Acta Crystallographica Section B
Structural Science
ISSN 0108-7681

CSDSymmetry: the definitive database of point-group and space-group symmetry relationships in small-molecule crystal structures

Jing Wen Yao,[a] Jason C. Cole,[a] Elna Pidcock,[a] Frank H. Allen,[a] Judith A. K. Howard[b] and W. D. Samuel Motherwell[a]*

[a]Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK, and [b]Department of Chemistry, University of Durham, South Road, Durham DH1 3LE, UK

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

Received 25 February 2002
Accepted 15 April 2002

An algorithm that perceives molecular symmetry has been applied to *ca.* 200 000 entries from the Cambridge Structural Database (CSD). For each molecule, the perceived point group, together with crystallographic properties such as space group, occupied Wyckoff positions and number of residues in the asymmetric unit, have been placed in a relational database, *CSDSymmetry*, using Microsoft *Access* software. Database queries can be constructed easily to find occurrences of any combination of molecular or crystallographic attributes, and thereby to answer questions on relative distributions. Some typical example queries are given. The inclusion of CSD reference codes enables direct visualization of search results using the Cambridge Crystallographic Data Centre's three-dimensional structure visualizer, *Mercury*.

1. Introduction

The question of why molecules pack to form particular crystal structures is of interest to many fields in science. For example, in crystal structure prediction and crystal engineering, establishing reliable correlations between molecular symmetry and crystallographic symmetry could prove invaluable (see *e.g.* Filippini & Gavezzotti, 1992), while in the design and synthesis of materials having non-linear optical properties, non-centrosymmetric space groups are required, and space-group control at the molecular level is highly desirable.

One approach to understanding the relationships between molecular and crystallographic symmetry is to gather statistics from existing crystal structures. Several published studies have considered molecular symmetry properties along with space-group frequency distributions (see, *e.g.* Scaringe, 1991; Zorky *et al.*, 1993; Wilson, 1993; Brock & Dunitz, 1994; Belsky *et al.*, 1995; Steiner, 2000 and references therein). In some of these studies, the distribution of space groups was examined with respect to the number of molecules per asymmetric unit (Z'), a property that is loosely correlated with molecular symmetry. Others have provided classifications based on the occupied Wyckoff positions across a wide range of space groups, defined as *structure classes* by Belsky *et al.* (1995). These publications have led to the formulation of some general rules of crystal packing (see Brock & Dunitz, 1994), such as 'mirror planes in mirror symmetric space groups are always occupied'. Other studies (Brock & Duncan, 1994; Lloyd & Brock, 1997), have focused on a set of molecules with a common structural motif and have examined their distribution through space groups and the occupancy of special positions in these space groups.

The Cambridge Structural Database (CSD) (Allen, 2002; Bruno *et al.*, 2002) has played an integral role in many of the studies cited. The CSD is the definitive database of published organic and metal-organic structures determined by X-ray and neutron diffraction techniques. The information contained within the database is primarily concerned with recording molecular structure as three-dimensional atomic coordinates with respect to a crystallographic unit cell and the space-group setting. However, it is recognized that molecular symmetry information is inherently present in the database and could be extracted; hence statistical analysis of molecular and crystallographic symmetry relationships could be performed using the entire database. A recent publication (Cole *et al.*, 2001) described a molecular-symmetry perception algorithm and its implementation in *RPluto* (Motherwell *et al.*, 1999).

The present paper describes the application of this algorithm to approximately 200 000 molecules retrieved from the CSD. The resulting data, together with crystallographic properties such as space group, symmetry of occupied Wyckoff positions (special or general positions) and Z', have been collated and entered into a relational database. Interrogation of this new database is possible with user-defined queries, and hence the database provides an extremely flexible source of symmetry-related information. Since the results of a query can be retrieved as a list of CSD reference codes (Refcodes), the corresponding molecules can be viewed easily in the Cambridge Crystallographic Data Centre's (CCDC's) three-dimensional structure visualizer, *Mercury* (Taylor & Macrae, 2001; Bruno *et al.*, 2002).[1]

2. Methodology

2.1. Choice of dataset

The April 2000 release of the CSD, version 5.19, containing 215 403 structures, was used to create a dataset from which to extract molecular and crystallographic information. The *Quest3D* program (Cambridge Crystallographic Data Centre, 1994) was used to select the dataset. Only those structures having three-dimensional coordinates were considered, and structures were further excluded if they (*a*) contained disordered molecules or ions,[2] (*b*) were polymeric (*catena* structures) or (*c*) contained connectivity matching errors, *i.e.* a 1:1 mapping of the chemical and crystallographic connectivities had been unsuccessful. Only one member of each Refcode family (*i.e.* structures having the same six-letter code) was retained to avoid bias in statistical analysis of the symmetry database. Crystal structures that represent different conformations of the same molecule are present within the CSD. If these crystal structures have the same six-letter Refcode, only one example will be retained within the symmetry database. A file of the surviving structures was created in the CCDC's FDAT format, which was then used for the extraction of symmetry and crystallographic information.

