Acta Crystallographica Section B
Structural Science
ISSN 0108-7681

Searching the Cambridge Structural Database for the 'best' representative of each unique polymorph

Jacco van de Streek

Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England

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

Received 30 March 2006
Accepted 25 May 2006

A computer program has been written that removes suspicious crystal structures from the Cambridge Structural Database and clusters the remaining crystal structures as polymorphs or redeterminations. For every set of redeterminations, one crystal structure is selected to be the best representative of that polymorph. The results, 243 355 well determined crystal structures grouped by unique polymorph, are presented and analysed.

1. Introduction

The Cambridge Structural Database (CSD; Allen, 2002), maintained by the Cambridge Crystallographic Data Centre (CCDC), is a database containing virtually all organic and metal-organic crystal structures ever published. The CCDC has several policies regarding which crystal structures are incorporated in the CSD, and this paper deals with the consequences of two of those policies:

(i) The CSD aims for as complete as possible coverage of the literature.

(ii) The CSD tries to report objectively what is presented in the literature.

There is a simple reason behind the second policy: it is almost impossible to distinguish an outlier due to an error from an outlier that happens to be an interesting new discovery like a very short bond length, a very short contact or an unusual metal coordination. Therefore, in order to guarantee that the CCDC does not impose their idea of what is right or wrong but leaves this judgement up to the crystallographers and the peer review process, suspicious crystal structures are not usually actively corrected without the authors' consent. An exception to this rule is made, and the crystal structure corrected, if the crystal structure is clearly in error, for example if it contains a 0.66 Å non-bonded C$\cdots$C contact, *and* an obvious cause for the anomaly can be found, for example if changing the space group from $P3_1$ to $P3_2$ makes the short contact disappear.

It is not always appreciated that as a consequence of these two policies the CSD contains some crystal structures that most crystallographers would consider questionable or even wrong. This is especially true for crystal structures that were published before the introduction of the crystallographic information file (CIF; Hall *et al.*, 1991), many of which had to be re-keyboarded, first prior to being printed and then a second time when being entered into the CSD. When carrying out searches in the CSD, all such entries will be included in the hits, possibly leading to outliers in statistical analyses that need to be removed.

The large size of the CSD, more than 350 000 crystal structures, and the exponential growth of this number prohibit

manual inspection of data sets used in statistical surveys. Therefore, crystallographers have written computer programs to validate crystal structures. *UNIMOL* (Allen *et al.*, 1974) and its successor *PreQuest* (CCDC, 2005*a*) implement a variety of checks including searching for voids and consistency of bond lengths. Hooft *et al.* (1996) have done this for the Protein Data Bank. Spek (2003) has devised a series of rigorous checks for single-crystal structures of small organic and organometallic compounds, which are now part of *checkCIF*, available on the internet (http://www.iucr.org). When new checks were added to *PreQuest*, however, these were not applied retrospectively to check the many thousands of existing CSD entries. Most of the tests implemented by Hooft *et al.* are specific to proteins and cannot be used to check crystal structures in the CSD. Most of the checks implemented by Spek rely heavily on the raw data, namely the structure factors (not available in the CSD), but those that do not can readily be used to check crystal structures in the CSD.

Although the completeness of literature coverage may seem an obvious strong point of the CSD, there are some consequences of this policy that are nevertheless not always appreciated. The presence of multiple determinations, and even republications of the same crystal structure, is not always helpful. When performing searches in the CSD, all of these entries will be returned, leading to possible bias in statistical analyses. Another problem occurs if an earlier determination is superseded by a later one, for example if Marsh publishes a space-group correction for the original determination (Marsh & Spek, 2001; Marsh, 2002): both the original structure and the corrected one will be returned, and it is up to the user to select the latter. A computer program has been written by van de Streek & Motherwell (2005) to distinguish redeterminations from true polymorphs, but only rudimentary checks on the correctness of the crystal structures were carried out.

In the present paper we will combine some of the checks on crystal structure quality devised by Spek with new checks that are more specific to the CSD. The crystal structures that pass these checks are then clustered as redeterminations or as polymorphs using the method of van de Streek & Motherwell (2005) with several improvements. From the remaining clusters, the 'best' representative is selected based on one of four criteria. The final result is a set of lists of CSD identifiers (refcodes) of well determined unique crystal structures that can be used as filters in *ConQuest* (Bruno *et al.*, 2002) or *Mercury* (Macrae *et al.*, 2006).

2. Methodology

We need to define the words *republication* (*rerefinement*), *redetermination*, *reinterpretation* and *polymorph* as applied in this paper to pairs of crystal structures. A crystal structure is a *republication* of another structure if the exact same determination is published more than once, without any additional crystallographic experimental data. These can generally be recognized by noting that the unit-cell parameters are identical, and that the papers have one or more authors in common. Sometimes the more recent publication has a lower *R*-factor, for instance if H atoms were not included in the first publication but added in the second. In this case the second crystal structure is a *rerefinement* of the first. A *redetermination* is the publication of a known crystal structure from a different set of experimental data, usually by different authors. A *reinterpretation* is the publication of a non-trivial correction to a previously published structure, usually from the same data but possibly based on different or additional experimental data. A *polymorph* is the publication of the crystal structure of a chemical compound for which at least one other crystal structure is known that has a different packing. (On rare occasions the distinction between these cases is not clear cut; for example if two refinements with different algorithms were published in the same paper to demonstrate the influence of the method used, or if the same crystal was measured at the same temperature with X-ray and with neutron radiation, and the structure was refined twice but using the same unit cell in both refinements.) In the CSD, republications, reinterpretations and polymorphs are flagged as such by means of the keywords XREF, REINT-OF and POLYMORPH, respectively.

Every chemical compound in the CSD is assigned an identifier called a 'refcode', consisting of six alphabetical characters. Multiple crystal structures of the same chemical compound (for example polymorphs, but also including redeterminations and republications) are distinguished by adding two further digits. For paracetamol, for instance, the CSD contains 25 entries, labelled HXACAN, HXACAN01, HXACAN02 *etc.* The refcode family is unique to the chemical compound, and all polymorphs of a compound can therefore be found in the same refcode family (ignoring the small number of misassignments that may be present). Partially or fully deuterated forms of a molecule are considered to be the same compound as the hydrogenated molecule and can be found in the same refcode family; racemates and enantiopure compounds are thermodynamically different compounds and are assigned to different refcode families, as are all other forms of stereoisomerism.

Our program methodology consists of three stages:

(i) Eliminate crystal structures that are very suspicious or downright wrong.

(ii) Distinguish between redeterminations and polymorphs to cluster all redeterminations per unique polymorph.

(iii) Per cluster, select the 'best' representative.

The work extended over about 2 years, and three versions of the CSD were used: the November 2003, November 2004 and November 2005 releases. All results in §6 were obtained with the November 2005 release.

3. Stage 1: eliminating incorrect crystal structures

Several criteria were applied, some are common to *CheckCif* (Spek, 2003). It was decided not to penalize against problems caused by H atoms, because they are such weak X-ray scatterers. H atoms are sometimes absent altogether from the coordinate list, but even when present they are sometimes placed in calculated positions based on the authors' inter-

pretation of the chemistry behind the diffraction data; and this interpretation is not always correct. Also, we avoid penalizing against problems caused by bond types because these are not always unambiguous, and the presence or absence of bonds does not change the results from the diffraction experiment; and, because the CSD also contains older crystal structures for which no positions of H atoms were reported, we cannot use the number of connections to C, N, O *etc.* to infer their hybridization state. A third category that perhaps should not be penalized against are problems caused by the inclusion of solvent molecules, but it turned out not to be possible to easily separate the effect of the solvent from the rest of the structure, and in this work problems caused by a rogue solvent molecule will cause the entire crystal structure to be eliminated; work is in progress to allow the solvent to be ignored. The criteria that were used to determine if a crystal structure can be reliably processed by a computer program are now discussed below.

