Acta Crystallographica Section B
Structural Science
ISSN 0108-7681

Conformational variability of molecules in different crystal environments: a database study

Ze F. Weng,[a,b] W. D. Sam Motherwell,[b] Frank H. Allen[b]* and Jacqueline M. Cole[a]

[a]Cavendish Laboratory, J. J. Thomson Avenue, Cambridge CB3 0HE, England, and [b]Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England

Correspondence e-mail: allen@ccdc.cam.ac.uk

Received 29 November 2007
Accepted 26 February 2008

A methodology is described for analysing the Cambridge Structural Database (CSD) in terms of molecular conformations. Molecular species that have more than a single occurrence across the complete CSD are identified, either as the sole crystal component or co-crystallized with other components. Cluster analysis, based on a root-mean-square fit of coordinates and chemical connectivity, is performed to identify conformational variance for each molecule. Results are analysed in terms of the number of discrete conformations observed *versus* the number of crystal environments and number of acyclic torsion angles in the molecule. Special subsets of environments are also analysed, namely polymorphs, co-crystals and solvates. In general, conformational diversity increases with an increasing number of different crystal environments and with an increasing number of flexible torsion angles. Overall, molecules with one or more acyclic flexible torsion angle are observed to exist in more than one conformation in *ca* 40% of cases. There is evidence that solvated molecules exhibit more conformational flexibility on average, compared with polymorphs and co-crystals.

1. Introduction

A common concern in using molecular conformations from crystal structures in computational chemistry applications, *e.g.* in molecular modelling, drug discovery *etc.*, is that these conformations may be affected by crystal packing forces. An earlier analysis (Allen *et al.*, 1996) determined the conformational preferences exhibited by 12 common chemical substructures in crystal structures retrieved from the Cambridge Structural Database (CSD; Allen, 2002). These experimental observations were then compared with conformations computed for gas-phase model molecules using high-level *ab initio* methods. The comparison indicated:

(i) that crystal structure conformations are generally a good indication of conformational preferences, *i.e.* they generally lie close to an energy minimum in the potential energy hypersurface, a fundamental tenet of the principle of structure correlation (Bürgi & Dunitz, 1994), and

(ii) that high-energy conformers are rarely observed in crystal structures.

Nevertheless, Allen *et al.* (1996) noted that there are a number of well known examples which are exceptions to this general rule. For example, some molecules are capable of forming intermolecular interactions that are sufficiently strong to compete with intramolecular forces and induce strained conformations. A clear example is the α-amino acids, where ideal gas-phase minimum energy conformations are far removed from crystal conformations because of the importance of forming strong hydrogen bonds between molecules,

rather than a single intramolecular hydrogen bond as in the gas phase (Cooper *et al.*, 2007). Other exceptions to the general rule occur when a molecule can adopt an alternative conformation to that observed in crystal structures with relatively little increase in the conformational strain energy. A typical example is biphenyl with H substituents at all four *ortho* positions (Brock & Minton, 1989). Here, planar or nearly planar conformations appear to be favoured systematically in crystal structures, despite having energies that are *ca* 6 kJ mol^{-1} above the gas-phase minimum energy conformation which exhibits an inter-ring dihedral angle of 44°.

Despite these exceptions, the CSD has been used to carry out many in-depth studies of conformational preferences focusing on specific molecular fragments or chemical functional groups (see *e.g.* Allen *et al.*, 1991*a*,*b*,*c*; Allen & Motherwell, 2002; Harris *et al.*, 2001; Dalhus & Görbitz, 2000; Starbuck *et al.*, 1999; Shankland *et al.*, 1998; Bakaj & Zimmer, 1999; Zimmer, 2001; Shenkin & McDonald, 1994; Norskov-Lauritsen & Bürgi, 1985), all of which give credence to the general rules noted above. However, such studies are not comprehensive at the molecular level and this study focuses on the conformations of complete molecules within the CSD. Those molecular species have been selected that have more than a single occurrence across the complete CSD, either as the sole crystal component or co-crystallized with other components, and the conformational variance of these occurrences has been studied. This paper presents the methodology of the CSD analysis in detail, together with summary results describing the conformational variance of molecules in different crystal environments.

2. Methodology

2.1. Dataset selection

Crystal structures from release V5.28 of the CSD (November 2006: 390 081 entries) were initially filtered to retain only those structures[1] that:

(i) are classed as 'organic' within CSD definitions (this eliminates compounds of transition metals, lanthanides, actinides or any of Al, Ga, In, Tl, Ge, Sn, Pb, Sb, Bi, Po),

(ii) have full three-dimensional coordinates recorded for all non-H atoms,

(iii) are not disordered,

(iv) are not polymeric (*catena*) structures,

(v) have $R < 0.10$, and

(vi) have a perfect match of their chemical and crystallographic connectivities.

2.2. Determining the number of crystal environments (N_{env}) for a given molecule

2.2.1. Two-dimensional molecular graph matching. Each CSD structure, designated by a CSD reference code of six letters and two digits (*e.g.* ABCXYZ02), has a two-dimensional chemical connectivity representation for each molecular or ionic component in the asymmetric unit. Chemical atom properties are represented as:

(i) element-type,

(ii) number of connected non-H atoms,

(iii) number of attached terminal H atoms, and

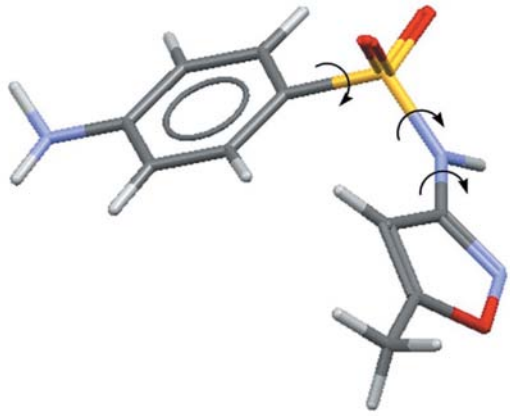
(iv) charge;

each chemical bond is assigned a coded CSD bond-type (Allen, 2002) and a cyclic/acyclic flag.

These molecular graphs are used to construct a list of equivalent two-dimensional connectivities for each unique component by comparing its graph representation against those for every other component in the retrieved dataset. Structures that are tautomers or have different protonation states, such as (4-(2-hydroxyethyl)-1-piperazine)ethanesulfonic acid (QIJZOY and QIJZOY01), are treated as different components.

For example, consider the molecule sulfamethoxazole (Fig. 1). Its chemical formula $C_{10}H_{11}N_3O_3S_1$ occurs in three CSD refcode families; two of these (SLFNMB, SMZTMP) contain sulfamethoxazole and one contains a structural isomer, a triazine derivative, which exists in the unique crystal environment ($N_{env} = 1$) of JABDOF. Chemical graph matching for sulfamethoxazole yields five occurrences in independent crystal contexts ($N_{env} = 5$): sulfamethoxazole itself occurs in four polymorphic forms having the same six-letter family code SLFNMB, which are further discriminated by the numerical suffixes 01, 02, 05, 06 for the four polymorphs. Sulfamethoxazole also occurs in a fifth crystal environment, as a co-crystal with trimethoprim in a structure having the refcode SMZTMP.

2.2.2. Duplicate structure determinations. To minimize statistical bias, duplicate structures – structures that are additional examples of an identical crystal environment for a



C10 H11 N3 O3 S1 | SMZTMP_2_0(i) | SLFNMB01_1_0(i) SLFNMB02_1_0(i) SLFNMB05_1_0(i) SLFNMB06_1_0(i)

Figure 1
Molecular structure of sulfamethoxazole in SMZTMP showing the three flexible acyclic torsion angles (see §2), together with the text representation of the conformation clustering from computed results record for the five occurrences of this molecule in different crystal environments in the CSD (see §3.1). The vertical bar '|' indicates boundaries of a cluster of molecules with the same conformation.

[1] Supplementary data for this paper are available from the IUCr electronic archives (Reference: BS5057). Services for accessing these data are described at the back of the journal.

given molecule – were removed. Duplicates may occur in the CSD as a result of the redetermination, reinterepretation or further refinement of structures. In the previous example of sulfamethoxazole, there are actually seven entries in the SLFNMB family, denoted as SLFNMB, SLFNMB01, SLFNMB02, ... SLFNMB06, of which three must be eliminated since these are duplicate determinations of one or other of the unique polymorphs. Recently, van de Streek (2006) has generated a list of the 'best (*i.e.* most accurate) examples' of each unique structure in the CSD, and this list was used to ensure the uniqueness of the crystal environments used in our conformational analysis.

2.2.3. Treatment of H and D atoms. H and D atoms require special treatment since some CSD entries do not contain explicit three-dimensional H coordinates, particularly in the earlier literature. Further, where present in X-ray diffraction studies, the three-dimensional H positions are often of low accuracy. In this analysis explicit H or D atoms were treated as implicit in the connectivity record, by encoding these atoms as a property of their connected atom. This approach means that conformational diversity due to substituents having only covalently bound H/D atoms is not considered here. However, such substituents, *e.g.* methyl groups, are known have low energy barriers to rotation (see *e.g.* Mo & Gao, 2007).

2.2.4. Stereoisomerism. For molecules which exist in two or more crystal environments (N_{env}), it is not guaranteed that each occurrence has the same stereochemistry. Thus, the three-dimensional structure of the two-dimensional connectivity graph of the first occurrence of a molecule is compared with those of successive molecules in the series to determine whether their stereochemistry is identical (i), enantiomeric (e) or diastereomeric (d). The algorithm for stereoisomer perception, based on the recognition of stereogenic centres, was implemented using the procedure outlined by Razinger *et al.* (1993) and is routinely applied within the CCDC's database building software to generate stereochemical cross-references that link the appropriate CSD entries. If a diastereomer was located in any molecule list generated *via* the two-dimensional connectivity comparison, it was used to construct a separate list for each different diastereomer. Thus, for the hexose sugars having the chemical formula $C_6H_{12}O_6$, this process will generate a separate molecule list, having its own N_{env}, for each diastereomeric hexose recorded in the CSD. Having separated out any diastereomers, stereoisomerism in the remaining list of N_{env} molecules is indicated by the addition of the flags (i) or (e) to indicate whether the CSD entry contains the original coordinates or the enantiomeric set. This indication was applied even for chiral molecules that crystallize in Söhncke space groups, since their absolute configurations may not have been determined by the diffraction experiment or properly assigned from external chemical evidence.

2.3. Conformational matching

The outcome of a conformational mapping *A*:*B* locates the best mapping of the three-dimensional atoms of molecule *B* onto those of molecule *A*. In some cases a two-dimensional structure will exhibit areas which have local topological symmetry (Allen *et al.*, 1991*b*), *e.g.* a phenyl substituent exhibits twofold symmetry about its connecting bond and the atom pairs in the *ortho* and *meta* positions are topologically equivalent. For these cases the appropriate topological symmetry permutations need to be applied to the three-dimensional atoms in order to find the optimum matching, *A*:*B*. Essentially, the atomic nomenclature of molecule *A* is imposed onto molecule *B* in order to obtain the best mapping *via* (1).