2.2. Molecular symmetry perception

A method for detecting approximate molecular symmetry in crystal structures has been developed previously (Cole *et al.*, 2001). In brief, the symmetry elements for each molecule are identified within certain geometrical tolerances, and the combination of the symmetry elements present for the molecule allows its assignment to a point group. The perception of molecular symmetry is performed using only the atomic coordinates, and no information regarding any coincident crystallographic symmetry is included in the analysis.

The molecular perception algorithm was implemented in *RPluto* through the *msym* command. *RPluto*, when provided with a list of retrieved CSD entries in FDAT format, can perform molecular symmetry identification as a batch process; the molecular point group, the Refcode, the number of residues, the space group (Hermann–Mauguin symbol), Z, Z' and the number of matrices used to detect the molecular symmetry are returned for each entry. The FDAT files of the retrieved dataset were submitted to the symmetry detection process and the point groups of 198 335 molecules were assigned. For this analysis of molecular symmetry in the CSD, the default geometric tolerances used on distance and angle (Cole *et al.*, 2001) were 0.1 Å and 5.0°, respectively. Molecules that contain fewer than three non-H atoms were not treated by the algorithm. The CSD bond-type code (a code that enumerates types of bond, *i.e.* single, double) was ignored and H atoms were disregarded for the purposes of symmetry detection, the implications of which are discussed in the next section.

2.3. Chemical considerations

Firstly it should be reiterated that structures tagged with the CSD disorder flag have not been included in the processed dataset. However, it is known that there are structures contained in the CSD that have been refined without proper resolution of disorder. Such structures often have highly variable geometry in regions of unresolved disorder. These structures are included in the processed dataset and may result in anomalous symmetry assignments.

As was mentioned above, the symmetry perception algorithm makes use of geometric tolerances for distance and angle. If the tolerances are set too tightly then very little symmetry will be perceived within a molecule. Conversely, if the tolerances are very large then symmetry will be determined to be present where it is not. The tolerances used in the analysis of this dataset were found to represent a reasonable compromise between 'missed symmetry' and 'false symmetry'.

In the current implementation of the symmetry perception algorithm, the CSD bond-type code is not considered. In some cases, this will lead to a 'false' point-group assignment. For example, 8-methoxynaphthalene-1-carboxylic acid [CSD entry MXNACX (Schweizer *et al.*, 1978)] is identified as having a mirror plane even though the carboxyl group is oriented perpendicular to the plane of the aromatic rings: the C—O bond is not distinguished from the C═O bond and the H atom is ignored. This type of assignment will occur with groups of the type B—A═B (where A may equal B) (Fig. 1). If there is a

[1] *RPluto* and *Mercury* can be downloaded for non-commercial research purposes from http://www.ccdc.cam.ac.uk/prods.

[2] Disorder is indicated as being present within the CSD if there is disorder of *any* atom sites, including solvent molecules.

symmetry element that passes through the central atom A (for example, a C_2 axis or a mirror plane), in the absence of bond-type information, the geometric tolerance of 0.1 Å is such that $A{=}B$ is not distinguishable from $A{-}B$.

To illustrate this point further, molecules (or fragments of molecules) found in the CSD that contain a $B{-}A{=}B$ group are shown in Fig. 1 along with the perceived symmetry. It can be seen that the decision to disregard bond-type information is justified in the cases of the nitrate ion and, in some instances, the carboxylate ion. However, in the examples illustrated by the molecules GOTQEL (Antorrena *et al.*, 1999) and DIVRIJ (Watson *et al.*, 1986) (Fig. 1), exclusion of bond-type information leads to false symmetry assignments. To establish how frequently the presence of a $B{-}A{=}B$ group perturbed the point-group assignment, a sample of 200 molecules was randomly selected from the CSD and the molecular point group determined by visual inspection was compared with the point group detected by the *msym* command implemented in *RPluto*. Only five examples were found where the formal point-group assignment was false.

Disregarding H atoms can lead to errors in molecular point-group assignment. For example, aqua-hydroxy-(μ_2-3,5-bis{bis[2-(diethylamino)ethyl]aminomethyl}pyrazole)-di-nickel(II)bis(tetraphenylborate) [Meyer *et al.* (1999), CSD entry HOCKEP], a binuclear nickel complex, is determined to have $C_2(2)$ symmetry when H atoms are ignored, thus identifying the two Ni atoms as being symmetry-equivalent. However, this symmetry is broken when H atoms are taken into consideration, as one nickel is coordinated to a water molecule and the other nickel is bound to a hydroxide. From the subset of 200 structures randomly selected from the CSD (see above), only three molecules were identified as belonging to a point group of lower symmetry once the H atoms were taken into consideration.