3.1. Space group/unit cell

Historically, some crystal structures were incorporated in the CSD for which only the unit-cell parameters were published, but no three-dimensional atomic coordinates. CSD entries without three-dimensional atomic coordinates sometimes have incomplete space groups, like *I****, and sometimes the unit-cell parameters are not present at all. In addition, it was checked that the unit-cell parameters agree with the crystal system of the space group, *e.g.* in $P2_1/c$, α and γ must be 90.0° (reassuringly, less than five errors of this type were encountered in the whole CSD).

3.2. Three-dimensional atomic coordinates present

This criterion insists that all non-H atoms have three-dimensional coordinates. H atoms, owing to their low X-ray scattering power, quite often used to be absent, and their presence is not required.

3.3. *R*-factor < 10%

Probably the most obvious quality measure is the *R*-factor. It was found that *R*-factors over 10% generally indicate that something is wrong with the crystal structure. It was also noticed that several of the false positives found when searching for polymorphs (van de Streek & Motherwell, 2005) were due to distortions in crystal structures with high *R* values, where 'high' usually meant higher than 10%. Therefore, removing crystal structures with an *R*-value of >10% reduces the number of false positives when trying to distinguish polymorphs from redeterminations.

Some CSD entries do not have an *R*-factor, usually because none was reported; this happens for example when Marsh publishes a reinterpretation (Marsh, 2002). In the program developed, such crystal structures without an *R*-factor pass this test.

The main disadvantage of using the *R*-factor as a quality measure is that information about the data-to-parameter ratio is lost, and this information is not available in the CSD. For completeness, we mention here the use of the e.s.d.s of the C—C bond lengths as an alternative quality measure for crystal structures (Allen *et al.*, 1995), but its merits were not investigated.

3.4. No disorder

Unfortunately, although CSD entries with disorder are flagged by the presence of a DISORDER field, the CSD does not yet store disorder information completely. The positions of disordered atoms are retained, but their occupancies are not. Moreover, the information on which pairs of atoms are related by disorder is lost. As a result, several of the manipulations and tests that we intend to perform would fail. Therefore, all CSD entries with disorder need to be eliminated. Of course, the presence of disorder does not in any way reflect a bad crystal structure determination, and this criterion unfortunately removes some in principle correct crystal structures. An exception is made if the disorder pertains only to H atoms, for example in a methyl group or because two H atoms are shared between two hydrogen-bonded carboxylate groups. No DISORDER keyword is added for disordered methyl groups, but in general the CSD does not distinguish between disorder in H atoms and disorder in non-H atoms; eliminating entries based on the presence of the DISORDER keyword would therefore eliminate both categories. A more involved approach was therefore needed.

If disorder is not symmetry-imposed, it is dealt with in one of two ways in the CSD: either the atoms with the lowest occupancies are discarded, or all atoms are entered but all but one of the alternatives must be entered as 'suppressed' atoms (marked by adding a question mark '?' to the atom label) that can be ignored. The first case does not need any special treatment, and CSD entries in which one of the occupancies was discarded will pass the test. The second case cannot be dealt with, and any CSD entry containing suppressed non-H atoms fails this test. If the disorder is symmetry-imposed, there are no suppressed atoms and this situation can therefore not be detected. However, the automatically generated symmetry-related atoms will cause many superfluous bonds that are very short and that do not exist in the two-dimensional connectivity. These should be caught by some of the other tests.

In many of these disordered crystal structures, the presence of disorder is due only to the presence of a disordered solvent molecule that serves as space-filler in an otherwise well behaved crystal structure.

3.5. No unmatched entries

Every CSD entry has a match flag, indicating whether it was possible to map every atom with three-dimensional coordinates to an atom in the structural formula (the two-dimensional connectivity). The main reason why this is not always possible is disorder, so this test is essentially an extension of the previous test.

Table 1
Breakdown of the 200 space-group changes that resolved intermolecular C···C contacts shorter than 2.7 Å.

Old space group	New space group	Frequency	Old space group	New space group	Frequency
$P2_1/n$	$P2_1/c$	37	$C2$	$I2$	1
$P2_1/c$	$P2_1/n$	35	$C2/c$	Cc	1
$P2_12_12_1$	$P2_12_12$	19	$C2/c$	$I2/c$	1
$P2_12_12_1$	$P2_12_12_1$†	18	Cc	Cn	1
$P2_1$/c	$P2_1/a$	9	Cc	Ia	1
$Pbca$	$Pcab$	6	$I2_1/a$	$I2/a$	1
$P2_1/c$	$P2/c$	5	$P2_1/b11$	$P112_1/b$	1
$P2_1/a$	$P112_1/a$	4	$P2_1/c$	$P112_1/n$	1
$P2_1/a$	$P2_1/n$	4	$P112_1/b$	$P112_1/n$	1
$P2_1/n$	$P2_1/a$	4	$P112_1/n$	$P112_1/a$	1
$P2_1$	$P2_1$†	3	$P112_1/n$	$P2_1/n$	1
$P2$	$P2_1$	3	$P2_1/a$	$P2/c$	1
$C2/c$	$I2/a$	2	$P2_12_12$	$P2_12_12_1$	1
$I2/a$	$C2/c$	2	$P22_12_1$	$P2_12_12_1$	1
$P2_1/n$	$P112_1/n$	2	$P3_1$	$P3_2$	1
$P2_1/a$	$P2_1/c$	2	$P3_2$	$P3_1$	1
$P2/c$	$P2_1/c$	2	$Pbc2_1$	$Pbc2_1$†	1
$P2/n$	$P2_1/n$	2	$Pc2_1n$	$Pc2_1n$†	1
$P2_1/a$	$P2/a$	2	$Pccn$	$Pccn$†	1
$P2_1/n$	$P2/n$	2	$P2_1nb$	$Pna2_1$	1
$P3_221$	$P3_121$	2	$F2dd$	$Fdd2$	1
$P4_1$	$P4_3$	2	Pn	Pn†	1
$P4_12_12$	$P4_32_12$	2	$R3$	$R3$‡	1
$P6_1$	$P6_5$	2	$P4_2/n$	$P4_2/n$†	1
$Pcab$	$Pbca$	2	$P2_1/m$	$P2_12_12$	1
$B2_1$	$B2_1$†	1	$P2_1$	$P2_1/a$	1
$A2/a$	$A2/n$	1	$P2_1$	$P2_1/n$	1

† Different choice of origin. ‡ Cell centring needed to be changed from hexagonal obverse to non-standard hexagonal reverse.

3.6. Chemical formula and structural formula consistent

The chemical formula that is stored with the entry should be consistent with the chemical formula that can be derived from the two-dimensional connectivity. H atoms are ignored.

3.7. Unit-cell volume

The volume of the unit cell can be calculated from the unit-cell parameters ($V_{\rm observed}$), but can also be estimated from the temperature-dependent average atomic volumes as published by Hofmann (2002) summed over all atoms in the unit cell and using the temperature at which the crystal structure was determined as recorded in the CSD ($V_{\rm estimated}$). In our first attempt, a histogram of the ratio $V_{\rm observed}/V_{\rm estimated}$ for all structures in the CSD was prepared. Based on this histogram, a structure should be rejected if the ratio $V_{\rm observed}/V_{\rm estimated}$ is greater than 1.4 or smaller than 0.7. Discrepancies can be caused by an unresolved solvent molecule, especially if it has been modelled using *SQUEEZE* (van der Sluis & Spek, 1990). An incorrect space group, for example $P1$ instead of $P\bar{1}$, is another likely cause. Closer inspection of the rejected structures suggested that these limits may be too strict for crystal structures containing metals. Metals can occur in more than one oxidation state, which can substantially change their contribution to the unit-cell volume, and their average atomic volumes therefore have a larger estimated standard deviation. The same is true on a smaller scale for the halogens, and this could be solved by making the average atomic volumes a function of oxidation state, but that was not attempted. In order to estimate the extent to which these oxidation states influence the results, separate histograms for organic and organometallic structures were prepared; these histograms clearly showed that volumes of organometallic compounds are predicted less accurately than those of organic compounds. Hofmann's paper lists the mean error Δv for the atomic volumes, and merging the two categories by switching the criterion from $V_{\rm observed}/V_{\rm estimated}$ to a criterion based on $(V_{\rm observed} - V_{\rm estimated})/(\Sigma\Delta v)$ was attempted, but these histograms still showed a significant difference between organic and organometallic structures. The only option remaining was to apply different limits depending on whether a crystal structure is organic or organometallic. The limits were determined from distributions of $V_{\rm observed}/V_{\rm estimated}$ accumulated using only those entries in the CSD that did not have disorder, were perfectly matched, and for which all three-dimensional atomic coordinates of all non-H atoms had been determined. These distributions are given in Fig. 1, and the limits derived from them are 0.80–1.40 for organics and 0.65–1.50 for organometallics. The limits are deliberately generous to avoid introducing too much bias.