While conformational mapping can be performed visually, *e.g.* by using a visualiser such as *Mercury* (Macrae *et al.*, 2006) to recognize and pair up appropriate atoms, this approach is impractical for systematic work. For this study a novel automated algorithm was developed. The *Tormat* program, which locates the optimal pair-wise associated atom sets and provides a visual and graphical comparison of matched molecules, is fully described by Weng *et al.* (2008). In summary:

(i) Given two molecules *M*1 and *M*2 with equivalent two-dimensional molecular graphs and identical three-dimensional stereoisomerism, perform molecular matching of *M*2 upon *M*1, let the results be R_1, R_2, R_3, ... R_n.

(ii) For each match R_i, superimpose the resultant coordinates upon *M*1, calculate the r.m.s. deviation (d_{ij}^{rms}) of the fit.

(iii) The best atom-pair mapping $\{a_1, a_2, a_3, ...a_N\}$ to $\{a'_1, a'_2, a'_3, ...a'_N\}$ is given by the match R_i with the lowest d_{ij}^{rms}.

The r.m.s. deviation (d_{ij}^{rms}) of the atoms, for the optimal rigid-body superposition of structures *i* and *j*, is given by

$$d_{ij}^{rms} = \left(\frac{\sum_{k=1}^{N} \left(X_{ik} - X_{jk}\right)^2 + \left(Y_{ik} - Y_{jk}\right)^2 + \left(Z_{ik} - Z_{jk}\right)^2}{N} \right)^{1/2} \quad (1)$$

where *X*, *Y* and *Z* are the Cartesian coordinates of the corresponding atoms, *k*, in structures *i* and *j*, and *N* is the number of matching atoms used in the rigid-body superposition algorithm. The algorithm used to superimpose two sets of coordinates is that of Kearsley (1989). Step (i) is essentially performed as a substructure matching problem which is NP-complete, therefore, an upper limit of 1 000 000 molecular matchings is imposed for practical reasons (Barnard, 1993).

Note on torsion angles: In earlier analyses of the conformational variance of small chemical fragments determined using clustering techniques (*e.g.* Allen *et al.*, 1991*a*,*b*), torsion angles were used to calculate the Euclidian distance d_{ij} between pairs of conformers. For determining the conformational variance of complete molecules across a significant range of molecular sizes (see §2.4), clustering using d_{ij}^{rms} is found to be more effective. However, it is informative to discuss the results (§3) in terms of the conformational flexibility of the molecules as indicated by the number of freely rotatable acyclic bonds that they contain and, in some cases, the torsion angle values that are adopted. Acyclic torsion

angles in molecules were identified using the same procedures as those of the CSD program *Mogul* (Bruno *et al.*, 2004); note that for a torsion $A1-A2-A3-A4$, the bond $A2-A3$ must be acyclic, but the two peripheral bonds $A1-A2$ and $A3-A4$ may either be acyclic or cyclic.

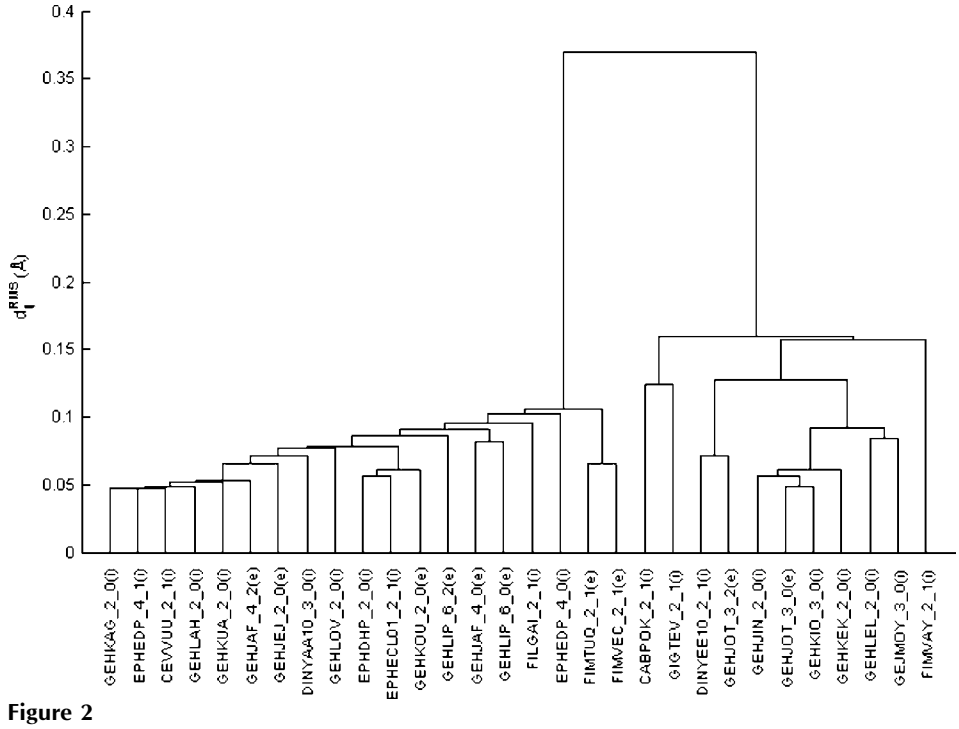


Figure 2
An example of a hierarchy tree for the 44 observations of ephedrine constructed using the single-linkage method.

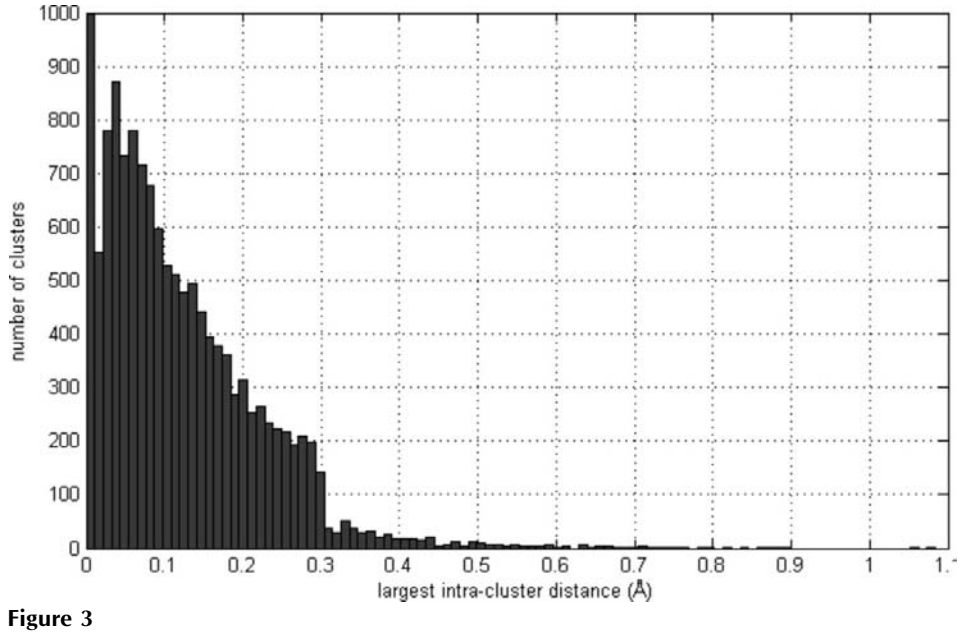


Figure 3
Histogram of the maximum intra-cluster distance *versus* the number of clusters for all molecular components in the CSD, each bar representing a range of 0.01 Å. At a cut-off value of $d_{ij}^{rms} = 0.30$ Å using the single linkage method, 111 379 out of 111 889 observations (99.54%) have a maximum intra-cluster distance of ≤ 0.30 Å, and 48 observations (0.04%) have a maximum intra-cluster distance of ≥ 0.60 Å. At the intra-cluster distance of 0–0.01 Å (truncated), the large number of observations (105 414) is mainly due to molecules having just a single observation.

2.4. Conformational clustering

Applying the described procedures generates a list of molecules with the best conformational matchings. Agglomerative, single-linkage clustering (Everitt *et al.*, 2001) based on d_{ij}^{rms} was used to partition the lists into classes of similar conformations using specially built code optimized for speed. The complete linkage method was also trialled in the analysis to try to force the formation of tight clusters with well defined boundaries. However, in view of the difficulty in selecting a suitable clustering cut-off value, the complete linkage method appeared less suitable than the single linkage approach when applied to whole-molecule conformations.

Once the hierarchy tree was constructed from the single-linkage clustering method (see *e.g.* Fig. 2 for the $N_{env} = 44$ observations of ephedrine), a simple d_{ij}^{rms} cut-off value was used to partition the observations into clusters, each of which represents a 'different' conformation. The definition of 'different' is subjective, and ideally a cut-off value should be chosen such that the observations are partitioned into conformational clusters, each of which corresponds to an energy minimum in the potential energy hypersurface. However, this process is complicated in the general case where the molecular class and size are not defined, and we generally have a small number of observations (N_{env}) per molecule. In the event, a d_{ij}^{rms} cut-off value was selected manually by examination of the conformational clusters formed for different cut-off values for a test set of 500 randomly selected molecules. It was found that a cut-off value of $d_{ij}^{rms} = 0.30$ Å for the single linkage algorithm is effective (Figs. 2 and 3), and applying this cut-off value to the hierarchy tree for ephedrine (Fig. 2) yields two clusters. Thorough examination of a wide variety of clusters in terms of intra-cluster torsion-angle differences showed that although there is no simple relationship between

d_{ij}^{rms} and the torsion angle differences for a given pair of molecules (i,j), we find that d_{ij}^{rms} = 0.30 Å corresponds to a maximum torsion angle difference of $\sim 30^\circ$.

A common problem associated with the single-linkage method is the occurrence of ellipsoidal cluster shapes or the 'chaining' of fairly distinct clusters (Allen *et al.*, 1991*b*; Leach, 1994). The maximum intra-cluster distance plotted against the number of clusters formed for the complete CSD dataset is given in Fig. 3. Overall, the chaining effect is observed as not significant: only 96 clusters (0.09%) contain a maximum intra-cluster distance $\geq$ 0.5 Å, and in the vast majority of clusters each member is linked to another by d_{ij}^{rms} < 0.30 Å. Possible factors for the effectiveness of the single-linkage algorithm are

(i) the relatively small number of conformers in any molecule list, and

(ii) that high-energy conformers are rarely observed in crystal structures (see *e.g.* Allen *et al.*, 1996).

When such conformers are found they can approach the top of an energy barrier and so provide a linkage to other conformational cluster(s) on the other side of the barrier. This 'chaining' effect may occasionally also be observed in molecules with very low energy barriers, such as mandelic acid which has a low barrier to carboxylate rotation, and which exhibits a maximum intra-cluster distance of 0.840 Å for one of the clusters.