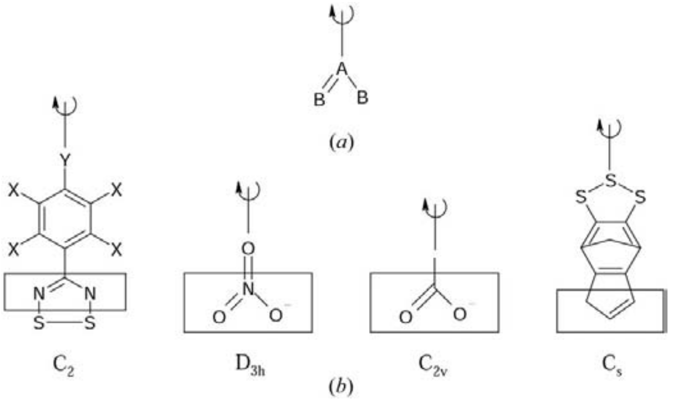


Figure 1
(a) Illustration of a $B{-}A{=}B$ group with a symmetry element that will not be found with the *RPluto* symmetry detection algorithm with the default tolerances of 0.1 Å and 5.0°. There will be other functional groups that behave similarly, not illustrated here. (b) A selection of molecules or fragments of molecules taken from the CSD, drawn using the CSD bond-type code convention. From left to right *para*-bromotetrafluorophenyl-1,2,3,5-dithiadiazolyl radical (Y = Br and X = F, CSD entry GOTQEL), nitrate ion, carboxylate group and *exo*-3,4,5-trithiatetracyclo-(5.5.1.0^{2,6}.0^{8,12})tridec-10-ene, (CSD entry DIVRIJ). The perceived molecular symmetry is appended to each structure. The $B{-}A{=}B$ group is highlighted with a rectangle, and a symmetry element $C_2(2)$ or $C_s(m)$, which might not be found by the symmetry detection algorithm at the default tolerances, is also shown.

2.4. Supplementing the dataset

It was thought to be useful to include information other than that generated by *RPluto*, and hence a local Fortran program was written to read the retrieved FDAT files and supplement the basic table. The further information extracted was space-group number, the number of atoms and the number of H atoms per molecule (or ion), the R-factor of the crystal structure, whether a metal atom is present in the molecule, the SIGCC flag,[3] and the symmetry of the occupied Wyckoff position(s). Unfortunately the CSD does not contain an easily-extractable marker or flag that details whether a structure is enantiomerically pure or not. Thus the dataset does not include any information regarding the chirality of the molecules or the optical activity of the structures.

The algorithm that determines the symmetry of the occupied Wyckoff position(s) uses symmetry-related atoms (*Satoms*) stored within the CSD entry: these are atoms generated by space-group symmetry that, together with atoms from the asymmetric unit, complete the chemical structure of a molecule or ion. Their coordinates and atom labels can be used to recognize the space-group symmetry operators that were used to generate the complete molecule. Additional checks are performed for symmetry operators that are known to be present in the space group but which are not detected through the presence of *Satoms*: for example, if a planar molecule is lying on a mirror plane. In these situations, the presence of the symmetry operator is tested by using the full coordinate set (Cole, 1995). It is known that a molecule can occupy a Wyckoff position of lower symmetry than its molecular point group; an example is illustrated in Fig. 2, and the choice of Wyckoff position occupied by a molecule may be of significant interest.

The additional data items were appended to the data generated by *RPluto* to complete the final table, exemplified for a selection of entries in Table 1. It should be noted that this table is assembled on a molecular basis; it contains an entry for each crystallographically unique molecule or ion in each crystal structure that has more than two non-H atoms. This table, which summarizes crystallographic information as well as molecular and crystal symmetry properties, is a very rich source of information, and the size of the dataset from which it was generated allows a thorough exploration of the possible relationships between crystallographic and molecular symmetry. However, in flat-file format, the information is not readily accessible.

2.5. Creation of the database

In order to interrogate these data with a wide range of queries, the table has been placed in a relational database

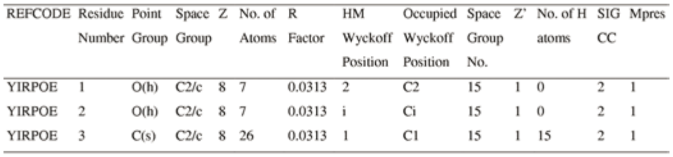
[3] SIGCC categorizes the mean structural precision for C—C bonds in each structure on a scale of 1–4; 1 indicates a high degree of precision and 4 indicates a low precision.