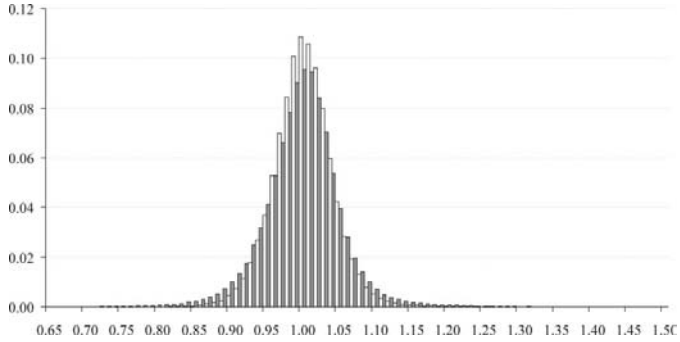

Figure 1
Distributions of the relative frequency of ratios of the observed and the estimated unit-cell volumes for organic (white) and organometallic (grey) compounds in the CSD. The histograms show that estimating unit-cell volumes is less accurate for organometallics than for organics.

3.8. No intermolecular C···C contacts < 2.7 Å

It can be surprisingly difficult to spot an incorrect space group. The asymmetric unit in such a case looks fine and, as long as the incorrect space group is consistent with the crystal system, the symmetry-generated molecules look fine too. Only the presence of short contacts indicates that the space group could be wrong. In order to validate this criterion, *ConQuest* was used to search the CSD for all non-bonded C···C contacts shorter than 2.7 Å. Even with a local installation of the CSD to avoid network traffic, this search took two days on a modern personal computer.

There are several reasons why a CSD entry can have a short C···C contact. In order of importance:

(i) *Disorder* of either the main molecule or a solvent.

(ii) *Catena compounds. ConQuest* does not take the 'wrap around' into account, and finds a short contact between two C atoms that should be bonded.

(iii) *An error in the space group.* For 200 CSD entries this author was able to find a space-group correction that made the short contacts disappear. Table 1 gives a breakdown of the types of error. These were all corrected in time for the November 2004 release of the CSD.

3.9. Geometry of homogeneous aromatic six-membered rings

The short C···C contact test mainly detects errors in the space group; it is not very reliable for detecting errors in the atomic coordinates of the asymmetric unit. Although it may not be possible to devise an algorithm that will detect every possible misprint in atomic coordinates, it was observed that it is generally easy to spot a suspicious phenyl ring or, more generally, a homogeneous aromatic six-membered ring. Two tests were investigated.

(i) *Deviation from the mean plane.* The deviation from the mean plane, *i.e.* a measure for planarity, turned out not to be a very good test for aromatic rings. The substituents on the ring can bend the ring quite considerably, *e.g.* for helices (Fig. 2). Tests showed that for phenyl rings, where at least five of the six substituents are H atoms, planarity is a good criterion, but obviously this criterion would not be applicable for most CSD entries. The planarity criterion also has the disadvantage that it does not detect aromatic rings that are 'sheared' due to two unit-cell axes having been swapped, as this does not affect the planarity of the ring (see Fig. 3). Therefore, it was decided not to incorporate this criterion.

(ii) *Distortion from sixfold symmetry*. The degree to which a six-membered aromatic ring corresponds to sixfold symmetry can be measured by calculating an orientational order parameter S_6 of the form

$$S_m = \left| \frac{1}{N} \sum_{j=1}^{N} \exp\left(-im\alpha_j\right) \right|^2, \quad m = 6, \tag{1}$$

for a set of angles $\{\alpha_j\}$, $j = 1 \ldots N$. The exact definition of the angles $\{\alpha_j\}$ and further mathematical details are given in Appendix *A*. This order parameter gives a value between 0.0 and 1.0; 1.0 meaning exact sixfold symmetry. A threshold value of 0.95 was tried, but visual inspection of the results showed that quite a few of the crystal structures that were eliminated were still fairly reasonable. A better threshold value is 0.90, with which 1711 CSD entries were eliminated. 500 of these were examined manually, and all 500 looked suspicious, indicating that sixfold symmetry is a very sensitive measure of the quality of a six-membered aromatic ring. For 25 of these 500 entries the geometries of aromatic rings that were clearly 'sheared' could be corrected either by swapping two unit-cell parameters or by changing the unit-cell angle β to $180^\circ - \beta$. An example is given in Fig. 3 for CSD refcode AKIKAG (Dubberley *et al.*, 2003), a recently published crystal structure with an *R*-value of 6.32%. All six aromatic rings in this crystal structure are visibly distorted, with an order parameter of just 0.346 for the aromatic ring shown in Fig. 3. All the distortions can be resolved by swapping the unit-cell parameters *a* and *c*, after which the order parameter for the ring in Fig. 3 becomes 0.999.

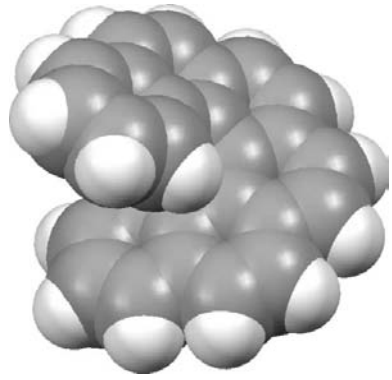

Figure 2
A helix consisting of homogeneous aromatic six-membered rings. The rings, though aromatic, are clearly non-planar. Molecule taken from CSD entry ABUNAM (Ermer & Neudorfl, 2001).

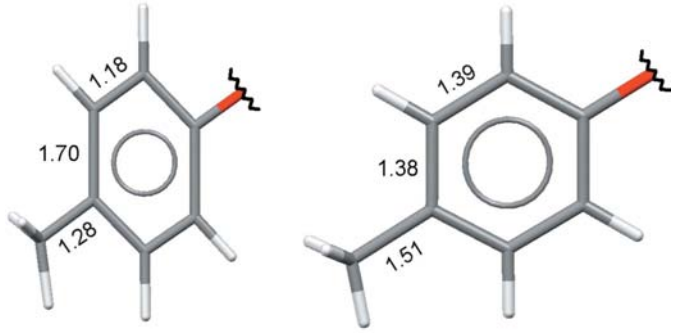

Figure 3
One of the aromatic rings in CSD refcode AKIKAG before (left) and after (right) swapping the unit-cell parameters *a* and *c*. In both figures the rings are in the plane of the paper. The numbers are bond lengths in Å. The black wavy line indicates where the rest of the molecule is attached.

Although the sixfold symmetry is also affected by substituents, especially by small-ring fusion (Allen, 1981), their impact is small and the order parameter for such systems does not fall below 0.970.

An obvious third test would involve the distribution of the aromatic C—C bond lengths or the distribution of the distances from the C atoms to their centroid. This test was not given further attention as the test based on distortion from sixfold symmetry appeared to be satisfactory.

Again, quite a few of the crystal structures are eliminated only because of the presence of an ill-determined solvent molecule such as benzene.

