Retrieval of Crystallographically-Derived Molecular Geometry Information

Ian J. Bruno, Jason C. Cole, Magnus Kessler, Jie Luo, W. D. Sam Motherwell, Lucy H. Purkis,
Barry R. Smith, and Robin Taylor*

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

Richard I. Cooper

Chemical Crystallography, University of Oxford, Chemistry Research Laboratory, Mansfield Road,
Oxford OX1 3TA, England

Stephanie E. Harris and A. Guy Orpen

School of Chemistry, University of Bristol, Bristol BS8 1TS, England

Received July 13, 2004

The crystallographically determined bond length, valence angle, and torsion angle information in the Cambridge Structural Database (CSD) has many uses. However, accessing it by means of conventional substructure searching requires nontrivial user intervention. In consequence, these valuable data have been underutilized and have not been directly accessible to client applications. The situation has been remedied by development of a new program (Mogul) for automated retrieval of molecular geometry data from the CSD. The program uses a system of keys to encode the chemical environments of fragments (bonds, valence angles, and acyclic torsions) from CSD structures. Fragments with identical keys are deemed to be chemically identical and are grouped together, and the distribution of the appropriate geometrical parameter (bond length, valence angle, or torsion angle) is computed and stored. Use of a search tree indexed on key values, together with a novel similarity calculation, then enables the distribution matching any given query fragment (or the distributions most closely matching, if an adequate exact match is unavailable) to be found easily and with no user intervention. Validation experiments indicate that, with rare exceptions, search results afford precise and unbiased estimates of molecular geometrical preferences. Such estimates may be used, for example, to validate the geometries of libraries of modeled molecules or of newly determined crystal structures or to assist structure solution from low-resolution (e.g. powder diffraction) X-ray data.

INTRODUCTION

The Cambridge Structural Database (CSD)[1] contains the results of about 320 000 (July 2004) crystal-structure determinations of organic and metallo-organic compounds. It has long served as a source of experimental information on molecular geometries for crystallographers, structural chemists, drug designers, etc. In the latter field, the most important geometrical parameters are torsion angles around rotatable bonds, since they influence the overall shapes of molecules—and hence their biological activities—far more than do bond lengths and valence angles. Many theoretical techniques for estimating torsional preferences are available but none is without problems. Supplementing energy calculations with a CSD-based conformational analysis therefore increases the confidence with which conclusions may be drawn. CSD torsional data have been used in programs (e.g. MIMUMBA,[2] et[3]) for generating low-energy conformations. The protein−ligand docking programs GOLD[4] and FlexX[5] use torsional distributions from the CSD to bias docking solutions toward low-energy ligand geometries. Several groups have published molecular superposition methods that utilize crystallographic information; for example, the CSD-derived MIMUMBA torsion-angle database is used in FlexS.[6] Crystal-packing forces occasionally bias conformations in the crystalline state, but there is ample evidence to show that molecular geometries in crystal structures are usually a good guide to conformational preferences in aqueous solution or at protein binding sites.[7,8]

For crystallographers and structural chemists, the main interest is often in bond lengths and angles. Comparison of the dimensions of a newly determined small-molecule crystal structure with the bond lengths and angles of similar structures in the CSD is useful as a check against refinement errors and to highlight unusual geometrical features. Average bond lengths and their standard errors from the CSD can be used to set up ligand dictionaries for restrained refinement of protein−ligand crystal structures or for restrained Rietveld refinement of powder-diffraction structures. CSD-based printed compilations of the means, medians, and standard deviations of organic and organometallic bond lengths[9,10] have been widely used for purposes such as this.[11]

Despite the value of the molecular geometry information in the CSD, there are some significant problems in its use. At a time when molecular and quantum mechanics calculations can be run “at the touch of a button”, deriving a distance or angle distribution from the CSD necessitates several steps. Suppose, for example, that a crystallographer has determined

* Corresponding author phone: +44 1223 336020; fax: +44 1223 336033; e-mail: taylor@ccdc.cam.ac.uk.