To gain some visual insight into the conformational differences between typical cluster members, superimposed structure pairs for the ephedrine example are shown in Fig. 4. Fig. 4(*a*) illustrates the importance of using the inversion operator when comparing enantiomeric pairs. In describing these conformational comparisons, we use the three torsion angles τ_1 = $H_3C-NH_2-C-CH_3$, τ_2 = $H_2N-C-C-OH$ and τ_3 = $HO-C-C_{ar}-C_{ar}$ (C_{ar} is aromatic carbon), with data given in the order of citation of the refcodes below. Thus, Fig. 4(*a*) shows that the conformation of (−)-ephedrine (EPHDHP) is essentially the same as that of the enantiomer (FIMVEC). Here, d_{ij}^{rms} = 0.228 Å, τ_1 = 177.3, −162.8° ($\Delta\tau_1$ = 19.9°), τ_2 = −73.6, −85.0° ($\Delta\tau_2$ = 11.4°), and τ_3 = −21.4, −21.0° ($\Delta\tau_3$ = 0.4°). These two molecules were judged as having the same conformation by the hierarchy tree shown in Fig. 2. However, if EPHDHP is compared with FIMVAY, a co-crystal of (−)-ephedrine with (+)-4-*ortho*-chlorophenyl-5,5-dimethyl-2,2-dioxo-1,3,2-dioxaphosphorinate (Fig. 4*b*), it can be seen that the conformational difference lies in the rotation of the methylamino group: d_{ij}^{rms} = 0.505 Å, τ_1 = 177.3, −72.2° ($\Delta\tau_1$ = 74.9°), τ_2 = −73.6, −71.7° ($\Delta\tau_2$ = 1.9°), and τ_3 = −21.4, −17.4° ($\Delta\tau_3$ = 4.0°).

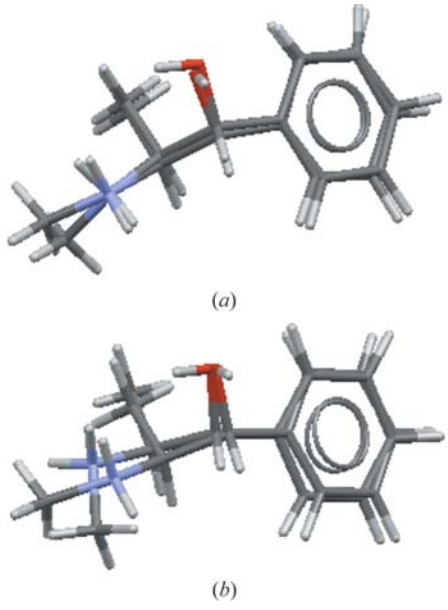


Figure 4
Superimposed molecules of ephedrine for (*a*) EPHDHP and FIMVEC in the same conformational cluster, and (*b*) EPHDHP and FIMVAY in different clusters.

3. Results and discussion

3.1. Generation and presentation of results

The procedures detailed in §2 operated on 112 816 organic crystal structures comprising 173 942 molecular 'components'. Of these, 512 molecular components (0.29%) were rejected because they required more than the maximum number of connectivity matches, or the setereochemistry could not be uniquely determined. Each line in the final computed results list corresponds to a unique molecular component, and identifies each of the N_{env} unique crystal environments as ZZZZZZnn_*o*_*p*(q). Here, ZZZZZZnn is the CSD refcode, '*o*' is the number of chemical components in the crystal, and '*p*' is the component number in the crystal as recorded in the CSD. Finally, '(q)' is a code letter (i) or (e) indicating if the original or enantiomeric structure must be used in the conformational comparison. After the assignment of molecules to conformational clusters, each partition into a new cluster is shown by the vertical bar character '|'. The legend to Fig. 1 shows an example of the final results list for the N_{env} = 5 examples of sulfamethoxazole.

In tabulating and discussing the results, the following nomenclature was used: N_{env} is the number of crystal environments, N_{obs} is the number of molecules that occur in each of the N_{env} environments, N_{conf} is the number of conformations adopted by the N_{obs} molecules, and %conf is the percentage of N_{obs} that adopt N_{conf} different conformational clusters. Thus, in line 3 of Table 2(*a*), 1687 molecules exist in N_{env} = 3 crystal environments and of these 57.0% adopt just one conformation, 26.4% adopt two different conformations and 16.6% adopt three different conformations.

3.2. Summary tables of molecular conformational variance

Table 1 shows the numbers of organic molecular components (N_{obs}) that exist in N_{env} different crystal environments in the CSD. This table, and other representations of the data, is dominated by the 87 394 molecules that exist in a single environment. Beyond that, there is a rapid fall-off from the 13 103 molecules that exist in two different environments, such that only 1687 molecules occur in three environments, and fewer than 60 molecules occur in any of the subdivisions

Table 1
The number of molecular components (N_{obs}) in organic crystal structures which exist in N_{env} different crystal environments.

Molecular formulae for components that occur in more than 30 environments are also given.

N_{env}	N_{obs}	Molecular formula	N_{env}	N_{obs}	Molecular formula
1	87394		52	2	C2 H2 O4
2	13103				C12 H8 N2
3	1687		53	2	C14 H19 N2 1+
4	972				C3 H7 N6 1+
5	270		54	1	H1 Cl1
6	221		55	2	C6 H15 N4 O2 1+
7	101				C12 H10 N2
8	110		56	1	C6 H6 N2 O1
9	70		57	1	C2 O4 2−
10	58		58	1	H1 O1 1−
11	55		60	2	C12 H28 N1 1+
12	44				F1 1−
13	27		61	3	C2 H6 N1 O4 S2 1−
14	22				C3 H5 N2 1+
15	21				C6 H3 N3 O6
16	18		63	1	C1 H3 O3 S1 1−
17	20		64	3	C2 H8 N1 1+
18	11				C4 H12 N2 2+
19	13				C6 H14 N1 1+
20	12		67	1	C36 H30 N1 P2 1+
21	9		68	1	C6 H4 S4
22	11		72	1	C2 F3 O2 1−
23	7		73	1	C18 H15 O1 P1
24	10		75	1	H12 Mg1 O6 2+
25	10		77	1	C10 H8 S8 1+
26	8		80	1	C8 H7 O3 1−
27	7		85	2	C24 H40 O5
28	8				C2 H3 O2 1−
29	8		86	1	H1 O4 S1 1−
30	4	C2 H6 N1 O2 1+	89	1	C20 H14 O2
		C3 H8 O1	93	1	C4 H3 O4 1−
		C14 H10	95	2	C8 H10; Rb1 1+
		C4 H12 N1 1+	96	1	C12 H24 N1 1+
31	3	C14 H19 N4 O3 1+	97	1	As1 F6 1−
		C20 H36 O6	98	1	C2 H1 O4 1−
		C6 H12	101	1	C24 H20 As1 1+
32	6	C2 H4 Cl2	104	1	C6 H14
		C31 H30 O4	106	1	C6 H16 N1 1+
		C10 H2 O6	107	1	C3 H8 O1
		C12 F4 N4	108	1	C1 N1 S1 1−
		C6 H4 O2	111	1	C5 H6 N1 1+
		C7 H7 N1 O2	112	1	C2 H10 N2 2+
33	3	C4 H10 O2	113	1	Ca1 2+
		C15 H12 N2 O1	118	1	C6 H6 O2
		C9 H14 N1 1+	121	1	C4 H5 O6 1−
34	5	C32 H32 O8	128	1	H3 O1 1+
		C4 H4 N2; C4 O4 2−	136	1	C5 H5 N1
		C7 H4 N1 O4 1−; N3 1−	145	1	I2
35	3	C1 H2 O2	149	1	C1 H4 N2 S1
		C12 H14 N2 2+	151	1	C8 H20 N1 1+
		C7 H4 N1 O3 S1 1−	152	1	O4 S1 2−
36	3	C18 H36 K1 N2 O6 1+	153	1	Li1 1+
		C8 H12 N2 O3	154	1	C2 H4 O2
		C12 H9 N2 1+	156	1	C12 H24 O6
37	3	C24 H41 N1 O4	157	1	C4 H8 O2
		C6 F4 I2	158	1	C10 H8 N2
		F6 Si1 2−	159	1	C4 H10 O1
38	4	C4 H5 O5 1−	164	1	C7 H7 O3 S1 1−
		C12 H12 N2	176	1	C1 H4 N2 O1
		C7 H11 N2 1+	181	1	H2 O4 P1 1−
		C7 H5 O2 1−	182	1	Cs1 1+
39	2	C30 H22 O2	187	1	C4 H12 N1 1+
		C2 H5 N1 O2	193	1	C16 H36 N1 1+
40	4	C14 H16 N1 1+	195	1	C8 H12 N1 1+
		C1 Cl4; H1 F1	222	1	C2 H6 O1 S1
		Sr1 2+	228	1	C24 H20 B1 1−
41	3	C4 H6 O4	239	1	C4 H8 O2
		C4 H3 O4 1−	249	1	C7 H8
		C3 H10 N1 1+	258	1	C6 H2 N3 O7 1−
42	2	C6 H12 N4	281	1	C1 H6 N3 1+
		H1 O4 P1 2−	286	1	C4 H8 O1
43	6	C12 H24 K1 O6 1+	291	1	C24 H20 P1 1+
		C4 H12 N1 1+	310	1	C12 H4 N4
		C10 H16 N1 O1 1+	314	2	H4 N1 1+
		C1 H3 N1 O2			I3 1−
		C6 H18 N3 S1 1+	327	1	C3 H7 N1 O1
		C7 H10 N1 1+	335	1	B1 F4 1−
44	1	C12 H8 N2	368	1	F6 P1 1−
45	1	C4 H10 N1 O1 1+	385	1	C1 F3 O3 S1 1−
46	3	Br3 1−	387	1	N1 O3 1−
		C19 H18 P1 1+	420	1	C2 H6 O1
		C5 H12 N1 1+	488	1	C3 H6 O1
47	2	Ba1 2+	564	1	K1 1+
		C1 H6 N1 1+	629	1	Na1 1+
48	3	C4 H1 O4 1−	648	1	C1 H1 Cl3
		C6 H6 O1	694	1	C1 H2 Cl2
		C7 H4 N2 O6	722	1	C2 H3 N1
49	1	C4 H4 O6 2−	781	1	C6 H6
50	4	C6 N4	928	1	Cl1 O4 1−
		C1 H1 O2 1−	1140	1	I1 1−
		C16 H10	1176	1	C1 H4 O1
		C7 H3 N2 O6 1−	1643	1	Br1 1−
51	3	C4 H4 O4	3475	1	Cl1 1−
		C1 S2	12 520	1	H2 O1
		H3 N1			

corresponding to ten or more crystal environments. Beyond N_{env} = 30, Table 1 lists the molecular formulae of chemical components that occur in very large numbers of crystal environments, and it is no surprise that the higher reaches of Table 1 contain common solvent molecules and counter-ions including, for example, dichloromethane (N_{env} = 694), methanol (N_{env} = 1176), and culminating with water which occurs in 12 520 different crystal environments in this subset of the CSD. Some common co-crystal agents are also represented at higher N_{env}, *e.g.* 7,7,8,8-tetracyanoquinodimethane ($C_{12}H_4N_4$) with N_{env} = 310.

Table 2(*a*) displays the overall conformational variance exhibited by all 103 986 molecules having $N_{env} \leq 10$ (the total sample of all molecules is 104 250). Excluding those molecules in a unique environment, 64% of the remaining 16 866 molecules having N_{env} from 2 to 30 adopt a single conformation, while 36% adopt ≥ 2 conformations. However, Table 2(*a*) includes structures having no acyclic torsion angles (N_{tor} = 0), a significant subset of 16 112 molecules whose conformational diversity is summarized in Table 2(*b*) (15 882 with $N_{env} \leq 10$). Of the 3360 molecules that exist in $N_{env} > 1$, 87% exhibit a single conformation. This is unsurprising since this subset is dominated by simple and polycyclic aromatics. The remaining 13% of molecules which do exist in more than one conformation principally comprise a wide variety of non-aromatic ring systems that are discussed below.