3.10. C—C bond lengths

Not every molecule has a homogeneous aromatic six-membered ring, but almost all of the entries in the CSD have a covalent C—C bond, where, to avoid penalizing against misassigned bond types, the C—C bond can be any bond type: single, double, triple, aromatic or delocalized. Allen *et al.* (1995) published a comprehensive survey of the quality of crystal structures based on the e.s.d.s in their C—C bond lengths, which is too detailed for our purposes; their survey also relied on the data itself being correct, whereas we are trying to detect errors in the data itself. In order to establish reasonable upper and lower limits, a histogram of all covalent

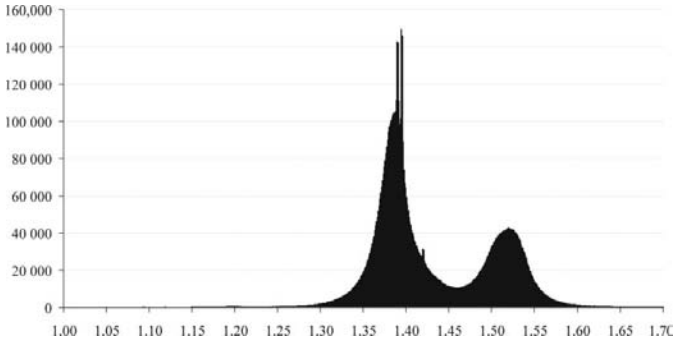

Figure 4
Distribution of C—C bond lengths in the CSD (Å). Note the barely visible peak around 1.2 Å corresponding to C≡C bonds. The upper and lower limits used as rejection criteria are 1.0 and 1.7 Å, respectively. See text for explanation of the three spikes.

C—C bond lengths in the CSD was generated (Fig. 4). *ConQuest*, *Vista* (CCDC, 2005*b*) and *MS-Excel* can no longer cope with this many data points ($>8 \times 10^6$) and a special-purpose program needed to be written. A bin width of 0.001 Å was chosen to allow visualization of fine detail. Based on the histogram, the limits chosen were 1.0 and 1.7 Å. These limits are deliberately fairly generous, so as to reduce the risk of eliminating possible interesting short or long C—C bonds as can be present in strained ring systems. The histogram shows a minor spike at 1.42 Å: this is presumably due to C—O single bonds in which the O atom was erroneously identified as a C atom. The spikes at 1.390 and 1.395 Å correspond to the use of constrained phenyl rings in the refinement.

3.11. C—O, C—N and N—N bond lengths

The procedure for bond lengths for other than C—C bonds is identical to the procedure for C—C bonds. The histograms are given in Figs. 5, 6 and 7. There is a minor spike at 1.390 Å in the C—N distribution that suggests that perhaps some aromatic C—C bonds were misinterpreted as aromatic C—N bonds. The upper and lower limits derived from these histograms are again 1.0 and 1.7 Å for all bonds.

3.12. Reinterpretations

Corrections to published structures, such as the space-group reinterpretations published by Marsh and others (Marsh & Spek, 2001; Marsh, 2002), are added to the CSD with the keyword REINT-OF and the refcode of the original CSD entry. The original entry is not corrected or removed, but a REINT-SEE keyword is added with a cross reference to the corrected crystal structure. Clearly, only the reinterpretation needs to be retained, the original entry is eliminated from the list.

3.13. Hard-coded list of exclusions

In order to allow the greatest possible flexibility, one of the tests is a comparison with a hard-coded list of refcodes that should be excluded. This allows reducing the number of false negatives without increasing the number of false positives.

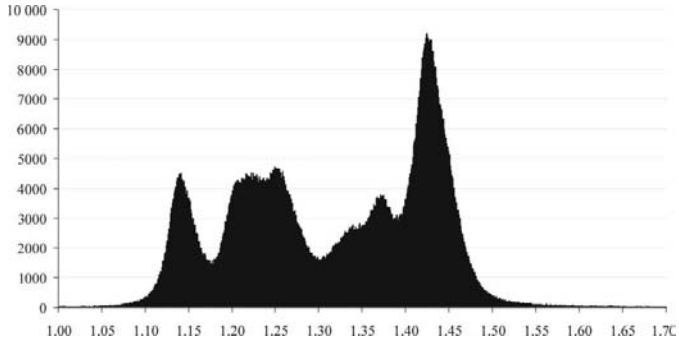

Figure 5
Distribution of C—O bond lengths in the CSD (Å). The upper and lower limits used as rejection criteria are 1.0 and 1.7 Å, respectively.

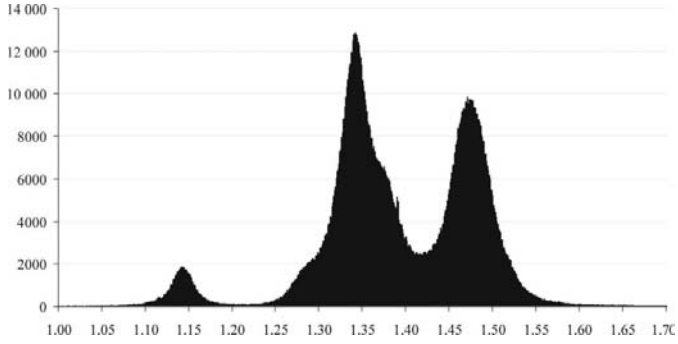

Figure 6
Distribution of C—N bond lengths in the CSD (Å). The upper and lower limits used as rejection criteria are 1.0 and 1.7 Å, respectively.

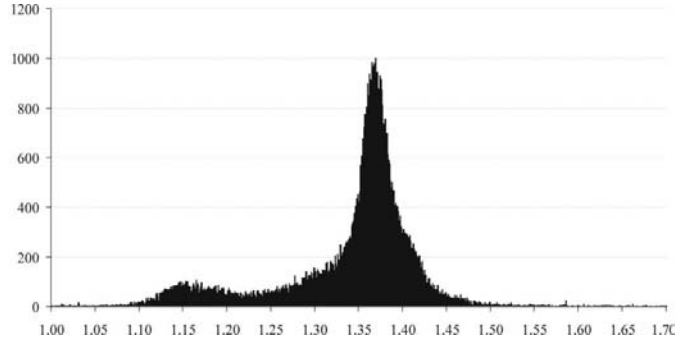

Figure 7
Distribution of N—N bond lengths in the CSD (Å). The upper and lower limits used as rejection criteria are 1.0 and 1.7 Å, respectively.

Once an entry is corrected, it should no longer be excluded by this list. Therefore, this list not only records the refcodes but also the dates they were added to the list: if the entry has been modified after it was added to the list, a warning is written out that prompts for another manual inspection to check if the entry should still be eliminated or not. Currently, this list is empty.

4. Stage 2: redeterminations *versus* polymorphs

As the CSD is intended to cover the literature exhaustively, redeterminations of crystal structures are incorporated

retaining all previous determinations. For a compound like paracetamol, for instance, this means that there are 25 CSD entries. However, not all of these are redeterminations: paracetamol has two polymorphs. Therefore, a method is needed to distinguish the polymorphs from the redeterminations. A short description will be given here; the details and validation of the method were described previously (van de Streek & Motherwell, 2005). For the current work, several improvements were introduced.

Problems with differences in choice of origin or space-group setting and missed symmetry or pseudo-symmetry can be conveniently solved by transforming the crystal structures to a one-dimensional function depending only on interatomic distances: a simulated powder diffraction pattern for example. The similarity of these simulated powder diffraction patterns is then calculated using a normalized weighted cross-correlation function (de Gelder *et al.*, 2001). This cross-correlation function is less sensitive to small peak shifts than ordinary point-by-point measures.

The expected unit-cell volume as calculated from Hofmann's average atomic volumes as used in one of the tests described above can also be used to normalize all unit-cell volumes to their room-temperature values. This volume normalization renders the unit cells of crystal structures determined at different temperatures more similar.

Simulated powder diffraction patterns are very sensitive to minor misprints in atomic coordinates: a missing minus sign in one of the three coordinates of a Cl atom is enough to completely alter a crystal structure's simulated powder diffraction pattern. This would cause the crystal structure to be perceived as 'different' with respect to redeterminations of the same polymorph, and the crystal structure would be classified, incorrectly, as a polymorph. Unit-cell parameters are more reliable in this respect, and we therefore also calculate the similarity of the stable reduced unit-cell *via* a 'reduced unit-cell diffraction pattern'. As it is the unit-cell parameters that determine the peak positions whereas it is the unit-cell contents (the atoms) that determine the intensities, the contribution of the atoms to the powder pattern can be removed by setting the magnitudes of all structure factors to a constant value. By also including those reflections that were systematically absent due to space-group symmetry, the contribution of the space group can be eliminated as well leaving a 'unit-cell diffraction pattern'. Omitting reflections generated by unit-cell centrings (if present) creates the diffraction pattern of the reduced unit cell, which is stable because of the projection onto a single axis.