10.1021/ci049780b CCC: $27.50 © 2004 American Chemical Society
Published on Web 10/13/2004

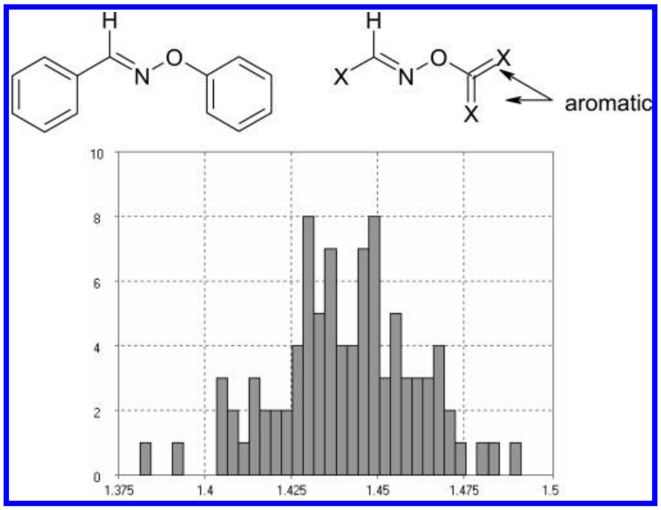


Figure 1. An example query molecule, together with a substructure that might be searched for to determine the likely value of the N−O bond length in the query (X = any atom except hydrogen); the resulting histogram of N−O bond lengths in CSD entries containing this substructure.

the structure of the molecule in Figure 1 and wishes to know whether the N−O bond length is unusual. The CSD must be searched for a substructure such as the one illustrated and a histogram (Figure 1) produced of the lengths of the N−O bonds in the hits. While straightforward, this is relatively time-consuming because the substructure has to be decided upon and drawn. Thus, few are likely to be willing to search for *all* bonds and angles in a molecule. Similarly, a modeler might perform CSD searches to check that the torsion angles of a computer-docked protein ligand are consistent with those seen in crystal structures, but this is impracticable for each of a library of many thousands of docked ligands. The printed bond-length compilations[9,10] are convenient for manual use but cannot be accessed directly by computer programs and include no angle information. Further, the compilations were based on a version of the CSD containing only about 50 000 crystal structures. Similarly, the CSD-derived torsion distributions used in programs such as MIMUMBA and GOLD are snapshots based on the data available several years ago.

We have therefore developed a program, Mogul, which addresses these problems. The aim of Mogul is to take a molecule—submitted either manually or by another computer program via an instruction-file interface—and perform substructure searches of the CSD for, typically, all its bonds, angles, and acyclic torsions. The crystallographically determined distributions of the bond lengths, valence angles, and torsion angles are returned to the user or client application. This requires the following: automatic generation of the search substructures; fast searching for the substructures and retrieval of the required geometrical information; remedial action if any substructure fails to find sufficient hits in the CSD; and presentation of results either graphically or in machine-readable form. These requirements are met as follows. Each *fragment* (i.e. bond, angle, or acyclic torsion) in the query molecule is assigned a set of key values, which collectively describe the substructural environment of the fragment. A search tree is traversed to find all fragments in CSD structures that have identical key values; this is approximately equivalent to performing a substructure graph-match. Should insufficient hits be obtained, a tree backtracking algorithm is combined with similarity calculations to find CSD fragments with approximately the same key values as the query fragment (conceptually similar to searching for a substructure containing one or more variable atom and/or bond types). Statistics (mean, standard deviation, histogram, etc.) for the dimension of interest are generated using the experimentally determined 3D coordinates of the hits. These may be viewed graphically or read by an external application. Further details of these steps are given in the remainder of this article.

SEARCH ALGORITHM

Keys, Key Components, and Key Types. The Mogul search algorithm requires a set of keys to be defined that can be used to describe the chemical environment of any given molecular fragment. Each key consists of one or more *key components* concatenated together. A given key will contain exclusively atom-based or fragment-based components. Examples of atom-based key components are as follows: atomic number (*AtomicNumber*); number of connected non-hydrogen atoms (*NonHCount*); and bond types of all bonds formed by the atom (*Bonds*). Fragment based components include the following: whether the fragment is cyclic (*LinearFragBondCyclicity*) and whether the fragment contains any metal atoms (*NonMetalAtoms*). A list of key components is given in Table 1. Many others can be envisaged. An atomic charge component is not currently used because charge assignment in the CSD is sometimes arbitrary. Components such as oxidation state will be necessary to produce well-defined bond-length distributions for bonds involving metals.