Table 1
Six entries in the table constructed by the systematic extraction of symmetry-related information from the CSD.

This table forms the basis of the relational database *CSDSymmetry*. The last column in the table (Mpres) is a flag that has the value of 1 if a metal atom is present and 0 otherwise.

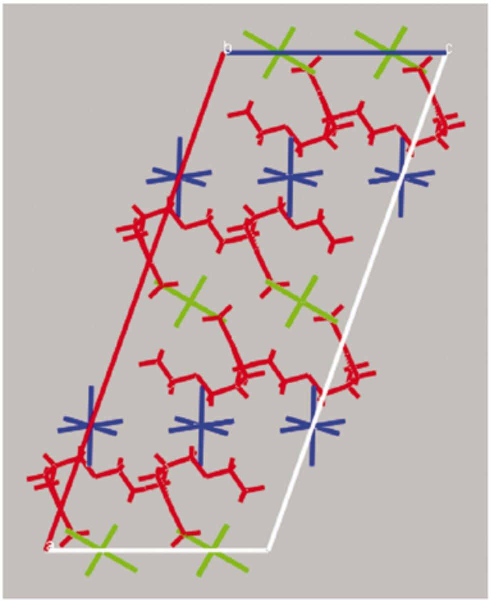
Refcode	Residue number	Point group	Space group	Z	Number of atoms	R-factor	HM Wyckoff position	Occupied Wyckoff position	Space-group number	Z'	Number of H atoms	SIGCC	Mpres
PIZYAY	1	$C(1)$	$Pna2_1$	4	41	0.057	1	$C1$	33	1	23	3	0
PIZYEC	1	$C(1)$	$P2_1/n$	4	44	0.046	1	$C1$	14	1	20	1	0
PIZYIG	1	$C(s)$	$P\underline{2}_1$	2	40	0.035	1	$C1$	4	1	20	2	0
PIZYOM	1	$C(1)$	$P\bar{1}$	2	87	0.027	1	$C1$	2	1	34	3	1
PIZZAZ	1	$C(i)$	$P2_1/n$	4	13	0.048	−1	Ci	14	1	0	0	1
PIZZAZ	2	$C(1)$	$P2_1/n$	4	17	0.048	1	$C1$	14	1	0	0	1

using Microsoft *Access* and supplemented with additional tables. The file containing the CSD-derived information was manipulated so that it could be read into *Access* as a plain-text file and as a single table, denoted as *CSDSymmetry*. Upon entry into *Access*, it was found that there were 106 entries for which the occupied Wyckoff position had not been determined. These records were removed from the database, and thus the database contains information for 198 229 molecules.



REFCODE	Residue Number	Point Group	Space Group	Z	No. of Atoms	R Factor	HM Wyckoff Position	Occupied Wyckoff Position	Space Group No.	Z'	No. of H atoms	SIG CC	Mpres
YIRPOE	1	O(h)	C2/c	8	7	0.0313	2	C2	15	1	0	2	1
YIRPOE	2	O(h)	C2/c	8	7	0.0313	i	Ci	15	1	0	2	1
YIRPOE	3	C(s)	C2/c	8	26	0.0313	1	C1	15	1	15	2	1

(*a*)



(*b*)

Figure 2
(*a*) Extract from *CSDSymmetry* showing the database entry for CSD Refcode YIRPOE (Childs *et al.*, 1994). There are three molecules listed for this crystal structure. (*b*) Crystal structure (Childs *et al.*, 1994) of 2-methyl-5-ethoxymethyl-1,3-dioxan-2-ylium hexachloroantimony (Refcode YIRPOE). The molecules are coloured by symmetry equivalence; the octahedral $SbCl_6^-$ ions occupy Wyckoff positions of $C_2(2)$ (green) and i ($\bar{1}$) (blue) symmetry.

The database has been supplemented with two further tables. A descriptive table for molecular point groups (*PointGroupInfo*) contains 38 common point groups (including non-crystallographic point groups such as C_5) and their component symmetry elements. Similarly, a descriptive table (*SpaceGroupInfo*) contains the symmetry of the Wyckoff positions that characterize each of the 230 space groups, and the table *SymmetryOps* contains the symmetry operators for each of the 230 space groups. The *CSDSymmetry* table is related to the auxiliary table *PointGroupInfo* through the field 'point group' and to the tables *SpaceGroupInfo* and *SymmetryOps* through the field 'space group number'. The inclusion of the descriptive tables allows a wide range of searches to be performed; for example, all space groups that contain a Wyckoff position with symmetry m (C_s), or all point groups containing a mirror plane, can be located easily.