While conformational clustering has been accomplished using d_{ij}^{rms}, thus identifying conformational diversity due to variations in both cyclic and acyclic torsion angles, it is chemically useful to summarize the observed conformational diversity by classifying molecules according to the number of

Table 2
Distribution of molecular conformations across different crystal environments for organic molecules in the CSD.

Column headings are explained in §3.1. All tables are truncated at $N_{env} \leq 10$.

N_{env}	N_{obs}	N_{conf}						
		1	2	3	4	5	6	≥ 7
(*a*) All molecular components								
1	87 394	87 394 (100%)	0	0	0	0	0	0
2	13 103	8511 (65.0%)	4592 (35.0%)	0	0	0	0	0
3	1687	962 (57.0%)	445 (26.4%)	280 (16.6%)	0	0	0	0
4	972	579 (59.6%)	228 (23.5%)	93 (9.6%)	72 (7.4%)	0	0	0
5	270	154 (57.0%)	53 (19.6%)	27 (10.0%)	22 (8.1%)	14 (5.2%)	0	0
6	221	141 (63.8%)	37 (16.7%)	19 (8.6%)	8 (3.6%)	10 (4.5%)	6 (2.7%)	0
7	101	65 (64.4%)	20 (19.8%)	6 (5.9%)	3 (3.0%)	1 (1.0%)	5 (5.0%)	1 (1.0%)
8	110	59 (53.6%)	28 (25.5%)	9 (8.2%)	5 (4.5%)	3 (2.7%)	4 (3.6%)	1 (0.9%)
9	70	42 (60.0%)	11 (15.7%)	3 (4.3%)	6 (8.6%)	3 (4.3%)	3 (4.3%)	0
10	58	35 (60.3%)	8 (13.8%)	5 (8.6%)	4 (6.9%)	3 (5.2%)	2 (3.4%)	0
(*b*) All molecular components having $N_{tor} = 0$								
1	12 682	12 682 (100%)	0	0	0	0	0	0
2	2247	1944 (86.5%)	303 (13.5%)	0	0	0	0	0
3	448	375 (83.7%)	57 (12.7%)	16 (3.6%)	0	0	0	0
4	228	204 (89.5%)	12 (5.3%)	9 (3.9%)	3 (1.3%)	0	0	0
5	83	69 (83.1%)	7 (8.4%)	1 (1.2%)	2 (2.4%)	4 (4.8%)	0	0
6	74	70 (94.6%)	2 (2.7%)	1 (1.4%)	0	1 (1.4%)	0	0
7	42	40 (95.2%)	1 (2.4%)	1 (2.4%)	0	0	0	0
8	34	27 (79.4%)	5 (14.7%)	1 (2.9%)	1 (2.9%)	0	0	0
9	23	20 (87.0%)	1 (4.3%)	1 (4.3%)	0	0	1 (4.3%)	0
10	21	19 (90.5%)	1 (4.8%)	1 (4.8%)	0	0	0	0
(*c*) All molecular components having $N_{tor} \geq 1$								
1	74 712	74 712 (100%)	0	0	0	0	0	0
2	10 856	6567 (60.5%)	4289 (39.5%)	0	0	0	0	0
3	1239	587 (47.4%)	388 (31.3%)	264 (21.3%)	0	0	0	0
4	744	375 (50.4%)	216 (29.0%)	84 (11.3%)	69 (9.3%)	0	0	0
5	187	85 (45.5%)	46 (24.6%)	26 (13.9%)	20 (10.7%)	10 (5.3%)	0	0
6	147	71 (48.3%)	35 (23.8%)	18 (12.2%)	8 (5.4%)	9 (6.1%)	6 (4.1%)	0
7	59	25 (42.4%)	19 (32.2%)	5 (8.5%)	3 (5.1%)	1 (1.7%)	5 (8.5%)	1 (1.7%)
8	76	32 (42.1%)	23 (30.3%)	8 (10.5%)	4 (5.3%)	3 (3.9%)	4 (5.3%)	1 (1.3%)
9	47	22 (46.8%)	10 (21.3%)	2 (4.3%)	6 (12.8%)	3 (6.4%)	2 (4.3%)	0
10	37	16 (43.2%)	7 (18.9%)	4 (10.8%)	4 (10.8%)	3 (8.1%)	2 (5.4%)	0

acyclic torsion angles (N_{tor}). Table 3 shows the overall distribution of molecules across all degrees of acyclic torsional freedom (N_{tor} = 1–77) determined in this study. Notably, the vast majority (85%) of molecules have at least one flexible acyclic torsion angle. The number of molecules with increasing acyclic torsion angles falls off rapidly as expected, with the exception of those with two torsions where there is an increase over those with one. Accumulating totals up to just three torsion angles encompasses 52% of the data, and accumulating to eight torsion angles covers 89%. Counting the number of acyclic torsion angles is of course a very crude measure of potential conformational variability, but we offer it as a parameter which should in general permit more local minima to occur in the conformational energy surface. The statistics take no account of the nature of a torsion angle, for example the single torsion angle in a biphenyl arguably has more effect on total molecular shape than *e.g.* the rotation of a terminal methoxy group on a naphthalene nucleus. Total molecular shape may have more importance in allowing packing energy to dominate in the competition between packing-energy minimization and internal molecular conformation energy.

Table 2(*c*) shows the overall breakdown of conformational diversity for all molecules having $N_{tor} \geq 1$. For those molecules that exist in more than one crystal environment, the extra degree(s) of conformational freedom over the (ring-limited) $N_{tor} = 0$ dataset of Table 2(*b*) is clear: in every case for N_{tor} = 2–10 in Table 2(*c*) we see that the percentage of molecules showing just one conformation is less for these more flexible molecules. For example, for $N_{env} = 2$, the 86.5% of Table 2(*b*) drops to 60.5% in Table 2(*c*). Furthermore, the percentage of molecules exhibiting just a single conformation is almost constant in the range 80–90% over all values of N_{env} in Table 2(*b*), but for molecules having $N_{tor} \geq 1$ (Table 2*c*) the percentage of single conformations falls from the 60.5% at $N_{env} = 2$, to around 45% for higher N_{env} values. Summation of the total dataset shows that 87.1% of molecules having $N_{tor} = 0$ exhibit just one conformation, but for molecules having $N_{tor} \geq 1$ this metric falls to 58.1%.

The probability of a molecule with a specific degree of torsional freedom to exhibit more than one conformation is obviously correlated positively with the number of opportunities to do so, *i.e.* with the number of crystal environments, N_{env}, reported to date in the CSD. The number of molecules

Table 3
Molecular flexibility: number of observations (N_{obs}) having N_{tor} acyclic torsion angles.

N_{tor}	N_{obs}	N_{tor}	N_{obs}
0	16 112	32	23
1	10 691	33	10
2	15 330	34	15
3	12 577	35	6
4	13 185	36	33
5	8445	37	5
6	8086	38	4
7	4629	39	8
8	4414	40	11
9	2637	41	9
10	2063	42	10
11	1202	43	1
12	1498	44	10
13	674	45	2
14	589	46	2
15	428	47	2
16	507	48	11
17	207	49	1
18	258	50	2
19	109	51	1
20	171	52	2
21	58	53	2
22	88	54	3
23	52	56	1
24	107	59	3
25	26	60	1
26	29	62	1
27	22	64	1
28	34	66	2
29	16	69	2
30	30	77	1
31	12	–	–

observed having $N_{env} = 1, 2, 3$ *etc.* falls off rapidly, showing that the focus of attention has been on individual molecules rather than on series of solvates or co-crystals. It might be expected that the proportion of molecules with $N_{env} = 2$ will decrease as more experiments reveal cases of $N_{env} = 3$ or higher. However, as the number of opportunities to find another conformation increases, *i.e.* as N_{env} increases, a rather surprising result emerges: the overall percentage of single conformations remains approximately constant at around 60% for the all-molecule sample, 87% for the $N_{tor} = 0$ set, and 47% for the $N_{tor} \geq 1$ set (see Table 2). This implies that even when more experiments are performed the above percentages will not change greatly, and are an approximate guide to the likelihood of more than one conformation being observed.

A further factor in any attempt to predict the number of conformers that a molecule will assume in crystal environments is that for N_{conf} conformations to be observed we must have $N_{env} \geq N_{conf}$. Thus, if all molecules having $N_{conf} = 2$ (5422) are selected and all molecules having $N_{env} \geq 2$ (16 952) are counted, then 32.0% exhibit exactly $N_{conf} = 2$. This normalization gives the likelihoods of formation of $N_{conf} = 2$–5 as 32, 16, 10 and 4%, respectively, but the numbers of observations for $N_{conf} = 4$ and 5 is rather too small to have statistical significance. These figures do not take into account the number of flexible torsion angles per molecule, but indicate that provided there is at least one flexible torsion angle (Table 2*c*) we can expect about 40% to show more than one conformation. This affords some measure of the fundamental problem facing crystal structure prediction for flexible molecules, although progress is being made for molecules with $N_{tor} = 2$ (Cooper *et al.*, 2007).

3.3. Patterns of conformational change

The analysis has generated a number of conformational diversity files classified on the basis of *e.g.* N_{tor}, or on chemical or crystal structure type. Obviously, in developing the complex automated procedure described in §2, numerous spot checks have been performed using the programs *Mercury* (Macrae *et al.*, 2006) and *Tormat* (Weng *et al.*, 2008). The following subsections summarize some immediate observations, but more complete analyses are currently in progress and will be presented for publication in the near future.

3.3.1. Molecules with $N_{tor} = 0$. Analysis of Table 2(*b*) shows that 12.9% of cyclic molecules, *i.e.* parent rings and those with only single-atom substituents, show conformational variety. The most significant observed conformational variability arises in large-ring compounds such as crown ethers and their analogues. Such rings are well known to be highly flexible with many conformations that fall into shallow minima separated by low-energy barriers in the relevant conformational hypersurface. Thus, dibenzo-24-crown-8 adopts 11 distinct conformations across 20 examples of the molecule and 18-crown-6 adopts 10 conformations across 156 occurrences.

Carbocyclic and heterocyclic rings of size ≥ 7 are also well represented, albeit with less variability than their larger analogues. NIMCUH is typical of the medium rings: there are two molecules in the asymmetric unit and all four conformationally flexible dithiacycloheptane rings adopt twist-chair (TC) conformations. However, in one of the independent molecules, one of the rings adopts a different TC variant along the relevant pseudo-rotation itinerary, at a very limited energy cost to the complete molecule (Allen *et al.*, 1993), thus placing the two molecules in different conformational clusters. Another group of compounds that is represented in Table 2(*b*) is strained ring assemblies involving fusion and bridging, *e.g.* FAJSIS, in which the cyclohexene ring adopts an envelope conformation, in contrast to the boat conformation adopted by this ring in the monohydrate (FAJSOY). These two cyclohexene conformers differ in energy by only 1.7 kJ mol^{-1} (Bucourt & Hainaut, 1965).