Once the set of keys is defined, the chemical environment of a given fragment can be described by evaluating the keys for specific atoms in the fragment or for the fragment as a whole, depending both on the nature of the components in the key (i.e. whether they are atom- or fragment-based) and on the *key type*. At present, two key types are used, *Standard* and *ConnectedAtomsNonH*. A *Standard* key is evaluated either for a complete fragment or a particular atom of the fragment, depending on the nature of the key component(s) that the key contains. A *ConnectedAtomsNonH* key can only contain atom-based components and is evaluated for all the non-hydrogen atoms bonded to a particular atom of the fragment. Table 2 gives an example. The key definitions and types are summarized in the first section of the table. The second section shows the results of applying the keys to the N−O bond fragment of Figure 1 or to particular atoms of that fragment. Key 1 is of *Standard* type and contains a fragment-based key component and is hence evaluated for the fragment as a whole (specifically, it is the bond type of the N−O bond, viz. 1). Keys 2 and 3 are of *Standard* type and comprise atom-based key components and are evaluated for specific atoms of the fragment. The remaining keys are of *ConnectedAtomsNonH* type and (necessarily) consist of atom-based key components; they are hence evaluated for all non-hydrogen atoms connected to specific atoms of the fragment. Applying the keys as indicated in the table results in nine different key values which, collectively, describe the chemical environment of the N−O bond to the level of precision allowed by this particular scheme.

Table 1. List of Key Components

key component	explanation
	Fragment
IncludesH	true if any fragment atom is hydrogen
LinearFragBondCyclicity	one or zero for each bond in fragment, depending on whether bond is cyclic or not, e.g. 1,1 for an O−C═C angle in furan
LinearFragBonds	bond types of all bonds in fragment, e.g. 1,2 for an O−C═C angle in furan
NonMetalAtoms	true if all fragment atoms nonmetallic
TermFragRing	for each bond in fragment (e.g. the two bonds of a valence angle), size of smallest ring containing bond; appended by one if bonds are in the same ring, zero if not; e.g. 5,5,1 for an O−C═C angle in furan
TermLinearFragBonds	bond types of outer bonds of fragment, e.g. of W−X and Y−Z bonds for a W−X−Y−Z torsion fragment
	Atom
AnyConnectedMetal	true if atom is bonded to a metal
AtomCyclicity	true if atom is cyclic and smallest ring containing atom has <9 atoms
AtomicNumber	e.g. 1 for hydrogen and deuterium
AtomicSymbol	e.g. H for hydrogen, D for deuterium
AtomSmallestRing	size of smallest ring containing atom (zero if atom is acyclic or smallest ring > 8 atoms)
Bonds	bond types of all bonds formed by atom, e.g. 1.1.2 for aldehyde carbonyl carbon
BondsNonH	bond types of all bonds formed by atom to non-hydrogen atoms, e.g. 1.2 for aldehyde carbonyl carbon
HCount	number of hydrogens attached to atom, e.g. 1 for aldehyde carbonyl carbon
NConnections	number of bonds formed by atom, e.g. 3 for aldehyde carbonyl carbon
NonHCount	number of bonds formed by atom to non-hydrogen atoms, e.g. 2 for aldehyde carbonyl carbon

Table 2. Some Typical Keys and the Results of Applying Them to the N−O Bond Fragment of Figure 1

Key Definitions

key	key type	components in key
1	Standard	LinearFragBonds
2	Standard	AtomicNumber + NConnections + HCount
3	Standard	Bonds
4	ConnectedAtomsNonH	AtomicNumber + NConnections + HCount
5	ConnectedAtomsNonH	Bonds

Results of Applying Keys to Specific Entities of the N−O Fragment of Figure 1

key	applied to	result
1	N−O fragment	1
2	N	7.2.0
2	O	8.2.0
3	N	1,2
3	O	1,1
4	N	6.3.1#8.2.0
4	O	6.3.0#7.2.0
5	N	1,1,2#1,1
5	O	1,ar,ar#1,2