The simulated powder pattern similarities described so far were treated in detail in the previous paper (van de Streek & Motherwell, 2005). It was not mentioned in the previous paper that manual inspection of unflagged pairs of polymorphs revealed that several of them turned out not to be polymorphs but either diastereomers or a racemate *versus* an enantiomerically pure compound. These are chemically different compounds and should therefore have been in different refcode families. Spotting these can be surprisingly difficult, especially if there are multiple chiral centres or if a racemate crystallizes in a space group without an improper symmetry element but with two molecules, enantiomers, in the asymmetric unit. Therefore, a special-purpose computer program was written to analyse all members of all refcode families for stereochemical consistency. This program uncovered 70 errors in the CSD, *i.e.* cases where stereochemically different compounds had been assigned to the same refcode family, all of which were corrected in time for the November 2004 release.

In the previous work we only used either the simulated powder diffraction pattern of the full structure, or the simulated pattern of the reduced unit cell. Both have their limitations, and in this work we included both, keeping the one that gave the greater similarity. This ensures that cases where missed symmetry led to a unit cell that is a multiple of the true cell are now properly identified as being the same crystal structures.

Incorrect element counts, usually the number of H atoms, interfere with the volume normalization and can cause the volume normalization to distort a crystal structure. Therefore, the second improvement introduced in this work is that both the similarity values with and without volume normalization (both of the full structure and of the reduced unit cell) were calculated and the greater was used.

Third, many of the cases where pairs of crystal structures were incorrectly identified as polymorphs turned out to be due to large differences in the temperatures of data collection. The temperature difference causes a difference in unit-cell volumes, *i.e.* in unit-cell parameters, which in turn causes peak shifts in the simulated powder diffraction patterns. As the thermal expansion of molecular crystals is generally anisotropic, and because it is unknown if the unit-cell angles for a given crystal structure increase or decrease as a function of temperature, the volume normalization mentioned above can only partially compensate for this effect. For substantial temperature differences, especially over 100 K, the volume normalization and use of a cross-correlation function are no longer sufficient to compensate for these peaks shifts, and structures that are visually similar have a low calculated similarity. Therefore, it was decided to bias the similarity measure to take this effect into account. At first glance it might seem that this can simply be done by lowering the threshold value above which two crystal structures are considered to be similar. However, this has the disadvantage that pairs of true polymorphs are also biased when they happened to have been determined at very different temperatures. Hence, it was not the similarity threshold that was made temperature-difference dependent, but the triangle width l that determines the tolerance to peak shifts. The formula $l = 2.0 + |\Delta T|/100.0$ K (where the unit of l is $^\circ$ 2θ) was initially tried, but for temperature differences greater than 100 K this formula yielded triangle widths that were so big ($l > 3.0^\circ$ 2θ) that crystal structures that are clearly different were given high similarities. It was therefore necessary to give the triangle width an upper limit of $l = 3.0^\circ$ 2θ. Pressure has an effect comparable with that of temperature, and an additional

term for pressure was also inserted. The triangle width l was therefore calculated as follows,

$$l = 2.0 + \min(1.0,\ |\Delta T|/100.0\,\mathrm{K} + |\Delta P|/10.0\,\mathrm{GPa}). \quad (2)$$

For every set of N refcodes in the same family that passed all of the tests in stage 1, an $N \times N$ similarity matrix was created. Based on previous work (van de Streek & Motherwell, 2005), a similarity value greater than 0.990 was assumed to indicate that the two crystal structures are the same, and a similarity smaller than 0.970 was assumed to indicate that the two crystal structures are polymorphs. Values between 0.970 and 0.990 were left as unknown, and where possible were assigned based on the remainder of the similarity matrix, by using the following two relationships: if $A = B$ and $B = C$ then $A = C$, and if $A = B$ and $A \neq C$ then $B \neq C$.

Any unknowns left after that stage were assumed to be different. This is not necessarily correct; the two crystal structures could represent two determinations of the same crystal structure at two very different temperatures, or one of the two crystal structures could have an undetected error in it; but the two crystal structures are apparently at least different enough so as not to bias surveys of the CSD any more.

The next step was checking if the similarity matrix was self-consistent. This is not always the case, especially if a temperature series has been measured and the crystal structures at two consecutive temperatures are still similar enough for the similarity to be detected, but the temperature difference between the two extremes of the range is too big for the two crystal structures to still be matched as being the same crystal structure. Therefore, any inconsistencies of the type $A = B$, $B = C$ but $A \neq C$ were resolved by setting $A = C$.

This clustering by means of a similarity matrix is the fourth improvement over the previous paper (van de Streek & Motherwell, 2005) introduced here. The previous paper only contained lists of pairs of polymorphs, including duplicates caused by redeterminations/republications, whereas the present paper contains a list of sets of refcodes clustered per unique polymorph. For paracetamol, for example, for which 25 determinations were published corresponding to only two unique polymorphs, the previous paper reported 56 pairs of polymorphs (because these included duplicates caused by the redeterminations). In an ideal world, these 56 pairs would be reduced to two clusters of refcodes: one cluster for each unique polymorph.

5. Stage 3: choosing the 'best' representative

At this point we have obtained a list of the most reliable crystal structures in the CSD and their breakdown into clusters of unique polymorphs. The next question most users of the CSD will ask is: given a cluster of redeterminations, which structure is 'the best'? At first glance this may seem like an impossible question to answer, because 'best' clearly depends on the context. Users interested in hydrogen bonds will prefer structures in which the H atoms have been located, especially *via* neutron studies. Users interested in comparing densities will prefer pairs of structures that have been determined at the same temperature, which almost invariably will mean that the structures were determined at room temperature. And some users will only accept a crystal structure as 'the best' if it has the lowest R-value (or alternatively e.s.d.s on C—C bond lengths), or if it was determined at the lowest temperature (so as to reduce the effects of thermal motion). In the author's experience, the criteria mentioned above are in practice by far the most common requirements. Therefore, although a single criterion will not satisfy all users' needs, and although the number of criteria is in principle infinite, the following four criteria probably cover the majority of user requirements:

(i) Lowest R-factor.

(ii) All H atoms present, and if everything else equal, neutron study.

(iii) Lowest temperature.

(iv) Room temperature.

It is possible that after applying one of the four criteria above we are still left with more than one crystal structure of a polymorph in a cluster. In that case it needs to be decided which criterion to use next and so on. Another issue that requires a little more thought is that R values and temperatures are continuous variables, and in order to establish equality a tolerance is needed. For example, two crystal structures with R-factors of 7.6% and 8.0% are really not that different, and if the 8.0% structure happened to be the one with all H atoms determined, that would probably be the one to keep. The following lists the tolerances used for comparisons and the order in which the criteria are applied.

List 1: *lowest R factor*
  R-factor, tolerance 1.0
  All H atoms determined
  Lowest temperature

List 2: *all H atoms determined*
  All H atoms determined
  Neutron radiation study
  R-factor

List 3: *lowest temperature*
  Lowest temperature, tolerance 20 K
  R-factor, tolerance 1.0
  All H atoms determined

List 4: *room temperature*
  Room temperature; in the CSD, this has a standard tolerance of 20 K
  R-factor, tolerance 1.0
  All H atoms determined

No unique polymorphs are eliminated at this stage: if no CSD entry with H atoms exists for list 2, the remaining criteria are applied to all entries. If no CSD entry determined at room temperature exists for list 4, then the structure that was determined at the temperature closest to room temperature is chosen, with a tolerance of 20 K. If for the same polymorph a CSD entry with and a CSD entry without an R-factor are available, the one with an R value (which is necessarily $<10\%$ at this stage) is chosen unless the entry without an R-factor was a reinterpretation, in which case it is always preferred over any entries with or without an R-factor. Using any of the lists as a subset to restrict searches to in *ConQuest* allows the

Table 2
Breakdown of the number of rejected CSD entries per test.

If a structure failed multiple tests, it is included multiple times in the table, but only added once to the total. Percentages are with respect to the total number of entries in the CSD.