An interesting 'interloper' into this subset is 1,1,4,4-tetramethylbut-2-yne-1,4-diol. Although acyclic, the molecule does not contain any acyclic torsions within our definitions (§2.3) due to the linearity of the four-carbon backbone. Conformational variability is akin to that of *e.g.* ethane, but with rotation of the Csp^3 centres about a three-bond linear chain. The parent molecule (NOXKAM, $Z' = 2$) has the two —OH groups in a *trans* arrangement, while in the other they are *gauche*. In its co-crystal with 1,3,5-tris(1-hydroxy-1-methylethyl)benzene (GAWSAY), one molecule again adopts the *trans*-OH conformation, while in the other molecule the two OH groups are each only $\sim 20^\circ$ away from being eclipsed with a methyl substituent. It is this conformational flexibility that

Table 4
Distribution of conformations for flexible organic molecules ($N_{tor} \leq 9$) across different crystal environments in the CSD.

Column headings are defined in §3.1. All tables are truncated at $N_{env} \leq 10$.

N_{env}	N_{obs}	N_{conf}						
		1	2	3	4	5	6	≥ 7
(*a*) Molecules having $N_{tor} = 1$								
1	8589	8589 (100%)	0	0	0	0	0	0
2	1550	1299 (83.8%)	251 (16.2%)	0	0	0	0	0
3	222	171 (77.0%)	40 (18.0%)	11 (5.0%)	0	0	0	0
4	131	102 (77.9%)	27 (20.6%)	2 (1.5%)	0	0	0	0
5	33	31 (93.9%)	2 (6.1%)	0	0	0	0	0
6	37	33 (89.2%)	4 (10.8%)	0	0	0	0	0
7	13	9 (69.2%)	4 (30.8%)	0	0	0	0	0
8	21	14 (66.7%)	7 (33.3%)	0	0	0	0	0
9	12	10 (83.3%)	2 (16.7%)	0	0	0	0	0
10	12	9 (75.0%)	2 (16.7%)	1 (8.3%)	0	0	0	0
(*b*) Molecules having $N_{tor} = 2$								
1	12 598	12 598 (100%)	0	0	0	0	0	0
2	2088	1530 (73.3%)	558 (26.7%)	0	0	0	0	0
3	274	163 (59.5%)	84 (30.7%)	27 (9.9%)	0	0	0	0
4	143	92 (64.3%)	38 (26.6%)	8 (5.6%)	5 (3.5%)	0	0	0
5	52	27 (51.9%)	16 (30.8%)	5 (9.6%)	4 (7.7%)	0	0	0
6	38	23 (60.5%)	11 (28.9%)	3 (7.9%)	1 (2.6%)	0	0	0
7	20	10 (50.0%)	7 (35.0%)	2 (10.0%)	1 (5.0%)	0	0	0
8	19	8 (42.1%)	7 (36.8%)	3 (15.8%)	1 (5.3%)	0	0	0
9	13	8 (61.5%)						
(*c*) Molecules having $N_{tor} = 3$								
1	10 548	10 548 (100%)	0	0	0	0	0	0
2	1579	1039 (65.8%)	540 (34.2%)	0	0	0	0	0
3	187	104 (55.6%)	62 (33.2%)	21 (11.2%)	0	0	0	0
4	118	71 (60.2%)	32 (27.1%)	10 (8.5%)	5 (4.2%)	0	0	0
5	36	16 (44.4%)	11 (30.6%)	7 (19.4%)	1 (2.8%)	1 (2.8%)	0	0
6	21	8 (38.1%)	6 (28.6%)	4 (19.0%)	3 (14.3%)	0	0	0
7	10	2 (20.0%)	6 (60.0%)	1 (10.0%)	0	1 (10.0%)	0	0
8	12	6 (50.0%)	4 (33.3%)	1 (8.3%)	0	1 (8.3%)	0	0
9	8	3 (37.5%)	2 (25.0%)	0	1 (12.5%)	1 (12.5%)	1 (12.5%)	0
10	5	2 (40.0%)	2 (40.0%)	0	0	1 (20.0%)	0	0
(*d*) Molecules having $N_{tor} = 4$								
1	11 161	11 161 (100%)	0	0	0	0	0	0
2	1613	931 (57.7%)	682 (42.3%)	0	0	0	0	0
3	185	61 (33.0%)	74 (40.0%)	50 (27.0%)	0	0	0	0
4	123	48 (39.0%)	43 (35.0%)	20 (16.3%)	12 (9.8%)	0	0	0
5	29	5 (17.2%)	11 (37.9%)	4 (13.8%)	6 (20.7%)	3 (10.3%)	0	0
6	14	2 (14.3%)	5 (35.7%)	2 (14.3%)	2 (14.3%)	3 (21.4%)	0	0
7	11	3 (27.3%)	2 (18.2%)	1 (9.1%)	2 (18.2%)	0	3 (27.3%)	0
8	9	2 (22.2%)	2 (22.2%)	0	2 (22.2%)	0	2 (22.2%)	1 (11.1%)
9	8	1 (12.5%)	1 (12.5%)	0	4 (50.0%)	1 (12.5%)	0	0
10	5	0	0	1 (20.0%)	0	1 (20.0%)	2 (40.0%)	0
(*e*) Molecules having $N_{tor} = 5$								
1	7227	7227 (100%)	0	0	0	0	0	0
2	1031	549 (53.2%)	482 (46.8%)	0	0	0	0	0
3	90	28 (31.1%)	33 (36.7%)	29 (32.2%)	0	0	0	0
4	61	24 (39.3%)	21 (34.4%)	8 (13.1%)	8 (13.1%)	0	0	0
5	10	2 (20.0%)	1 (10.0%)	1 (10.0%)	6 (60.0%)	0	0	0
6	7	2 (28.6%)	2 (28.6%)	2 (28.6%)	0	1 (14.3%)	0	0
7	1	0	0	0	0	0	1 (100%)	0
8	5	1 (20.0%)	1 (20.0%)	1 (20.0%)	0	1 (20.0%)	0	0
9	0	0	0	0	0	0	0	0
10	4	1 (25.0%)	0	1 (25.0%)	1 (25.0%)	1 (25.0%)	0	0
(*f*) Molecules having $N_{tor} = 6$								
1	6972	6972 (100%)	0	0	0	0	0	0
2	920	460 (50.0%)	460 (50.0%)	0	0	0	0	0
3	93	27 (29.0%)	34 (36.6%)	32 (34.4%)	0	0	0	0
4	58	13 (22.4%)	22 (37.9%)	12 (20.7%)	11 (19.0%)	0	0	0
5	13	3 (23.1%)	2 (15.4%)	4 (30.8%)	2 (15.4%)	2 (15.4%)	0	0
6	11	1 (9.1%)	2 (18.2%)	2 (18.2%)	1 (9.1%)	2 (18.2%)	3 (27.3%)	0
7	1	1 (100%)	0	0	0	0	0	0

Table 4 (continued)

N_{env}	N_{obs}	N_{conf}						
		1	2	3	4	5	6	≥ 7
8	4	1 (25.0%)	1 (25.0%)	0	0	1 (25.0%)	1 (25.0%)	0
9	4	0	1 (25.0%)	1 (25.0%)	0	0	1 (25.0%)	0
10	3	1 (33.3%)	1 (33.3%)	0	1 (33.3%)	0	0	0
(*g*) Molecules having N_{tor} = 7								
1	4031	4031 (100%)	0	0	0	0	0	0
2	508	232 (45.7%)	276 (54.3%)	0	0	0	0	0
3	47	9 (19.1%)	18 (38.3%)	20 (42.6%)	0	0	0	0
4	27	8 (29.6%)	7 (25.9%)	6 (22.2%)	6 (22.2%)	0	0	0
5	3	0	3 (100%)	0	0	0	0	0
6	7	1 (14.3%)	2 (28.6%)	1 (14.3%)	0	0	3 (42.9%)	0
7	0	0	0	0	0	0	0	0
8	4	0	1 (25.0%)	1 (25.0%)	1 (25.0%)	0	1 (25.0%)	0
9	1	0	0	0	0	1 (100%)	0	0
10	1	0	0	0	1 (100%)	0	0	0
(*h*) Molecules having N_{tor} = 8								
1	3870	3870 (100%)	0	0	0	0	0	0
2	460	182 (39.6%)	278 (60.4%)	0	0	0	0	0
3	42	8 (19.0%)	11 (26.2%)	23 (54.8%)	0	0	0	0
4	32	8 (25.0%)	10 (31.3%)	6 (18.8%)	8 (25.0%)	0	0	0
5	3	0	0	1 (33.3%)	1 (33.3%)	1 (33.3%)	0	0
6	3	0	1 (33.3%)	1 (33.3%)	0	1 (33.3%)	0	0
7	1	0	0	0	0	0	1 (100%)	0
8	1	0	0	1 (100%)	0	0	0	0
9	1	0	1 (100%)	0	0	0	0	0
10	0	0	0	0	0	0	0	0
(*i*) Molecules having N_{tor} = 9								
1	2317	2317 (100%)	0	0	0	0	0	0
2	274	99 (36.1%)	175 (63.9%)	0	0	0	0	0
3	22	7 (31.8%)	8 (36.4%)	7 (31.8%)	0	0	0	0
4	15	2 (13.3%)	5 (33.3%)	3 (20.0%)	5 (33.3%)	0	0	0
5	3	0	0	2 (66.7%)	0	1 (33.3%)	0	0
6	1	0	0	1 (100%)	0	0	0	0
7	0	0	0	0	0	0	0	0
8	1	0	0	1 (100%)	0	0	0	0
9	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0

underpins the value of alkynols of this general type in the assembly of hydrogen-bonded complexes of transition metals (Braga *et al.*, 1997).

3.3.2. Molecules with $N_{tor} \geq 1$. Data for these molecules are summarized in Table 2(*c*) and represent 80.1% of the organic molecules in this dataset that can exhibit conformational diversity, whether from rotations about acyclic bonds or from changes in ring conformations, or both. Table 4(*a*)–(*i*) presents a numerical summary of molecular conformational diversity classified according to values of N_{tor} in the range 1–9. Inspection of Table 4 shows that the proportion of molecules with $N_{env} \geq 2$ that adopt just one conformation decreases steadily as N_{tor} increases.

Trans and gauche conformers: The most common type of rotational energy profile about a single bond is that of staggered conformations defined by torsion angles of $\sim 180^\circ$ [*anti* or *trans* (*t*)] or $\pm 60^\circ$ [+*gauche* (+*g*), −*gauche* (−*g*)], commonly seen in the trimodal torsion angle plots generated by *e.g.* the *Mogul* program (Bruno *et al.*, 2004). This type of rotational variation is responsible for the 26 different conformations exhibited by the 193 instances of the tetra-*n*-butylammonium ion, the component with the greatest conformational variety in the complete $N_{tor} \geq 1$ subgroup. While the fully extended *tt* conformation is the most prevalent for individual *n*-butyl units, almost every possible variation involving *gt* and *gg* twists is also present. Conformational (*t*,*g*) variety in the C_{17} side-chains of steroids is well documented (Allen *et al.*, 1991*c*; Duax *et al.*, 1980) and is reflected here in the six discrete conformers adopted by the 64 observations of cholic acid. It is also no surprise to see that amino acids such as L-arginine, L-histidine and L-tryptophan also occur high in lists ordered by the number of conformational clusters, adopting 24, 9 and 5 conformers over 55, 24 and 7 occurences. A simpler example is provided by methoxycarbonylcholine, which adopts a fully extended (*tt*) conformation in its iodide (MCHOLI01), but a folded *tg* conformation in its picrate (GEBMIJ) as discussed by Frydenvang & Jensen (1996), who also discuss similar conformational variations in other choline esters. Similarly, L-ascorbic acid adopts the *tt* conformation in both independent molecules in LASCAC12, but a *gt* conformation in its co-crystals with 1-isoquinoline-carboxylic acid (ERAVAU) and L-serine (SERASC10), almost certainly so as to form the most effective hydrogen-bonded units in all three structures.