Table 3. Key Scheme Currently Used for Valence Angles X−Y−Z

Key Definitions

key	key type	components in key
1	Standard	LinearFragBonds
2	Standard	AtomicNumber
3	Standard	TermFragRing
4	Standard	AtomSmallestRing
5	Standard	BondsNonH
6	Standard	NConnections + HCount
7	ConnectedAtomsNonH	AtomicNumber + NConnections
8	ConnectedAtomsNonH	AtomicNumber + NConnections + HCount
9	ConnectedAtomsNonH	BondsNonH
10	Standard	IncludesH
11	Standard	NonMetalAtoms

Filters

key	applied to	action
10	X−Y−Z	discard if true (rejects if X, Y or Z = H)
11	X−Y−Z	discard if false (rejects if X, Y or Z = metal)

Chemical-Environment Specification and Tree Definition

key	applied to	used to index tree levels
1	X−Y−Z fragment	1
2	Y, X, and Z (separately)	2 (Y), 3 (X), 4(Z)
3	X−Y−Z fragment	5
4	Y	6
5	Y, X, and Z (separately)	7 (Y), 8 (X), 9 (Z)
6	Y, X, and Z (separately)	10 (Y), 11 (X), 12 (Z)
7	Y, X, and Z (separately)	13 (Y), 14 (X), 15 (Z)
8	Y, X, and Z (separately)	16 (Y), 17 (X), 18 (Z)
9	Y, X, and Z (separately)	19 (Y), 20 (X), 21 (Z)

Key values frequently consist of several parts, either because the key contains several components and/or because a single component is evaluated for several entities. For example, key 2 contains three components, and its value for the nitrogen atom in Figure 1 is therefore the three-part string 7.2.0 (atomic number 7; bonded to two atoms; of which none is hydrogen). Key 3 has only one component but is multipart because its value consists of the bond types of *all* the bonds formed by an atom (e.g. has the value 1,1 for the oxygen atom of Figure 1). Key 4 has three components and, being of type *ConnectedAtomsNonH*, is evaluated for all non-hydrogen atoms connected to a given atom.

Use of Keys for Fragment Classification; Mogul Distributions. A complete *key scheme* (comprising the key definitions and the information about how they are to be applied) provides a means of both filtering and classifying fragments. Taking valence angles as an example, the key scheme of Table 3 was used to evaluate keys for every valence angle in the CSD. Those failing the filters were rejected; in practice, this meant eliminating angles X−Y−Z where any of X, Y, Z is a metal or a hydrogen. The surviving angles were grouped into *distributions*, such that all valence angles in a given distribution had identical key values. When archiving the distributions, information was stored about the geometrical values of the angles and the CSD entries whence they came. The resulting set of distributions comprise the Mogul angle library. Bond and torsion libraries were set up similarly; the key schemes used were similar but not identical to that shown in Table 3.

The existing suite of key components is inadequate for classifying metal-containing fragments (for example, we cannot subdivide by oxidation state). Also, most hydrogen-atom positions in the CSD are of low precision. As with angles, therefore, bonds (X−Y) and torsions (W−X−Y−Z) were excluded if any of W, X, Y, Z were metal or hydrogen atoms. Since torsion angles in a ring are highly interdependent, filtering on the *LinearFragBondCyclicity* key was used to exclude cyclic torsions from the torsion library (more precisely, torsions in rings of size 8 or less). Pi bonds between metals and poly-hapto-coordinated ligands count when determining which atoms should be included in the evaluation of keys of type *ConnectedAtomsNonH*. Consequently, for example, C−C−C valence angles in cobalt- and iron-coordinated cyclopentadienyl rings would be assigned to different distributions from each other. Ring-closure constraints have an important effect on valence angles so two keys reflecting ring size are used in the angle key scheme. One (key 3 of Table 3) comprises the fragment-based key component *TermFragRing*. The other (key 4) comprises the component *AtomSmallestRing* and is applied to the central atom of the valence angle. Consequent on the use of both keys, all of the angles shown in A−D of Figure 2 will be assigned to different distributions.

Search Trees. Searching in Mogul involves evaluating the keys for a query fragment and finding the Mogul distribution with the same set of key values. This is almost (see below) equivalent to performing a substructure search. Fast search speeds are achieved by use of a tree indexed on successive key values. Each leaf of the tree points to the location on disk where the distribution is stored containing the CSD fragments with the same key values as the query.