Criterion	Rejected	Percentage
Unmatched entries	57 522	16.2%
No three-dimensional coordinates	56 171	15.8%
Disorder	48 085	13.5%
Volume discrepancy	37 950	10.7%
R-factor > 10%	15 076	4.2%
C—C bond lengths	8236	2.3%
Chemical formula	5374	1.5%
Space group/unit cell	3722	1.0%
Short C···C contacts	2037	0.6%
Aromatic six-membered rings	1853	0.5%
C—O bond lengths	1641	0.5%
REINT-SEE	1293	0.4%
C—N bond lengths	1052	0.3%
N—N bond lengths	400	0.1%
Total rejected	111 860	31.5%

application of additional filtering criteria using the full search capabilities of *ConQuest*.

If after applying all criteria still more than one structure remains, the entry with the most recent year of publication is chosen, based on the simple assumption that diffractometers and structure solution software improve over time.

6. Results and discussion

6.1. Stage 1: eliminating incorrect crystal structures

From this point, all results are as obtained with the November 2005 version of the CSD. The number of CSD entries rejected per test are given in Table 2. The result is a list of 243 204 refcodes of CSD entries that passed all tests (see supplementary material[1]). After failing a test, the remaining tests are still applied to that entry and entries can therefore fail more than one test. Several tests are correlated: a crystal structure without unit-cell parameters cannot possibly have (meaningful) fractional atomic coordinates, for example. For some of the tests, failing the test does not necessarily imply that the CSD entry is of low quality: *e.g.* the presence of disorder depends on the compound and not on the quality of the crystal structure determination. Conversely, it is not guaranteed that a crystal structure that passed all tests is error-free: the limits used in this work were deliberately fairly generous to ensure that a maximum of diversity was maintained. Hooft *et al.*, for example, in their work on protein structures (Hooft *et al.*, 1996), mention on one of the web pages referred to *via* their paper (http://swift.cmbi.ru.nl/gv/pdbreport/checkhelp/intro.html) that for C—C bonds a bond length of 1.66 Å is off from the average of 1.53 Å by 6.4σ, whereas in their work they used 4σ ($\sim$1.6 Å) as the upper limit. However, some highly strained molecules contain a legitimate C—C bond length between 1.6 and 1.7 Å, and we wanted to avoid removing these interesting cases. Therefore, in this work 1.7 Å was used as the upper limit, which is obviously much more generous. The presence of any remaining errors is not necessarily that harmful: the real outliers (values of 0.0 Å, for example) have been removed, giving the remaining errors a much better chance of averaging out.

There are some obvious tests that have not yet been included, such as a test for short C···O contacts, or any test involving valence angles. These may well be added in the future. The main reason why they have not yet been added is the time it takes to validate the results, which usually involves manually looking at several hundred CSD entries to check for false positives. Anisotropic displacement parameters (ADPs) are not yet stored in the CSD, and can therefore not be used to detect misassigned element types.

It seems wasteful to reject one third of the CSD, and one might ask if it is possible to correct those rejections that pertain to errors (as opposed to *e.g.* disorder) instead of rejecting them; indeed, the reader may have counted 300 corrections to CSD entries in this paper already. However, the difference between *detecting* an error and *resolving* an error should be noted: it is far easier to spot that a C atom is in a suspicious position than to determine why it is where it is, let alone to understand where it should have been. This often requires going back to the original publication, and ideally even to the original raw diffraction data, which is a time-consuming process that will result in a satisfactory correction for only a small fraction of the cases. Work is in progress to enable the software to ignore problems caused by solvents or by disorder.

Fig. 8 shows the number of reliable determinations per compound (according to our tests). This figure is presented here only for comparison with Fig. 12 in the next section.

Fig. 9 shows the breakdown of the number of rejected CSD entries by year of publication as a percentage of the total number of entries in the CSD that were published that year. To put this in perspective, Fig. 10 shows the total number of

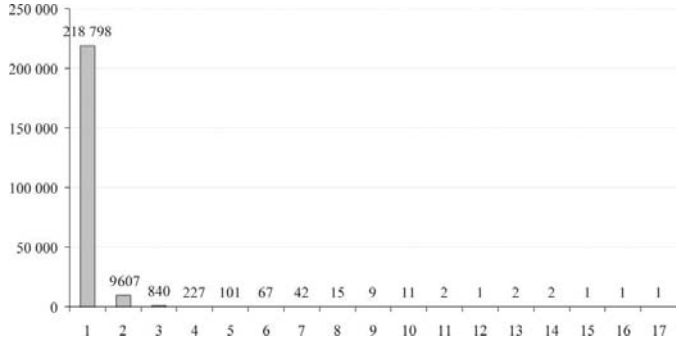

Figure 8
The frequencies of the number of determinations per compound, summed over all its polymorphs, that passed all the tests in stage 1. Five outliers, two compounds with 21 determinations and three with 22, 25 and 27 determinations, have been omitted for clarity. The raw data, including individual refcodes, can be found in the supplementary material[1].

[1] Supplementary data for this paper are available from the IUCr electronic archives (Reference: RY5004). Services for accessing these data are described at the back of the journal.

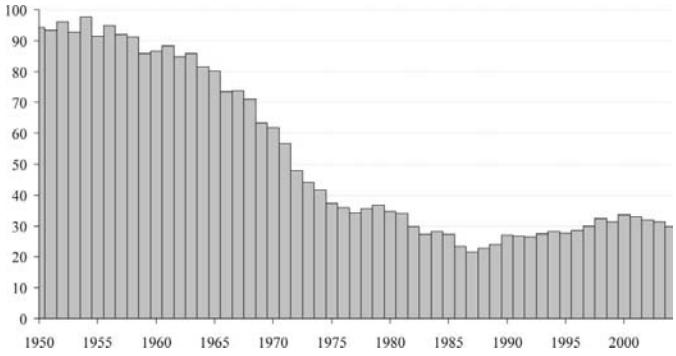

Figure 9
Percentage of rejected entries per year of publication. Percentages are with respect to the total number of CSD entries for that year of publication.

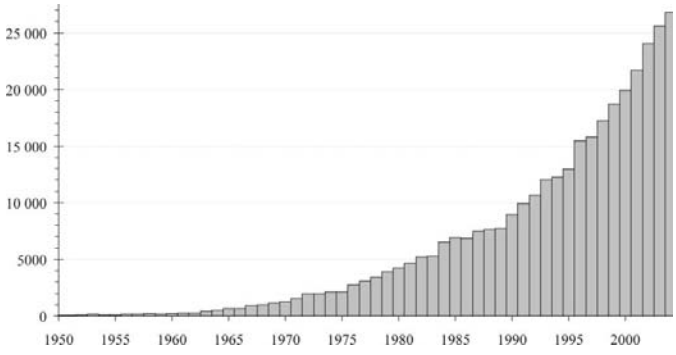

Figure 10
Number of entries in the CSD by year of publication.

entries in the CSD by year of publication. It can be seen that between 1965 and 1975 the quality of crystal structures improved considerably, and that since about 1975 approximately 70% of all published crystal structures passed all tests. Allen & Taylor (2005) report a figure of 10% for misprints in typeset crystallographic data (a significant source of error until the mid-1980s, when electronic depositions started to become more common), and Marsh (1997) reports the same figure for crystal structures erroneously described in space group *Cc* (of course, our tests cannot spot missed symmetry, it is used here only as an indication of the number of human errors per crystal structure). Combined with the fact that approximately 20% of the structures in the CSD are flagged as having disorder suggests that as a rough guide, of the 30% that fails the tests, approximately 10% is caused by human error in the published crystal structure, the remaining 20% being mainly due to problems caused by disorder. It is the author's estimate that half of the disordered structures will no longer cause a problem, *i.e.* pass all the tests, when disorder is stored properly in the CSD.