The database has facilities that permit analysis of the full table contents or of subsets resulting from queries. For example, results can be sorted in descending order of occurrence, grouped, counted or manipulated with user-defined equations. An illustrative example query is: *'Find the point group with the greatest number of molecules in the space group P2₁/c, and present the result as a percentage of the total number of molecules in the database.'* This requires that all molecules found in crystal structures belonging to space group $P2_1/c$ are collated by their point group, the number of molecules per point group is counted, and the counts are sorted in descending order (the answer is C_1, see Fig. 3). The number of molecules of $C_1(1)$ symmetry found in $P2_1/c$ is then divided by the total number of molecules in the database *via* a user-defined equation to give the answer of 11.4%. Further, those molecules that belong to point group C_1 and space group $P2_1/c$ can be gathered as a Refcode list. The three-dimensional structure visualizer *Mercury* 1.0 (Taylor & Macrae, 2001; Bruno *et al.*, 2002) accepts a list of Refcodes as input (from a .gcd file) so that the results of the query can be viewed. In suitable cases, Microsoft *Excel* can also be used to present query results in graphical form (Fig. 3).

Once the *CSDSymmetry* database has been downloaded, it can be customized by the user in many ways: additional fields can be added to the existing tables or whole new tables can be added to the database. The results from a query can be saved and this subset, a user-defined database, can be used as a basis for more selective queries. Importantly, the results of any chemical or substructural search in *ConQuest* (Bruno *et al.*, 2002) can be intersected with the *CSDSymmetry* database. Thus, a *ConQuest* search for the presence of a carboxylic acid group will return a list of Refcodes that can be entered into the *CSDSymmetry* database. Where a Refcode from the *ConQuest* search matches a Refcode in the *CSDSymmetry* database, crystallographic and molecular symmetry information can be extracted and analyzed.

2.6. Validation

Independent examinations of the CSD by several authors over the last few years have yielded symmetry-related correlations and their results provide suitable queries with which to perform an initial validation of *CSDSymmetry*. Brock & Dunitz (1994) have outlined a number of correlations observed between space group and molecular symmetry. For example, the authors state 'inversion centres are favourable' and go on to show that centrosymmetric molecules often occupy inversion centres in space groups. A search of *CSDSymmetry* (with the criterion that the number of residues must not be greater than one) reveals that there are 18 008 molecules that belong to point groups containing the symmetry element i ($\bar{1}$). Of these, 17 152 molecules (95.2%) crystallize in space groups containing a Wyckoff position of symmetry i ($\bar{1}$), and 15 156 molecules (88.4%) utilize the crystallographic inversion centre.

Brock & Dunitz (1994) also state 'groups with threefold axes do not usually occur unless axes are located within molecules of the appropriate symmetry'. This conclusion was drawn from space groups having special positions of C_3 symmetry only. A search of *CSDSymmetry* revealed 2351 crystal structures where the space group contains a threefold axis; the presence of other special positions was permitted. There are 1158 structures with an occupied Wyckoff position of symmetry $C_3(3)$, and all of these special positions are occupied by molecules belonging to a point group that contains a threefold rotation axis. Brock & Dunitz (1994) also note 'twofold rotation axes are sometimes occupied and sometimes not'. Of the 198 229 molecules within the database, 24 332 (12.3%) are found within space groups that contain a twofold axis. Of these, 8760 molecules (36.0%) are located on the twofold axis.

Thus, concepts that were established from relatively small and hand-edited datasets, and which have been corroborated by a number of authors, are fully supported by the more extensive dataset contained within the *CSDSymmetry* database.

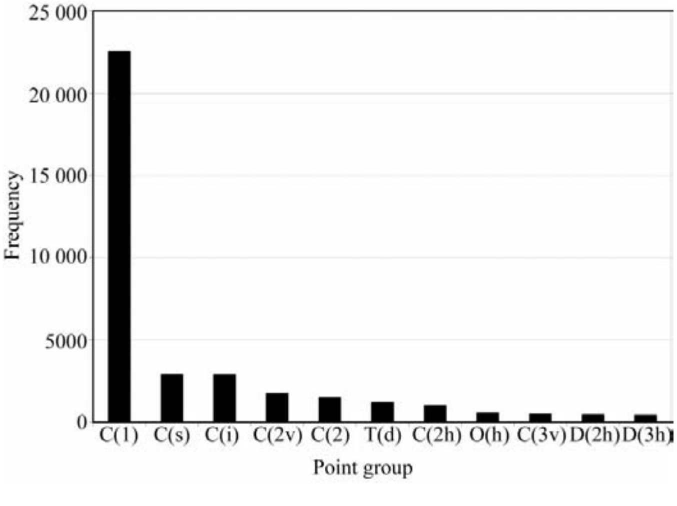


Figure 3
Histogram plotted from data retrieved using *CSDSymmetry* showing the 11 most prevalent molecular point groups and their frequencies in crystal structures belonging to the space group $P2_1/c$.