Key Canonicalization and Hashing. Since key values often consist of several parts, it is sometimes necessary for them to be canonicalized following arbitrary but consistently applied conventions. For example, irrespective of how atoms were ordered or labeled, the value of key 3 of Table 2 would always be returned as 1,2 for the nitrogen atom of Figure 1, not 2,1. Also, there are rules governing, e.g., which of the atoms in an X−Y−Z valence angle is considered X and which Z. The canonicalization rules are mundane but quite complex. For example, it is important in the angle key scheme (Table 3) that the values obtained when keys 1 and 3 are applied to the fragment are canonicalized in concert, not separately: if they are not, a query angle such as that shown in E of Figure 2 might hit both of the angles F, G rather than just the former. We decided, however, to canonicalize the values of *ConnectedAtomsNonH*-type keys independently of one another. This has the consequence that a Mogul search may not be exactly equivalent to a test on subgraph isomorphism. For example, the two bond fragments shown in H and I of Figure 2 are not isomorphous yet have the same Mogul key values because, after canonicalization, their *ConnectedAtomsNonH*-type keys are identical. We regard this as acceptable as such fragments will invariably be very similar.

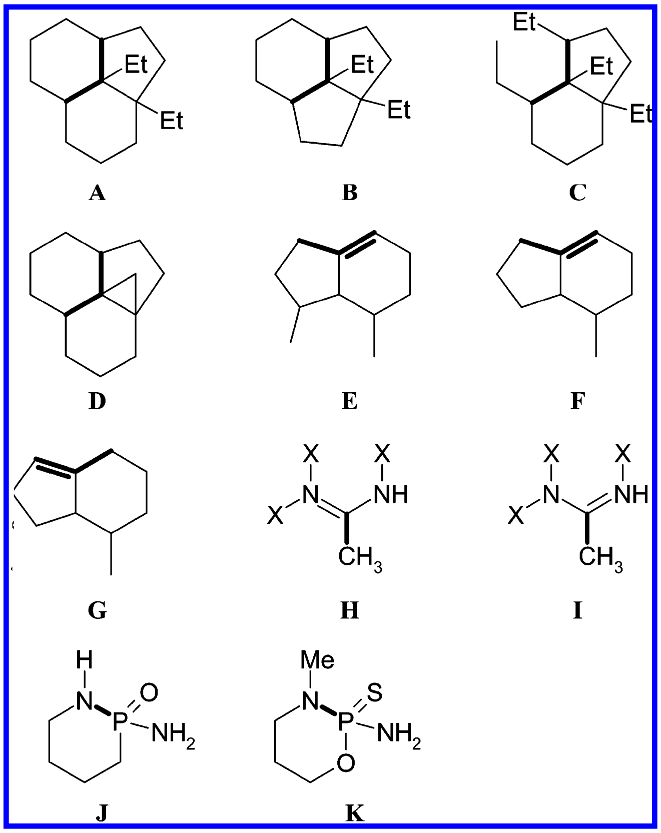


Figure 2. Example structures and substructures referred to in the text; bond or angle fragment of interest shown in bold, X = any atom except hydrogen.

Each key value is originally computed as a (possibly long) string. To save space, the string is hashed to a 32-bit integer. It is these integer values that are used in indexing and searching Mogul trees. Experiments indicate that the frequency of collisions is negligible, so no check is made.

Exact and Generalized Searching. The key scheme used in the Mogul bond library is such that a search for the N−O bond in Figure 1 would be equivalent to searching for the substructure shown in the figure, i.e., would include information extending two bonds out from the fragment atoms N and O. The key schemes used for the angle and torsion libraries also correspond to this degree of fragment-environment specification. At this level of precision, the results of searches tend to be very relevant to the query fragment and therefore a good guide to geometrical preferences. The concomitant disadvantage is that relatively few hits may be obtained. Table 4 gives a breakdown of the number of distributions in the Mogul libraries and the number of observations they contain. Many distributions contain small numbers of observations. Also, many chemically possible fragments have never been characterized by crystallography and are therefore absent from the libraries. Frequently, therefore, Mogul searches performed as described above (termed *exact searches*, since they find exactly one Mogul distribution) produce very few hits.