With solvents causing so many problems (short contacts, strange bond lengths and distorted phenyl rings), an estimate of the number of solvated structures is needed. Görbitz & Hersleth (2000) published statistics on the number of solvent inclusions in organic and organometallic crystal structures, which allows us to estimate how many of the problems are possibly due to solvent molecules. For 168 112 unique refcodes (only one member of each refcode family was included) they found 15 848 hydrates, or 9.4%, and 21 148 solvates (excluding hydrates), or 12.6%. Combined, this amounts to 22.0% solvates (including hydrates). With 20% of the CSD entries exhibiting disorder, we know that the number for the solvates must be greater (van der Sluis & Kroon, 1989), but we also know that it is not 100%. This suggests that between 20 and 100% of the solvated structures are disordered. If this number is chosen to be halfway, it is estimated that 60% of the solvated structures are disordered; this number agrees very well with the number of 56% found by Nangia & Desiraju (1999). Then, as a crude approximation, 60% of 22% of the structures in the CSD have a problem caused by a disordered solvent molecule. Taking into account that the relative number of solvates increases over time and that the data of Görbitz & Hersleth are 6 years old, this suggests that approximately 15% of the CSD entries are eliminated due to problems caused by the solvent, which is half of all the rejected structures.

It goes without saying that the elimination process has created a biased subset of the CSD, and it is probably impossible to quantify the effect of this bias, positive or negative, on statistical surveys of small-molecule crystal structures. Endless numbers of histograms showing the differences in distributions in the CSD and the subset generated in this paper could be prepared: for the distribution of space groups, molecular weight, number of atoms, elements, stereochemistry, polymorphs *etc.* We would like to confine ourselves to only one: the number of flexible torsion angles (Fig. 11).

A flexible torsion angle is defined as any torsion angle involving four contiguous atoms where the bond between the two central atoms is a single bond that is not part of a ring system and where the two central atoms are bonded to each other and to at least one other non-H atom (so as to avoid counting a methyl group as flexible). This number is calculated for every bonded unit in the asymmetric unit (which may be a molecule, a counter-ion or a solvent), and the greatest number per CSD entry is added to the histogram. CSD entries containing a polymer or a *catena* compound were not included. The distributions in Fig. 11 clearly show that crystal structures of compounds with many flexible torsion angles are

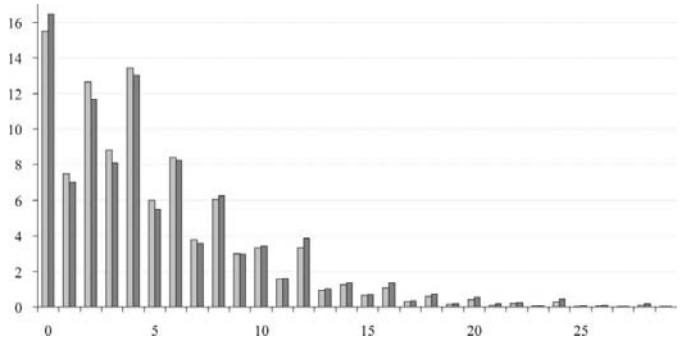

Figure 11
Comparison of the relative frequency of the number of flexible torsion angles in the CSD (dark grey) and after removing suspicious crystal structures (light grey).

systematically under-represented in the new subset. (There appear to be more compounds with zero flexible torsion angles in the CSD than in the filtered set, but this includes those cases where the automatic calculation of the number of flexible torsion angles failed, *e.g.* because no two-dimensional connectivity was available.) This is presumably because greater molecular flexibility increases the probability that the crystal structure is disordered.

Table 3
Results of the clustering step for some well studied polymorphic systems.

Compound	Refcode family	Entries in CSD	Surviving entries	Polymorphs found	Polymorphs correct†
Paracetamol	HXACAN	25	21	3	2
Glycine	GLYCIN	36	25	3	3
Benzene	BENZEN	11	6	3	2
Oxalic acid dihydrate	OXACDH	30	27	2	2
Carbamazepine	CBMZPN	7	6‡	4	4
Sulfathiazole	SUTHAZ	7	7	5	5
Piracetam	BISMEV	6	5	4	4

† The correct number of polymorphs is the number of polymorphs that the clustering algorithm should have found among the surviving entries; it is not necessarily the correct number of polymorphs as present in the CSD, due to the eliminations in stage 1. ‡ Although CBMZPN03 passed all the tests, the empty channels in the structure look suspicious.

6.2. Stage 2: clustering per unique polymorph

The result is a file that contains the refcodes of each cluster (see supplementary material); there turn out to be 231 782 unique clusters. Fig. 12 shows the number of polymorphs (which may be 1, in which case the term 'polymorph' is not technically correct) per compound that passed all the tests in stage 1. In their talks at the 35th International School of Crystallography in Erice, Italy, in 2004, Professor M. U. Schmidt and Professor S. R. Byrn showed unpublished graphs of the distributions of the number of polymorphs in well studied systems (organic pigments and pharmaceuticals, respectively). Both graphs suggested a Poisson distribution with $\lambda = 2$ polymorphs per compound (due to the Poisson distribution being defined as starting at 0 rather than at 1, the interpretation of 'polymorphs' in this context is 'number of *additional* different crystal packings discovered after the first crystal structure has been found', but exact numbers need not concern us here). Comparing these distributions of the number of polymorphs in well studied systems with Fig. 12 very clearly reminds us that crystallography is traditionally used to investigate molecular geometries, in which the packing of the molecules, let alone the existence of more than one packing, was merely a by-product. The following quote from a recent electronic teaching pamphlet (http://journals.iucr.org/iucr-top/comm/cteach/pamphlets/21/index.html) by Gavezzotti & Flack (2005) is appropriate here:

> 'The age of intramolecular structural chemistry is declining for small molecules. There is very little that can be added to the average intramolecular geometrical data collected by use of the Cambridge Structural Database; anything at variance with these well established averages is most probably wrong. Long experience has shown that discussing electronic effects in terms of molecular geometry alone is a tricky business. So, if you are an X-ray diffractionist, instead of looking at your molecule, try looking at your crystal. There is plenty to be discovered, at a low cost and with perfectly high confidence, by looking at what molecules do when they interact with each other, and single-crystal X-ray diffraction is still the best technique for this purpose.'

Moreover, by appreciating how interesting molecular packing can be, X-ray diffractionists will hopefully be encouraged to start actively looking for whether alternative crystal packings of the same compound exist, hopefully leading to the discovery of more polymorphs.

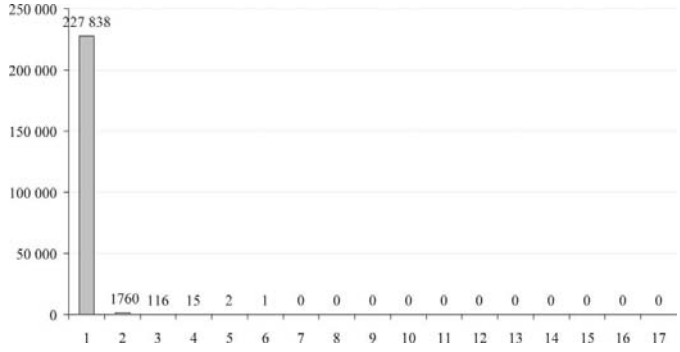

Figure 12
The number of different (as detected by our program) crystal packings per compound (*i.e.* 'polymorphs' if that number is greater than 1) that passed all the tests in stage 1. The scale is the same as that of Fig. 8. The individual refcodes can be found in the supplementary material.

The clusters of 83 refcode families with multiple determinations were manually examined in detail. The results for some well studied polymorphic compounds are given in Table 3. The false positive (two redeterminations being classified as polymorphs) for paracetamol is due to a high-pressure study at 4.0 GPa. The false positive for benzene turned out to be the result of a misprint in the unit-cell parameter *c* in the original publication of BENZEN04 (Fourme *et al.*, 1971) that had been faithfully copied into the CSD; this error has now been corrected in time for the November 2006 release of the CSD. The table illustrates that false positives are much more common than false negatives (two polymorphs being classified as redeterminations). For the 83 refcode families, the combined number of determinations is 386, for which our program finds 268 unique polymorphs. The correct number is 250, and the difference is built up of 19 false positives and one false negative. Apart from the 19 spurious clusters and one incorrect merge, none of the crystal structures had been assigned to the wrong cluster, *i.e.* there was no cross-contamination between clusters.

