of course, if the RR approximation becomes increasingly less used, application of the atom-clash criteria reported herein becomes increasingly straightforward.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information. Lists of the CSD and PDB reference codes of the structures from which the data sets in Table 1 were taken; first and 0.1th percentiles based on a CSD subset selected with the more stringent criteria (R factor ≤ 5%, no disordered structures). This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: robin@justmagnolia.co.uk.

■ ACKNOWLEDGMENT

Molecular Networks GmbH and the Cambridge Crystallographic Data Centre are thanked for providing access to CORINA and the CSD, respectively. The author is grateful to John Liebeschuetz for helpful comments.

■ REFERENCES

(1) Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R. Three-Dimensional Pharmacophore Methods in Drug Discovery. *J. Med. Chem.* **2010**, *53*, 539–558.

(2) Young, D. C. Docking. In *Computational Drug Design: A Guide for Computational and Medicinal Chemists*; Wiley: Hoboken, NJ, USA, 2009; pp 133−160.

(3) Day, G. M.; Cooper, T. G.; Cruz-Cabeza, A. J.; Hejczyk, K. E.; Ammon, H. L.; Boerrigter, S. X. M.; Tan, J. S.; Della Valle, R. G.; Venuti, E.; Jose, J.; Gadre, S. R.; Desiraju, G. R.; Thakur, T. S.; van Eijck, B. P.; Facelli, J. C.; Bazterra, V. E.; Ferraro, M. B.; Hofmann, D. W. M.; Neumann, M. A.; Leusen, F. J. J.; Kendrick, J.; Price, S. L.; Misquitta, A. J.; Karamertzanis, P. G.; Welch, G. W. A.; Scheraga, H. A.; Arnautova, Y. A.; Schmidt, M. U.; van de Streek, J.; Wolf, A. K.; Schweizer, B. Significant progress in predicting crystal structures of small organic molecules—a report on the fourth blind test. *Acta Crystallogr., Sect. B: Struct. Sci.* **2009**, *65*, 107–125.

(4) 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.

(5) Griewel, A.; Kayser, O.; Schlosser, J.; Rarey, M. Conformational Sampling for Large-Scale Virtual Screening: Accuracy versus Ensemble Size. *J. Chem. Inf. Model.* **2009**, *49*, 2203–2311.

(6) 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.

(7) 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.

(8) Pavlov, T.; Todorov, M.; Stoyanova, G.; Schmieder, P.; Aladjov, H.; Serafimova, R.; Mekenyan, O. Conformational Coverage By a Genetic Algorithm: Saturation of Conformational Space. *J. Chem. Inf. Model.* **2007**, *47*, 851–863.

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

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

(11) 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.

(12) Zhu, F.; Agrafiotis, D. K. Self-Organising Superimposition Algorithm for Conformational Sampling. *J. Comput. Chem.* **2007**, *28*, 1234–1239.

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

(14) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of automatic three-dimensional model builders using 639 X-ray structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000–1008.

(15) 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.

(16) Boström, J.; Greenwood, J. R.; Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J. Mol. Graphics Modell.* **2003**, *21*, 449–462.

(17) Diller, D. J.; Merz, J. M., Jr. Can we separate active from inactive conformations? *J. Comput.-Aided Mol. Des.* **2002**, *16*, 105–112.

(18) 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.

(19) Perola, E.; Charifson, P. S. Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganisation upon binding. *J. Med. Chem.* **2004**, *47*, 2499–2510.

(20) 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.

(21) Allen, F. H.; Bruno, I. J. Bond lengths in organic and metal-organic compounds revisited: X-H bond lengths from neutron diffraction data. *Acta Crystallogr., Sect. B: Struct. Sci.* **2010**, *66*, 380–386.

(22) Verdonk, M. L.; Mortenson, P. N.; Hall, R. J.; Hartshorn, M. J.; Murray, C. W. Protein-Ligand Docking against Non-Native Protein Conformers. *J. Chem. Inf. Model.* **2008**, *48*, 2214–2225.

(23) Bondi, A. Van der Waals Volumes and Radii. *J. Phys. Chem.* **1964**, *68*, 441–451.

(24) Rowland, R. S.; Taylor, R. Intermolecular Nonbonded Contact Distances in Organic Crystal Structures: Comparison with Distances Expected from van der Waals Radii. *J. Phys. Chem.* **1996**, *100*, 7384–7391.

(25) Burling, F. T.; Goldstein, B. M. Computational Studies of Nonbonded Sulfur-Oxygen and Selenium-Oxygen Interactions in the Thiazole and Selenazole Nucleosides. *J. Am. Chem. Soc.* **1992**, *114*, 2313–2320.

(26) Burling, F. T.; Goldstein, B. M. A Database Study of Nonbonded Intramolecular Sulfur-Nucleophile Contacts. *Acta Crystallogr., Sect. B: Struct. Sci.* **1993**, *49*, 738–744.

(27) Nicklaus, M. C.; Wang, S.; Driscoll, J. S.; Milne, G. W. A. Conformational changes of small molecules binding to proteins. *Bioorg. Med. Chem.* **1995**, *3*, 411–428.

(28) Hao, M.-H.; Haq, O.; Muegge, I. Torsion Angle Preference and Energetics of Small-Molecule Ligands Bound to Proteins. *J. Chem. Inf. Model.* **2007**, *47*, 2242–2252.

(29) Butler, K. T.; Luque, F. J.; Barril, X. Toward accurate relative energy predictions of the bioactive conformation of drugs. *J. Comput. Chem.* **2009**, *30*, 601–610.

(30) 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: Struct. Sci.* **2002**, *58*, 389–397.

(31) Brameld, K. A.; Kuhn, B.; Reuter, D. C.; Stahl, M. Small Molecule Conformational Preferences from Crystal Structure Data. A Medicinal Chemistry Focused Analysis. *J. Chem. Inf. Model.* **2008**, *48*, 1–24.

(32) 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.