The problem is overcome by performing a *generalized search* when an exact search would find insufficient hits. During tree traversal, if the path taken runs out before

Table 4. Number and Size of Distributions in Bond, Valence-Angle, and Torsion Libraries

	type of distribution		
no. observations in distribution, N	bond	valence angle	torsion
$N < 5$	310456	910238	239703
$5 \leq N < 10$	44284	108661	41864
$10 \leq N < 50$	40221	86599	37070
$50 \leq N < 100$	5624	9956	4162
$100 \leq N < 10000$	6470	9676	3879
$N = 10000$[a]	76	85	20

[a] Any distribution containing >10000 observations is reduced to exactly 10000 by random selection.

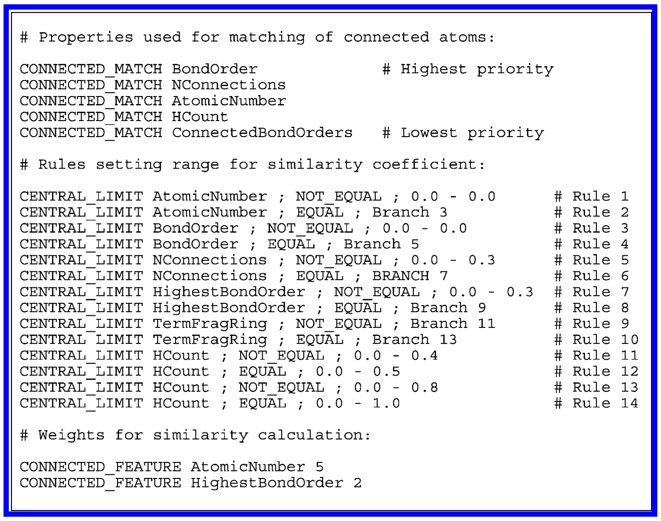
reaching the leaf level (i.e. an exactly matching distribution), or if that distribution contains insufficient observations, the algorithm backtracks to the preceding level. It then examines the distributions corresponding to *all* the leaves beyond that point. A coefficient is calculated for each distribution measuring the similarity between the query fragment and the fragment to which that distribution corresponds. Distributions exceeding a similarity threshold are collected and pooled. If there are still insufficient observations in the pool, another backtracking step is performed and the procedure repeated. This continues until enough observations are collected or it ceases to become chemically sensible to backtrack further.

Because of this methodology, the order in which key values are used to index the tree must be carefully chosen. Defining the *bottom* of the tree as being its leaves, key values coding for relatively insignificant features of the fragment environment must be used to index tree levels toward the bottom. Backtracking from the bottom upward then generalizes the least important features first. Conversely, the most important features (for angles, these include the bond types of X−Y and Y−Z and the atomic numbers of X, Y, and Z) are used to index levels toward the top of the tree. Some features are so important that it is not sensible to allow them to vary at all; thus, there is a limit to how far backtracking is allowed during a generalized search (to level 13 in the case of the angle tree).

Search Performance and Storage Requirements. Elapsed times for exact searches on every single bond, angle, and torsion of a molecule are typically a few seconds on a modern PC. R-factor filtering (see below) adds a noticeable but not serious overhead. Generalized searches, however, can take much longer, depending on the number of separate distributions that have to be accessed. For this reason, it is difficult to estimate how long a search for all fragments in a given molecule will take; a large molecule might be finished in a few seconds if all its fragments are common and only require exact searches. However, even a small molecule might take a few minutes if it contains fragments that require generalized searches that necessitate examining thousands of individual distributions. Search times can be controlled by setting a number of customizable parameters, e.g. the minimum number of hits desired for each fragment. Mogul data files occupy a total of about 625 megabytes, not counting the CSD itself.