The author believes that the elimination of unreliable crystal structures in stage 1 is necessary for the clustering in stage 2 to be reliable. An incorrect crystal structure differs from a correct one, and if not eliminated, these incorrect structures will therefore show up as 'polymorphs'. This leads

to a catch-22 situation if one wants to generate a list of *all* best representatives of unique polymorphs in the CSD, not just those that passed all the tests in stage 1: if unreliable crystal structures are eliminated first, clustering those that remain is reliable, but we will have lost some polymorphs in the process. If, on the other hand, unreliable structures are not eliminated first, no polymorphs are lost, but the incorrect structures will not fit in with any of the existing clusters (because they are different from the correct structures) and will therefore be classified as separate 'polymorphs'.

Some CSD users may be interested not only in polymorphs of the pure compound but also in solvates (including hydrates) of that compound, and even polymorphs of those solvates. In the current work, no attempt was made to include these, but future work in this direction is underway.

6.3. Stage 3: choosing the 'best' representative

The result of stage 3 is four lists of 231 918 refcodes each (see supplementary material) that can be used in *ConQuest* or *Mercury*.

It is important to note that (*a*) all the crystal structures that are competing to be the best representative already passed all our tests, and so can be expected to meet at least a minimal set of quality criteria, and (*b*) all the crystal structures that are competing to be the best representative are already compared as being very similar, and for statistical purposes which one is chosen to represent them should not be too critical.

7. Conclusions

A computer program is described that filters the contents of the CSD based on quality tests, groups the crystal structures that pass into sets of unique polymorphs and chooses the 'best' representative based on one of four criteria to arrive at four lists of unique polymorphs of overall acceptable quality. 70% of the crystal structures in the CSD pass all quality tests. The author estimates that roughly 10% of the crystal structures are rejected due to the presence of an avoidable human error and that the remaining 20% are caused by disorder. Half of the disordered structures, *i.e.* 10%, can probably be made to pass all the tests if the complete disorder information was available in the CSD. Half of the crystal structures that are rejected, *i.e.* roughly 15%, are due to a problem with a solvent molecule. These figures are crude estimates and not very accurate. The main shortcoming of the current procedure is that the influence of solvent molecules cannot be separated out from the main molecules, and work is underway to address this. The next big improvement would be for the CSD to store disorder in full.

The main three problems with published small-molecule crystal structures encountered in this work are: (i) problems caused by solvents, (ii) problems caused by disorder, partially due to inadequate storage in the current CSD, and (iii) the tradition in small-molecule crystallography to search for 'the' crystal structure of a compound sufficient to characterize the molecular geometry, rather than for all possible polymorphs, causing polymorphism to be under-represented.

The computer program described in this paper can be run on a yearly basis and updated lists of refcodes provided to CSD users.

APPENDIX *A*
Orientational order parameter of a homogeneous six-membered aromatic ring

The average plane through the C atoms is calculated and the positions of the C atoms are projected onto the plane. The centroid of the C atoms (which lies on the plane) is chosen as the origin. An arbitrary axis in the plane is chosen (as will be shown, the final result is independent of the choice of axis), and the angles $\{\alpha_j\}$, $j = 1 \ldots 6$ are calculated as the angles between the arbitrary axis and the position vectors of the projected C atoms with respect to the centroid. The distribution of these $\{\alpha_j\}$ values is then expanded in the set of functions $\{\exp(im\alpha)\}$, $i = (-1)^{1/2}$, $m = 0, 1, \ldots \infty$, and only the expansion coefficient for $m = 6$, corresponding to sixfold symmetry, is kept. The square of the norm of that expansion coefficient is then our order parameter S_6,

$$S_m = \left| \frac{1}{N} \sum_{j=1}^{N} \exp\left(-im\alpha_j\right) \right|^2, \quad m = 6, \tag{3}$$

where N is the number of α_js; in this case 6. Equation (3) can be rewritten as

$$\begin{aligned} S_m &= \left| \frac{1}{N} \sum_{j=1}^{N} \exp(-im\alpha_j) \right|^2 \\ &= \frac{1}{N^2} \left\{ \left[\sum_{j=1}^{N} \cos\left(m\alpha_j\right) \right]^2 + \left[\sum_{k=1}^{N} \sin(m\alpha_k) \right]^2 \right\} \\ &= \frac{1}{N^2} \sum_{j=1}^{N} \sum_{k=1}^{N} \left[\cos\left(m\alpha_j\right) \cos(m\alpha_k) + \sin\left(m\alpha_j\right) \sin(m\alpha_k) \right] \\ &= \frac{1}{N^2} \sum_{j=1}^{N} \sum_{k=1}^{N} \cos\left[m\left(\alpha_j - \alpha_k\right) \right]. \end{aligned} \tag{4}$$

In other words, the final result depends only on the differences between two α_js and is therefore independent of the choice of axis. However, note that (3) has complexity O(N), whereas the final expression in (4) has complexity O(N^2).

Drs Sam Motherwell and Frank Allen are gratefully acknowledged for their careful reading of the manuscript and their helpful suggestions.

References

Allen, F. H. (1981). *Acta Cryst.* B**37**, 900–906.
Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.
Allen, F. H., Cole, J. C. & Howard, J. A. K. (1995). *Acta Cryst.* A**51**, 95–111.
Allen, F. H., Kennard, O., Motherwell, W. D. S., Town, W. G., Watson, D. G., Scott, T. J. & Larson, A. C. (1974). *J. Appl. Cryst.* **7**, 73–78.
Allen, F. H. & Taylor, R. (2005). *Chem. Commun.* pp. 5135–5140.

Bruno, I. J., Cole, J. C., Edgington, P. R., Kessler, M., Macrae, C. F., McCabe, P., Pearson, J. & Taylor, R. (2002). *Acta Cryst.* B**58**, 389–397.
CCDC (2005*a*). *PreQuest: A program for the validation of crystal structure and chemical information for entry to the CSD*. Version 5.27. CCDC, 12 Union Road, Cambridge, UK.
CCDC (2005*b*). *VISTA: A program for the display and analysis of geometrical and numerical information retrieved from the CSD*. Version 2.1b. CCDC, 12 Union Road, Cambridge, UK.
Dubberley, S. R., Friedrich, A., Willman, D. A., Mountford, P. & Radius, U. (2003). *Chem. Eur. J.* **9**, 3634–3654.
Ermer, O. & Neudorfl, J. (2001). *Helv. Chim. Acta*, **84**, 1268–1313.
Fourme, R., André, D. & Renaud, M. (1971). *Acta Cryst.* B**27**, 1275–1276.
Gavezzotti, A. & Flack, H. (2005). *Crystal Packing*, http://journals.iucr.org/iucr-top/comm/cteach/pamphlets/21/index.html.
Gelder, R. de, Wehrens, R. & Hageman, J. A. (2001). *J. Comput. Chem.* **22**, 273–289.
Görbitz, C. H. & Hersleth, H.-P. (2000). *Acta Cryst.* B**56**, 526–534.
Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *Acta Cryst.* A**47**, 655–685.
Hofmann, D. W. M. (2002). *Acta Cryst.* B**58**, 489–493.
Hooft, R. W. W., Vriend, G., Sander, C. & Abola, E. E. (1996). *Nature (London)*, **381**, 272.
Macrae, C. F., Edgington, P. R., McCabe, P., Pidcock, E., Shields, G., Taylor, R., Towler, M. & van de Streek, J. (2006). *J. Appl. Cryst.* **39**, 453–457.
Marsh, R. E. (1997). *Acta Cryst.* B**53**, 317–322.
Marsh, R. E. (2002). *Acta Cryst.* B**58**, 893–899.
Marsh, R. E. & Spek, A. L. (2001). *Acta Cryst.* B**57**, 800–805.
Nangia, A. & Desiraju, G. R. (1999). *Chem. Commun.* pp. 605–606.
Sluis, P. van der & Kroon, J. (1989). *J. Cryst. Growth*, **97**, 645–656.
Sluis, P. van der & Spek, A. L. (1990). *Acta Cryst.* A**46**, 194–201.
Spek, A. L. (2003). *J. Appl. Cryst.* **36**, 7–13.
Streek, J. van de & Motherwell, S. (2005). *Acta Cryst.* B**61**, 504–510.