A small number of conformations of the eclipsed type were also observed, for example, about the bonds from each ring to the exocyclic C atoms in the highly strained crystal structure of the di-adamantyl structure DUSMOT, almost certainly due to heavy steric overcrowding in an already strained system. A closer examination of the results shows that eclipsed conformations are most frequently due to poorly determined solvent structures.

Two further examples, linking the discussions above and below, concern two crystal components that have a very high ratio of conformational variants to their numbers of occurrences in this subset. First, 1,6-bis(*o*-chlorophenyl)-1,6-diphenylhexa-2,4-diyne-1,6-diol (*e.g.* in GIKZAB) occurs in nine different conformations in just 11 different crystal environments: the parent molecule and ten clathrates. By comparison with the corresponding but-2-yne diol (NOXKAM), this molecule has an even longer linear spacer between the terminal Csp^3 centres, thus allowing it to react conformationally to the hydrogen-bonding requirements of a wide range of co-crystal (clathrate) partners. Secondly, the citrate ion (*e.g.* in CITEND01) occurs in 12 environments and also adopts nine different conformations, due entirely to rotational variance of the —COOH and $-COO^-$ groups, again providing flexibility in hydrogen-bond formation.

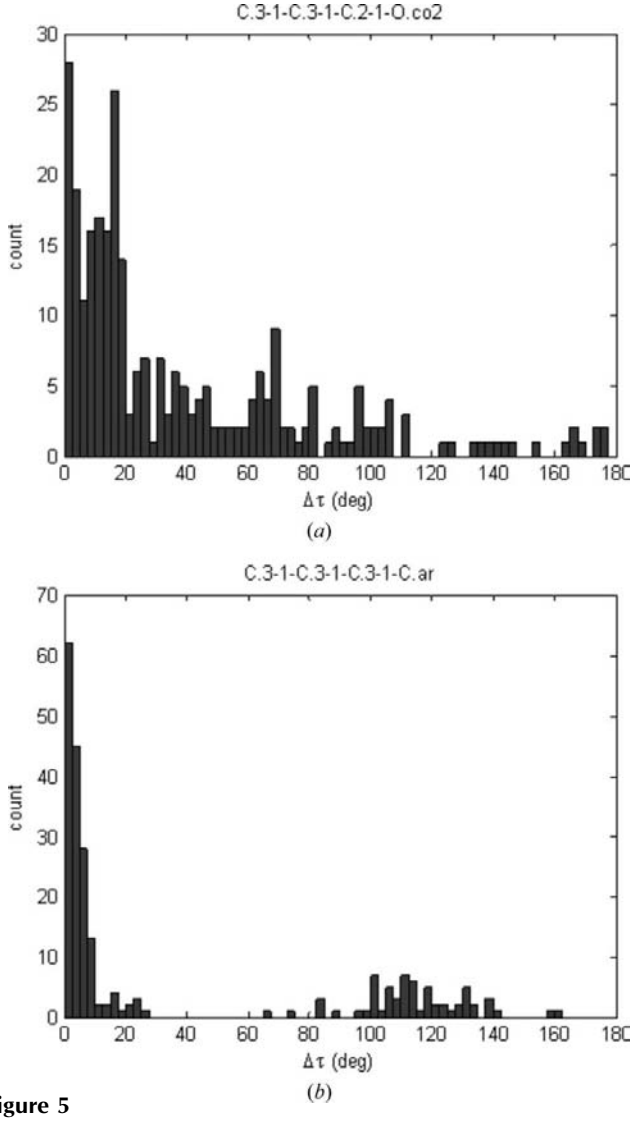


Figure 5
Histograms of selected torsion angles in different molecular conformations: (*a*) the angle defining the twist of a COOH in $Csp^3-Csp^3-C(=O)-OH$ fragments, (*b*) the angle defining the twist of a phenyl group in $Csp^3-Csp^3-Csp^3-C_{ar}$ fragments.

Rotations of planar functional groups: The presence of carboxylate rotamers is one of the most frequent causes of conformational diversity in the $N_{tor} \geq 1$ dataset. Thus, twists of close to 180° are observed, *e.g.* in comparing the three independent molecules of quinolinium-4-carboxylic acid in COBPAK and BEQXAW, and a very wide range of rotations is exhibited by naphthalene-1,4-dicarboxylic acid in OFUHUS and WUKROJ by internal comparisons and comparisons with the parent molecule in NAPDCX. The broad range of —COOH rotation is displayed in Fig. 5(*a*), which plots differences in $Csp^3-Csp^3-C(=O)-OH$ torsions between pairs of identical molecules in different crystal environments. This figure should be contrasted with, *e.g.* comparable data for $Csp^3-Csp^3-Csp^3-C_{ar}$ torsion angles presented in Fig. 5(*b*). Flipping of similar substituents is also observed in instances of the 3-amidopyridinium ion, where all examples have the amido group essentially coplanar with the pyridinium ring, but three examples (LACTEO, VUFPIV and IPOZAO) have the amido-O *cis* to the ring N, and three more (EMINUJ, TAFBUX, VAXLIP) have a *trans* arrangement. Complete 180° flipping of planar substituents is also observed for aldehyde groups, *e.g.* in the four independent molecules of pyrrole-2,5-dicarbaldehyde (DIZGOI), two of which have their —CHO groups *cis*,*cis* with respect to the ring-N, while the other two have a *cis*,*trans* arrangement. Meanwhile, rotations of nitro groups attached to phenyl rings are also common, for example in the isomolecular pair GATCAG and KIXZUM.

N_{tor} = *n subgroups*: For N_{tor} = 1, there are limited opportunities for conformational variability, and the most conformationally diverse molecule, ethylenediamine, has 19 exemplars forming four clusters, principally fully staggered about C—C, but with some *gauche* variants. There may also be problems related to the poor resolution of such a small molecule in the presence of much larger co-components. Otherwise, the maximum number of variants observed is 3, due to rotations of planar substituents (*e.g.* nitro) as noted above, or the adoption of staggered and (different) *gauche* conformers about single bonds. For N_{tor} = 2, the dihydrogen diphosphate ion has 16 occurrences spread over eight conformational clusters, due to a variety of rotations of the terminal $P(=O)(-OH)(O^-)$ groupings. The most highly diverse organic component in this class is malonic acid, with 16 occurrences in six clusters. This is due to the variety of rotamers that may be adopted by the terminal —COOH groups.

For N_{tor} > 2, the maximum number of observed conformational variants begins to rise quite significantly. Thus, for N_{tor} = 3, the most diverse molecule is the L-histidinium ion, with 24 occurrences spread over nine different clusters, and already noted above. There are 104 instances of hexane solvent, 96 exhibiting the minimum energy fully extended

Table 5
Distribution of conformations for organic molecules for specially constructed subsets of the complete dataset.

Column headings are defined in §3.1. All tables are truncated at $N_{env} \leq 10$

N_{env}	N_{obs}	N_{conf}					
		1	2	3	4	5	≥ 6
(*a*) Polymorphs (all data)							
1	0	0	0	0	0	0	0
2	834	583 (69.9%)	251 (30.1%)	0	0	0	0
3	287	175 (61.0%)	71 (24.7%)	41 (14.3%)	0	0	0
4	111	74 (66.7%)	19 (17.1%)	10 (9.0%)	8 (7.2%)	0	0
5	38	30 (78.9%)	6 (15.8%)	2 (5.3%)	0	0	0
6	23	18 (78.3%)	3 (13.0%)	2 (8.7%)	0	0	0
7	8	7 (87.5%)	1 (12.5%)	0	0	0	0
8	2	2 (100%)	0	0	0	0	0
9	3	3 (100%)	0	0	0	0	0
10	2	2 (100%)	0	0	0	0	0
(*b*) Polymorphs ($N_{tor} \geq 1$)							
1	0	0	0	0	0	0	0
2	579	347 (59.9%)	232 (40.1%)	0	0	0	0
3	213	107 (50.2%)	66 (31.0%)	40 (18.8%)	0	0	0
4	68	32 (47.1%)	19 (27.9%)	9 (13.2%)	8 (11.8%)	0	0
5	20	12 (60.0%)	6 (30.0%)	2 (10.0%)	0	0	0
6	8	4 (50.0%)	2 (25.0%)	2 (25.0%)	0	0	0
7	6	5 (83.3%)	1 (16.7%)	0	0	0	0
8	2	2 (100%)	0	0	0	0	0
9	2	2 (100%)	0	0	0	0	0
10	0	0	0	0	0	0	0
(*c*) Co-crystals (all data)							
1	1799	1799 (100%)	0	0	0	0	0
2	459	347 (75.6%)	112 (24.4%)	0	0	0	0
3	144	96 (66.7%)	32 (22.2%)	16 (11.1%)	0	0	0
4	102	64 (62.7%)	26 (25.5%)	9 (8.8%)	3 (2.9%)	0	0
5	63	43 (68.3%)	11 (17.5%)	7 (11.1%)	2 (3.2%)	0	0
6	47	33 (70.2%)	4 (8.5%)	4 (8.5%)	3 (6.4%)	1 (2.1%)	2 (4.3%)
7	33	21 (63.6%)	8 (24.2%)	4 (12.1%)	0	0	0
8	29	21 (72.4%)	2 (6.9%)	4 (13.8%)	1 (3.4%)	1 (3.4%)	0
9	19	11 (57.9%)	3 (15.8%)	3 (15.8%)	2 (10.5%)	0	0
10	13	6 (46.2%)	2 (15.4%)	1 (7.7%)	3 (23.1%)	0	0
(*d*) Co-crystals ($N_{tor} \geq 1$)							
1	1350	1350 (100%)	0	0	0	0	0
2	278	176 (63.3%)	102 (36.7%)	0	0	0	0
3	93	49 (52.7%)	29 (31.2%)	15 (16.1%)	0	0	0
4	66	29 (43.9%)	25 (37.9%)	9 (13.6%)	3 (4.5%)	0	0
5	38	19 (50.0%)	11 (28.9%)	6 (15.8%)	2 (5.3%)	0	0
6	30	18 (60.0%)	3 (10.0%)	4 (13.3%)	2 (6.7%)	1 (3.3%)	2 (6.7%)
7	21	10 (47.6%)	7 (33.3%)	4 (19.0%)	0	0	0
8	19	11 (57.9%)	2 (10.5%)	4 (21.1%)	1 (5.3%)	1 (5.3%)	0
9	13	5 (38.5%)	3 (23.1%)	3 (23.1%)	2 (15.4%)	0	0
10	9	2 (22.2%)	2 (22.2%)	1 (11.1%)	3 (33.3%)	0	0
(*e*) Hydrates and solvates (all data)							
1	6309	6309 (100%)	0	0	0	0	0
2	1116	644 (57.7%)	472 (42.3%)	0	0	0	0
3	113	55 (48.7%)	29 (25.7%)	29 (25.7%)	0	0	0
4	85	45 (52.9%)	25 (29.4%)	10 (11.8%)	5 (5.9%)	0	0
5	21	10 (47.6%)	5 (23.8%)	2 (9.5%)	4 (19.0%)	0	0
6	12	7 (58.3%)	1 (8.3%)	2 (16.7%)	1 (8.3%)	1 (8.3%)	0
7	11	9 (81.8%)	2 (18.2%)	0	0	0	0
8	6	6 (100%)	0	0	0	0	0
9	4	1 (25.0%)	2 (50.0%)	1 (25.0%)	0	0	0
10	3	0	0	2 (66.7%)	0	1 (33.3%)	0
(*f*) Hydrates and solvates ($N_{tor} \geq 1$)							
1	5416	5416 (100%)	0	0	0	0	0
2	943	501 (53.1%)	442 (46.9%)	0	0	0	0
3	90	40 (44.4%)	23 (25.6%)	27 (30.0%)	0	0	0
4	71	34 (47.9%)	25 (35.2%)	8 (11.3%)	4 (5.6%)	0	0
5	17	6 (35.3%)	5 (29.4%)	2 (11.8%)	4 (23.5%)	0	0
6	9	5 (55.6%)	1 (11.1%)	1 (11.1%)	1 (11.1%)	1 (11.1%)	0
7	8	6 (75.0%)	2 (25.0%)	0	0	0	0
8	3	3 (100%)	0	0	0	0	0
9	4	1 (25.0%)	2 (50.0%)	1 (25.0%)	0	0	0
10	3	0	0	2 (66.7%)	0	1 (33.3%)	0