3. Example applications of *CSDSymmetry*

Some further examples of the use of *CSDSymmetry* have been outlined below to exhibit the utility of the database. These examples are meant to be illustrative rather than an exhaustive demonstration of the types of queries that can be put to the database. Detailed statistics and specialized studies of symmetry relationships determined using *CSDSymmetry* will be presented in later publications.

Example 1: *What space groups are represented by molecules of common point groups*? In the field of crystal structure prediction, it may be informative to know what impact the molecular point group has on the choice of space group (Belsky *et al.*, 1995; Scaringe, 1991). For the entire database, the five most populated space groups are (in descending order) $P2_1/c$, $P\bar{1}$, $P2_12_12_1$, $P2_1$ and $C2/c$. These five space groups account for 81.7% of the structures contained within the database; $P2_1/c$ (space-group number 14) alone accounts for 36.6% of the database. For a molecule of the point group $C_2(2)$,[4] the ten most populated space groups are (in descending order) $C2/c$, $P2_1/c$, $P\bar{1}$, $Pbcn$, $P2_12_12_1$, $P2_1$, $P2/c$, $Pbca$, $C2$ and $P2_12_12$. Of these space groups, $C2/c$, $Pbcn$, $P2/c$, $C2$ and $P2_12_12$ contain a Wyckoff position of symmetry $C_2(2)$, and approximately 92% of the molecules are found residing on such an axis. The three most populated of these space groups ($C2/c$, $Pbcn$ and $P2/c$) also contain a centre of inversion. The next three space groups that contain a twofold axis are $C2$, $P2_12_12$ and $Fdd2$; these space groups only contain the C_2 Wyckoff position and no centre of inversion. Why do molecules of C_2 symmetry seem to prefer space groups with a twofold axis and an inversion centre? Wilson (1993) stated that if a special position of a crystal structure is occupied then the packing of the crystal is '... as if the symmetry of the space group were degraded to that of a subgroup lacking the molecular symmetry in question'. The maximal non-isomorphic subgroups (with the C_2 axis removed) of $C2/c$, $Pbcn$ and $P2/c$ include $P\bar{1}$ and $P2_1/c$, the two most populated space groups of the entire database. The maximal non-isomorphic subgroups generated on removal of the C_2 axis

[4] The database does not include information on the optical activity of the compounds and so no distinction is made between enantiomerically pure C_2 structures and racemic C_2 structures.

from $C2$, $P2_12_12$ and $Fdd2$ are $P1$, $P2_1$ and Cc. These space groups are less common than $P2_1/c$ and $P\bar{1}$. Therefore, the observation that C_2 molecules prefer space groups that contain Wyckoff positions of $C_2(2)$ and i ($\bar{1}$) symmetry is a consequence of 'accidental symmetry': really it is a preference for the packing motif found in the space groups $P2_1/c$ or $P\bar{1}$ that is expressed. In other words, a molecule with C_2 symmetry finds favourable close packing in space group $P2_1/c$. Maintaining the same packing motif, but with its twofold axis aligned with a C_2 special position, the space group becomes $C2/c$ with a half-molecule as the asymmetric unit.

For a molecule of point group $C_s(m)$, the five most populated space groups in descending order are $P2_1/c$, $P\bar{1}$, $Pnma$, $P2_12_12_1$ and $C2/c$. The promotion of the space group $Pnma$ from the tenth most populated in the entire database to the third most populated for molecules of $C_s(m)$ symmetry can be explained by the following: $Pnma$ is the only space group of the top ten most populated space groups that contains a Wyckoff position of symmetry $C_s(m)$ in the database. Of the 1315 molecules of $C_s(m)$ symmetry that are found in $Pnma$, only 42 (3.2%) do not crystallize on the special positions $C_s(m)$ or $C_{2v}(mm)$. Again it is interesting to note that a maximal non-isomorphic subgroup of $Pnma$ (with the inversion centre removed) is $P2_12_12_1$, the third most populated space group of the *CSDSymmetry* database.

Example 2: *Are the most popular space groups the same for metal-containing molecules and purely organic molecules*? The five most popular space groups for *CSDSymmetry* entries that do not contain a metal atom are the same as for the entire database, *i.e.* $P2_1/c$, $P\bar{1}$, $P2_12_12_1$, $P2_1$ and $C2/c$ in descending order. If the criterion that a metal atom is present is imposed[5] and the query is performed again, the five most popular space groups, in descending order, are $P2_1/c$, $P\bar{1}$, $C2/c$, $P2_12_12_1$ and $Pbca$. Therefore, in the metal-containing dataset, the space group $P2_1$ (no special positions) has become less prominent and $Pbca$, a space group with an inversion centre, has been promoted. A histogram showing the distributions of metal-containing and non-metal-containing molecules is shown in Fig. 4. Searching *CSDSymmetry* reveals that of the 112 198 molecules for which the metal-present flag is equal to one, half the dataset (56 610 records, 50.4%) are of $C_1(1)$ point-group symmetry. This figure is closer to 70% for the organic dataset (86 031 structures, 59 216 belong to point group C_1). Consequently, approximately half of the metal-containing dataset (48.6%) belong to point groups with some symmetry: at least an inversion centre and/or a twofold rotation axis and/or a mirror plane. For the organic dataset, substantially less than half (only 30.8%) of molecules belong to point groups with symmetry. Thus, in broad terms, it can be concluded that metal-containing molecules exhibit more symmetry than purely organic molecules. The popularity of space group $Pbca$ in the metal-containing dataset is perhaps an indication that symmetric molecules have a preference for space groups with symmetry.