conformation, and eight others with (different) *gauche* twists along the carbon chain. Following that are the eight conformers exhibited across the 17 examples of 3,5-dicarboxycyclohexane-1-carboxylate (*e.g.* in JEDPEO and JEDPUE), due to varied orientations of the —COO(H) groups which enable them to form the most effective hydrogen-bonding interactions with other component(s). For $N_{tor} = 4$, the most conformationally diverse molecule is the diyne (*e.g.* in GIZKAB) discussed above. This is followed, in complexity order, by the L-methioninium ion with 10 examples spread over nine clusters, and by a tetraethyloctahydroxy-calix(4)arene, having nine examples spread over eight clusters.

For $N_{tor} = 5$, the most conformationally diverse molecule is trimethoprim in its protonated form (*e.g.* in CABYIO), with 31 occurrences spread over 11 clusters. Surprisingly, very little of this variation is due to differing orientations of the O—CH_3 groups, but arises almost exclusively from significant differences in the pair of torsion angles about the two bonds linking the ring systems. A more detailed analysis of conformational variability in trimethoprim and its co-crystals is now being undertaken. The second most diverse molecule in this class is the citrate ion, noted above, while the third is the hexamethylene-diammonium ion. Here, 29 examples are spread over 7 clusters, with 21 ions showing the fully extended conformation, and eight others showing different *gauche* twists. For $N_{tor} = 6$, the greatest conformational diversity is exhibited by protonated L-arginine (55 examples, 25 clusters), and by the bis(triphenylphosphine)iminium ion $(Ph_3)P{=}N^+{=}P(Ph_3)$ with 67 examples spread over 19 conformational clusters, and arising from differing orientations of the terminal phenyl rings on the almost linear P=N=P system.

3.4. Selected subsets

Crystal engineering, with a specific focus on improving the development and delivery of active pharmaceutical ingredients in crystalline forms, is a major current research interest at the

CCDC (Chisholm *et al.*, 2006). This background therefore directs interest to three specific subsets of the complete dataset generated in this work:

(i) *Polymorphs*: Chemical compounds that exist in two or more crystalline forms (and which may, of course, contain > 1 chemical component).

(ii) *Co-crystals*: Crystals comprising at least two different uncharged chemical components which are not water or a solvent listed by Görbitz & Hersleth (2000), although such solvent molecules may also be present. Further, the two major chemical components may not be identical or enantiomorphs of each other.

(iii) *Hydrates and solvates*: Crystals containing a major component (*i.e.* not a solvent or water) together with at least one water or solvent molecule. Apart from the major component, all other components must be water or solvents which occur in the solvent list published by Görbitz & Hersleth (2000).

Conformational diversity data for these three subsets are collected in Table 5, with two groupings being presented in each case:

(i) for all molecules in the subset, and

(ii) for those molecules in the subset having $N_{tor} \geq 1$.

It is this latter grouping that is the most appropriate for assessing conformational diversity, and data for those components that exist in two or three crystal environments in each subset are compared in Table 6.

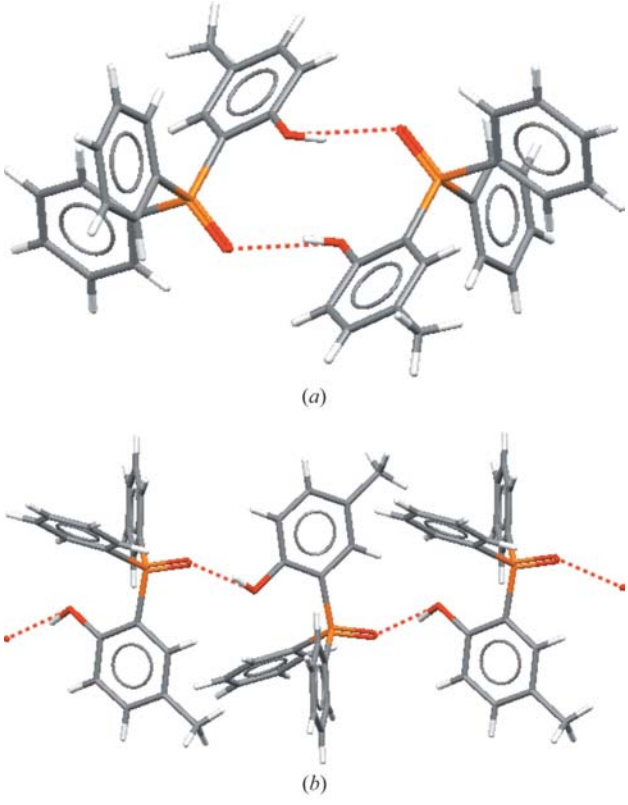


Figure 6
Two polymorphs of (2-hydroxy-5-methylphenyl)diphenylphosphine oxide: WOBWIT (top) shows hydrogen-bonded dimers while WOBWIT01 (bottom) shows hydrogen-bonded chains.

3.4.1. Polymorphs. Table 5(*b*) shows that polymorphs of compounds with $N_{tor} \geq 1$ which are found in just two environments actually exhibit two conformations for 40% of the sample. For those polymorphs observed in three environments, the number having $N_{tor} \geq 1$ and exhibiting two or three different conformations rises further to *ca* 50%. Table 6 shows that these data are very similar to data for the complete set of organic molecules having $N_{tor} \geq 1$. These figures of 40 and 50% are notably higher than the average for all molecules with $N_{env} = 2$, namely 32% (see §3.2). Beyond that, the numbers of compounds that exist as four or more polymorphic forms are too small to have statistical significance.

The polymorphic compounds that exhibit the highest number of conformational clusters (four) are all highly flexible acyclic molecules or derivatives of large rings, with some structures having $Z' > 1$. BUJBEN, VAZCOP and GOXWAR are typical examples and it is unsurprising that these molecules exhibit conformational diversity within their polymorphs. Conformational variation between polymorphs is often needed so as to form the alternative hydrogen-bonding motifs that differentiate the crystal packings. Thus (Fig. 6), WOBWIT exists as hydrogen-bonded dimers while its polymorph WOBWIT01 exhibits a chain motif.

3.4.2. Co-crystals. For co-crystals, 37% of compounds with $N_{env} = 2$ and $N_{tor} \geq 1$ exist in two conformers (Table 5*d*). When $N_{env} = 3$ the number existing in two or three conformers rises to 47%. These results are very similar to those shown by polymorphic compounds, as discussed above, and are also very similar to data for the complete set of organic molecules having $N_{tor} \geq 1$ (§3.2). The most conformationally diverse and charge-neutral co-crystal component is 18-crown-6, which exhibits nine different conformations across 148 co-crystal structures. Other components that have high numbers of different conformers in co-crystals include long-chain compounds such as decanedioic (sebacic) acid and tetraethyleneglycol dimethyl ether, together with rigid-core molecules such as benzene-1,2,4,5-tetracarboxylic acid.

3.4.3. Hydrates and solvates. For the set of solvated molecules that have $N_{env} = 2$, Table 5(*e*) shows that 42% adopt two conformations, and this rises to 47% for those molecules having $N_{tor} \geq 1$. For solvated molecules having $N_{env} = 3$ and $N_{tor} \geq 1$, 56% exhibit two or three different conformations. The most frequently solvated molecule in the CSD is cholic acid which exists in 29 different solvated forms, which between them exhibit six different conformations of the C_{17} side chain, in agreement with previous detailed conformational analyses (Allen *et al.*, 1991*c*; Duax *et al.*, 1980). The next most solvated compound is another cholane derivative (see *e.g.* GUXFUA), which has ten different solvates recorded in the CSD and exhibits five conformational variants of the C_{17} side-chain. Other molecules in the hydrates and solvates list that exhibit high conformational diversity are those with easily rotatable functional groups, *e.g.* —COOH, and a variety of peptides; the conformational variety of both of these classes of compounds has already been noted several times in this paper. The need

Table 6
Comparison of conformational diversity in structural subsets (i)–(iii) with each other and with the overall results for all molecules with $N_{tor} \geq 1$ for molecular components existing in $N_{env} = 2$ and $N_{env} = 3$ environments.

Column headings are defined in §3.1.

	$N_{env} = 2$			$N_{env} = 3$			
	N_{obs}	N_{conf}		N_{obs}	N_{conf}		
Subset		1	2		1	2	3
All molecules	10 856	6567 (60.5%)	4289 (39.7%)	1239	587 (47.4%)	388 (31.3%)	264 (21.3%)
(i) Polymorphs	579	347 (59.9%)	232 (40.1%)	213	107 (50.2%)	66 (31.0%)	40 (18.8%)
(ii) Co-crystals	278	176 (63.3%)	102 (36.7%)	93	59 (52.7%)	29 (31.2%)	15 (16.1%)
(iii) Hydrates and solvates	943	501 (53.1%)	442 (46.9%)	90	40 (44.4%)	23 (25.6%)	27 (30.0%)

for conformational change in hydrates and solvates is almost always due to the spatial requirements of different hydrogen-bonding schemes which dominate their extended crystal structures. However, conformational variability is not always necessary, since the same conformation of a solvated molecule can often form different hydrogen-bonded motifs with different solvates. This point is illustrated in Fig. 7 which shows a drug molecule in three solvated forms with different hydrogen-bond environments. In this case the hydrate (SULSUX) has one conformation, while the other two (HEZNEF, HEZNOP) have the same conformation, despite the formation of different hydrogen bonds and HEZNOP incorporating two independent and bulky solvent molecules into the crystal structure.