Example 3: *Do high symmetry space groups prefer highly symmetric molecules*? There are 1420 molecules with more than five atoms found within space groups that contain both of the special positions of symmetry $C_s(m)$ and $C_2(2)$. Of these 1420 molecules, 1079 (76.0%) belong to high-symmetry point groups [defined as point groups other than $C_i(\bar{1})$, $C_s(m)$, $C_2(2)$ or $C_1(1)$]. There are 2299 molecules with more than five atoms found in structures with no crystallographic symmetry (space group $P1$). Of these structures, only 167 (7.2%) belong to high-symmetry point groups (as defined above); a very large proportion (84.1%) belong to the point group $C(1)$. Therefore there seems a clear preference for high-symmetry space groups to accommodate molecules of a high symmetry. The space group $P1$, with no crystallographic symmetry, has a clear preference for molecules with low or no symmetry. It was established in the previous example that metal-containing complexes are found in point groups characterized by operations other than the identity operation more often than organic molecules. In *CSDSymmetry*, 54.3% of the entries with more than five atoms contain a metal atom. Of the 1420 molecules with more than five atoms found within space groups that contain both of the special positions of symmetry $C_s(m)$ and $C_2(2)$, a large proportion (70.9%) contain a metal atom. The molecules encompassed by the space group $P1$ are predominantly organic; only 32.6% contain a metal atom. The lower than expected representation of symmetric metal complexes in $P1$ (and conversely the higher than expected proportion of metal complexes in symmetric space groups) is an indication that distinct relationships between point-group and space-group symmetry exist. These relationships will be expounded on in later publications.

Example 4: *Which point groups have the greatest proportion of molecules in non-centrosymmetric space groups*? As mentioned earlier, crystallization in a non-centrosymmetric

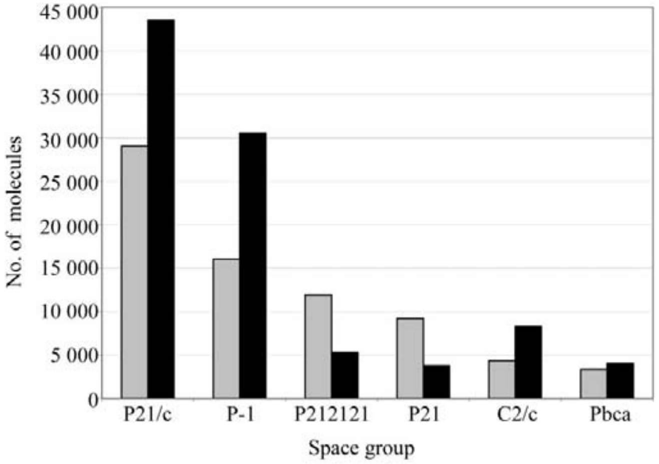


Figure 4
Histogram showing the distribution of metal-containing and organic molecules through a subset of space groups. The space groups shown are the six most populated for the entire *CSDSymmetry* database. Black: metal-containing molecules. Grey: organic molecules.

[5] The metal-present flag (Mpres) is Refcode-specific. Thus, if a structure contains four residues and three residues contain a metal atom, all entries in *CSDSymmetry* for the Refcode have the metal-present flag set to one. A subset of *CSDSymmetry* can be generated for which the number of residues in the structure equals one, and therefore if Mpres = 1, a metal is present in that residue.

Table 2
The results of a query constructed in the Microsoft *Access* database, *CSDSymmetry*, that determines which point groups contain the highest proportion of their molecules in non-centrosymmetric space groups.