3.4.4. Comparison of polymorphs, co-crystals and solvates. Apart from the intrinsic interest of each of the three individual subsets, it was also of interest to see if there were any significant differences in conformational variability when the subsets were compared one with another and with the data for all molecules. Table 6 shows that the subsets of polymorphs (i) and co-crystals (ii) exhibit very similar conformational diversity patterns. Thus, when $N_{env} = 2$, *ca* 60% of components adopt a single conformation, and *ca* 40% adopt two different conformations; when $N_{env} = 3$, then *ca* 50% maintain a single conformation, *ca* 30% exhibit two different conformations and *ca* 20% exhibit three different conformations. These diversity patterns are strikingly similar to those exhibited by the 'all molecules' grouping. However, conformational diversity data for hydrates and solvates [subset (iii)] do appear to differ, in having nearly 47% of major molecular components exhibiting two conformations when $N_{env} = 2$, and 30% of components exhibiting three different conformations when $N_{env} = 3$. These figures are *ca* 7–10% higher than for subsets (i) and (ii), and for the 'all molecules' grouping.

It would appear that the major molecular components in hydrates and solvates (iii) are better able to adjust their conformations to accommodate small, and usually hydrogen-bond-forming, solvate or water molecules into the crystal lattice so as to stabilize their crystal packings, *i.e.* the net reduction in lattice energy offsets any small conformational energy increases in many cases. In the case of polymorphs (i) the components normally have only identical molecular components with which to pack, while in the case of co-crystals (ii) the second molecular component is normally much

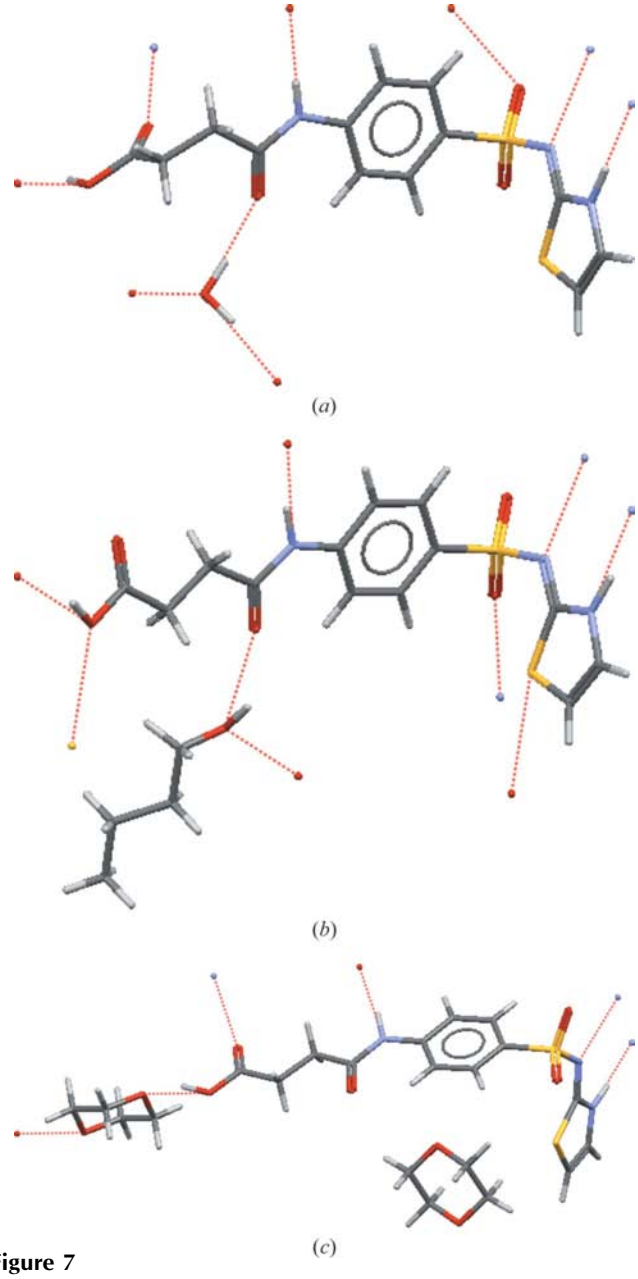


Figure 7
Examples of conformational differences in different solvates of 4-succinylamino-*N*-2-thiazolylbenzenesulfonamide: (top) in its monohydrated form (SULSUX), (middle) in its 1-butanol solvate (HEZNEF) and (bottom) in its 1,4-dioxane solvate (HEZNOP). Hydrogen bonds are shown as dotted lines.

larger than water or any of the common solvent molecules and component pairs are often of similar size. Hence subsets (i) and (ii) appear to exhibit conformational behaviour that is more similar to the norm exemplified by 'all molecules' than to the solvates and hydrates of subset (iii). However, Tables 5(*e*) and (*f*) show that 86% of all solvated primary components have inherent conformational flexibility, *i.e.* they have $N_{\rm tor} \geq 1$, but for polymorphs and co-crystals, this percentage is rather lower at 69 and 76%, respectively.

4. Conclusion

This paper has laid the foundations for a thorough conformational analysis of the organic component molecules in the CSD. The methodology for analysing the database for multiple occurrences of these molecular components and for determining their different conformations using cluster analysis has been described. The discussion presents a broad initial survey of the complete dataset, and of various subsets representing:

(i) various degrees of perceived conformational flexibility based on the number, $N_{\rm tor}$, of freely rotatable acyclic bonds, and

(ii) various structural groupings that are important in crystal engineering and pharmaceutical materials development.

These initial surveys have concentrated on those molecules that exhibit maximal conformational diversity, and it is observed that these molecules are, as expected, those which can adopt a wide variety of almost equi-energetic conformers, *e.g.* large ring compounds, small peptides and compounds having freely rotatable functional group substituents. There have been no surprises in the current analysis. Rather it has provided further reassurance that crystal conformations are indeed good guides to conformational flexibility *in vacuo* and in solution, in line with the previous computational and database study of Allen *et al.* (1996). Nevertheless, a number of the data listings generated for the subsets are worthy of considerable further analysis. This work is now being carried out and will be reported in due course. The methodology has also been run against the metal-organic section of the CSD, and results from this survey are being collated.

However, the current analysis has not attempted to answer one key question: are there instances of conformations for the same molecule in the CSD which differ significantly in energy? This requires that the conformational clusters obtained in this study are analysed further through energy calculations, and using a level of theory that is appropriate for such a large computation. Prototype calculations are currently being performed before undertaking full-scale processing which will aim to answer the key question posed above. Again, the outcomes of this work will be reported as soon as results become available.

ZFW thanks the Sims Empire Trust, the Cambridge Commonwealth Trust and the CCDC for financial support of a studentship. JMC acknowledges the Royal Society for a University Research Fellowship and St Catharine's College for a Senior Research Fellowship. Dr Greg Shields and Dr Jacco van de Streek are thanked for their software contributions to the determination of three-dimensional stereoisomerism.

References

Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.
Allen, F. H., Doyle, M. J. & Taylor, R. (1991*a*). *Acta Cryst.* B**47**, 29–40.
Allen, F. H., Doyle, M. J. & Taylor, R. (1991*b*). *Acta Cryst.* B**47**, 41–49.
Allen, F. H., Doyle, M. J. & Taylor, R. (1991*c*). *Acta Cryst.* B**47**, 50–61.
Allen, F. H., Harris, S. E. & Taylor, R. (1996). *J. Comput. Aided Mol. Des.* **10**, 247–254.
Allen, F. H., Howard, J. A. K. & Pitchford, N. A. (1993). *Acta Cryst.* B**49**, 910–928.
Allen, F. H. & Motherwell, W. D. S. (2002). *Acta Cryst.* B**58**, 407–422.
Bakaj, M. & Zimmer, M. (1999). *J. Mol. Struct.* **508**, 59–72.
Barnard, J. M. (1993). *J. Chem. Inf. Comput. Sci.* **33**, 532–538.
Braga, D., Grepioni, F., Walther, D., Heubach, K., Schmidt, A., Imhof, W., Gorls, H. & Klettke, T. (1997). *Organometallics*, **16**, 4910–4917.
Brock, C. P. & Minton, R. P. (1989). *J. Am. Chem. Soc.* **111**, 4586–4592.
Bruno, I. J., Cole, J. C., Kessler, M., Luo, J., Motherwell, W. D. S., Purkis, L. H., Smith, B. R., Taylor, R., Cooper, R. I., Harris, S. E. & Orpen, A. G. (2004). *J. Chem. Inf. Comput. Sci.* **44**, 2133–2144.
Bucourt, R. & Hainaut, D. (1965). *Bull. Soc. Chim. Fr.* pp. 1366–1378.
Bürgi, H.-B. & Dunitz, J. D. (1994). *Structure Correlation.* Weinheim: VCH Publishers.
Chisholm, J., Pidcock, E., van de Streek, J., Infantes, L., Motherwell, W. D. S. & Allen, F. H. (2006). *CrystEngComm*, **8**, 11–28.
Cooper, T. G., Jones, W., Motherwell, W. D. S. & Day, G. M. (2007). *CrystEngComm*, **9**, 595–602.
Dalhus, B. & Görbitz, C. H. (2000). *Acta Cryst.* B**56**, 715–719.
Duax, W. L., Griffin, J. F., Rohrer, D. C. & Weeks, C. M. (1980). *Lipids*, **15**, 783–792.
Everitt, B. L. M., Ash, A., Schober, J. & Leese, M. M. (2001). *Cluster Analysis.* London: Hodder Arnold.
Frydenvang, K. & Jensen, B. (1996). *Acta Cryst.* B**52**, 184–193.
Görbitz, C. H. & Hersleth, H.-P. (2000). *Acta Cryst.* B**56**, 526–534.
Harris, S. E., Pascual, I. & Orpen, A. G. (2001). *J. Chem. Soc. Dalton Trans.* pp. 2996–3009.
Kearsley, S. K. (1989). *Acta Cryst.* A**45**, 208–210.
Leach, A. R. (1994). *J. Chem. Inf. Comput. Sci.* **34**, 661–670.
Macrae, C. F., Edgington, P. R., McCabe, P., Pidcock, E., Shields, G. P., Taylor, R., Towler, M. & van de Streek, J. (2006). *J. Appl. Cryst.* **39**, 453–457.
Mo, Y. & Gao, J. (2007). *Acc. Chem. Res.* **40**, 113–119.
Norskov-Lauritsen, L. & Bürgi, H. B. (1985). *J. Comput. Chem.* **6**, 216–228.
Razinger, M., Balasubramanian, K., Perdih, M. & Munk, M. E. (1993). *J. Chem. Inf. Comput. Sci.* **33**, 812–815.
Shankland, N., Florence, A. J., Cox, P. J., Wilson, C. C. & Shankland, K. (1998). *Int. J. Pharm.* **165**, 107–116.
Shenkin, P. S. & McDonald, D. Q. (1994). *J. Comput. Chem.* **15**, 899–916.
Starbuck, J., Docherty, R., Charlton, M. H. & Buttar, D. (1999). *J. Chem. Soc. Perkin Trans. 2*, pp. 677–691.
Streek, J. van de (2006). *Acta Cryst.* B**62**, 567–579.
Weng, Z. F., Motherwell, W. D. S. & Cole, J. M. (2008). *J. Appl. Cryst.* Submitted for publication.
Zimmer, M. (2001). *Coord. Chem. Rev.* **212**, 133–163.