Point group	% of molecules found in space groups that do not contain a centre of inversion	Number of molecules in *CSDSymmetry*
C_4	42.9	70
C_3	41.1	1410
C_1	30.7	115826
S_4	29.5	1162
D_3	26.0	580
T_d	24.4	6220
D_2	23.6	977
C_2	23.1	13474
D_{5h}	23.1	96
D_{2d}	21.9	1420

space group is a requirement for non-linear optical behaviour. The number of molecules, grouped by point group, contained in space groups that do not contain a centre of inversion has been determined and compared with the number of molecules found in each point group for the entire *CSDSymmetry* database. A user-defined equation was formulated to calculate the percentage of molecules of a particular point group that are found in non-centrosymmetric space groups: the ten point groups with the highest percentage of molecules in non-centrosymmetric space groups are given in Table 2. A criterion was placed on the query that there had to be more than 30 occurrences of the point group in the whole database. No criterion concerning the number of atoms was applied. The molecular point group that contains the highest proportion (43%) of its molecules within non-centrosymmetric space groups is C_4(4). Is this an indication that the synthesis of molecules with C_4(4) symmetry is a promising route to materials that form in non-centric space groups and that may behave as non-linear optics?

4. Conclusions

Molecular and crystallographic symmetry properties have been extracted from the CSD and collected together in a relational database. This is the most complete collation of observed molecular and crystallographic symmetry properties to date, comprising nearly 200 000 entries. The software package Microsoft *Access*, which was used to create the *CSDSymmetry* database, provides a very flexible way of creating queries and analyzing the data contained within. The database can be customized by the user with the addition of further fields or tables. The inclusion of CSD Refcode entries in the database allows the user to (*a*) match the results of *ConQuest* searches with information contained within the database and (*b*) view the results of queries in the molecular visualization package *Mercury* 1.0. It is hoped that the collation and publication of the information contained within this database will provide the scientific community with the tool it needs to access and investigate relationships between molecular and crystallographic symmetry properties.

5. Technical details

The entire symmetry database is contained in one Microsoft *Access* file, CSDSymmetry.mdb. This file is approximately 60 MB. A self-extracting archive (20 MB) has been created, which contains the database and a short help file. This help file contains two further example searches with a step-by-step guide as to how to perform them. Microsoft *Access* has its own help facilities that give information on how to use the software package and detailed information on how to construct queries. The download package will be free for non-commercial research purposes and will be available at the URL http://www.ccdc.cam.ac.uk in due course.

JAKH thanks the Engineering and Physical Sciences Research Council (UK) for the award of a Senior Research Fellowship.

References

Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.

Antorrena, G., Davies, J. E., Hartley, M., Palacio, P., Rawson, J. M., Smith, J. N. B. & Steiner, A. (1999). *Chem. Commun.* pp. 1393–1394.

Belsky, V. K., Zorkaya, O. N. & Zorky, P. M. (1995). *Acta Cryst.* A**51**, 473–481.

Brock, C. P. & Duncan, L. L. (1994). *Chem. Mater.* **6**, 1307–1312.

Brock, C. P. & Dunitz, J. D. (1994). *Chem. Mater.* **6**, 1118–1127.

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.

Cambridge Crystallographic Data Centre. (1994). *Quest3D – A Program for Searching the Cambridge Structural Database.* Cambridge Crystallographic Data Centre, Cambridge, UK.

Childs, R. F., Kang, G. J., Wark, T. A. & Frampton, C. S. (1994). *Can. J. Chem.* **72**, 2084–2093.

Cole, J. C. (1995). PhD thesis. University of Durham, England.

Cole, J. C., Yao, J. W., Shields, G. P., Motherwell, W. D. S., Allen, F. H. & Howard, J. A. K. (2001). *Acta Cryst.* B**57**, 88–94.

Filippini, G. & Gavezzotti, A. (1992). *Acta Cryst.* B**48**, 230–234.

Lloyd, M. A. & Brock, C. P. (1997). *Acta Cryst.* B**53**, 780–786.

Meyer, F., Kaifer, E., Kircher, P., Heinze, K. & Pritzkow, H. (1999). *Chem. Eur. J.* **5**, 1617–1630.

Motherwell, W. D. S., Shields, G. P. & Allen, F. H. (1999). *Acta Cryst.* B**55**, 1044–1056.

Scaringe, R. P. (1991). *Electron Crystallography of Organic Molecules*, edited by J. R. Fryer and D. Dorset, pp. 85–113. Dordrecht: Kluwer.

Schweizer, W. B., Proctor, G., Kaftory, M. & Dunitz, J. D. (1978). *Helv. Chim. Acta*, **61**, 2783–2808.

Steiner, T. (2000). *Acta Cryst.* B**56**, 673–676.

Taylor, R. & Macrae, C. F. (2001). *Acta Cryst.* B**57**, 815–827.

Watson, W. H., Jain, P. C., Bartlett, P. D. & Ghosh, T. (1986). *Acta Cryst.* C**42**, 332–334.

Wilson, A. J. C. (1993). *Acta Cryst.* A**49**, 795–806.

Zorky, P. M., Potekhin, K. A. & Dashevskaya, E. E. (1993). *Acta Chim. Hung.* **130**, 221–233.