SIMILARITY CALCULATION

Similarity Coefficient Definition. Molecular similarity is a well studied area.[12] However, our requirement when performing generalized searching is unusual in that it involves comparing two fragment environments (not molecules) and with a particular view to estimating how likely it is that the geometrical dimensions of the fragments will be similar. The method used reflects the mental processes we believe experienced chemists go through in assessing similarity and comprises three steps: atom matching; assessment of critical differences; and assessment of minor differences. The details of the calculation can be customized by a configuration file. Figure 3 shows the essential elements of the file used for defining the bond similarity calculation. The use of the file will be illustrated with reference to the bond fragment environments shown in J and K of Figure 2.



```
# Properties used for matching of connected atoms:

CONNECTED_MATCH BondOrder              # Highest priority
CONNECTED_MATCH NConnections
CONNECTED_MATCH AtomicNumber
CONNECTED_MATCH HCount
CONNECTED_MATCH ConnectedBondOrders    # Lowest priority

# Rules setting range for similarity coefficient:

CENTRAL_LIMIT AtomicNumber ; NOT_EQUAL ; 0.0 - 0.0      # Rule 1
CENTRAL_LIMIT AtomicNumber ; EQUAL ; Branch 3           # Rule 2
CENTRAL_LIMIT BondOrder ; NOT_EQUAL ; 0.0 - 0.0         # Rule 3
CENTRAL_LIMIT BondOrder ; EQUAL ; Branch 5              # Rule 4
CENTRAL_LIMIT NConnections ; NOT_EQUAL ; 0.0 - 0.3      # Rule 5
CENTRAL_LIMIT NConnections ; EQUAL ; BRANCH 7           # Rule 6
CENTRAL_LIMIT HighestBondOrder ; NOT_EQUAL ; 0.0 - 0.3  # Rule 7
CENTRAL_LIMIT HighestBondOrder ; EQUAL ; Branch 9       # Rule 8
CENTRAL_LIMIT TermFragRing ; NOT_EQUAL ; Branch 11      # Rule 9
CENTRAL_LIMIT TermFragRing ; EQUAL ; Branch 13          # Rule 10
CENTRAL_LIMIT HCount ; NOT_EQUAL ; 0.0 - 0.4            # Rule 11
CENTRAL_LIMIT HCount ; EQUAL ; 0.0 - 0.5                # Rule 12
CENTRAL_LIMIT HCount ; NOT_EQUAL ; 0.0 - 0.8            # Rule 13
CENTRAL_LIMIT HCount ; EQUAL ; 0.0 - 1.0                # Rule 14

# Weights for similarity calculation:

CONNECTED_FEATURE AtomicNumber 5
CONNECTED_FEATURE HighestBondOrder 2
```

Figure 3. Part of the configuration file for the bond similarity calculation.

The first step is to pair the atoms of J with those of K. The pairing of the atoms involved in the bond itself is obvious: P onto P, N onto N. (If the bond were between two atoms of the same elemental type, both pairings would be tried and the one accepted that gave rise to the highest similarity.) In pairing peripheral atoms, it is obvious that a phosphorus-bound atom in J can only be paired with a phosphorus-bound atom in K and similarly for the nitrogen-bound atoms. Subject to this constraint, each peripheral atom in J is compared with each in K and their similarity estimated. The most similar pair of atoms are paired off, then the next most similar, and so on until no more pairing is possible. The similarity of a given atom pair is estimated by setting up a bit string, each bit corresponding to a particular atom property, e.g. element type or coordination number. A bit is set on if the property value which it represents is the same in the two atoms, otherwise it is set off. The property considered most important corresponds to the left-most bit, the next most important to the next-to-left-most bit, and so on. The similarity of the atom pair is taken as the arithmetic value of the bit string. The first section of the configuration file defines the order of importance that atom properties have in this comparison. In Figure 3, the order is defined as the following: highest bond order of bonds formed by atom; total number of bonds formed by atom; atomic number; hydrogen count; and bond orders of all bonds formed by atom. For example, the double-bonded oxygen of J will be paired with the sulfur of K, not the oxygen, because highest bond order is more important than atomic number.

In the second step, critical differences between the two fragment environments are sought. If any are found, a limit is placed on the maximum possible value that the similarity coefficient can have. These limits are specified by the *rules* in the second section of the configuration file. For example, rule 1 of Figure 3 states that if the bonds are not between the same element pair the maximum possible value for the similarity coefficient is zero. Rule 5 states that if the total connectivities of either of the atoms in the bond are not the same (e.g. $H_3\mathbf{C}-\mathbf{C}H_3$ compared with $H_3\mathbf{C}-\mathbf{C}H{=}CH_2$), then the coefficient cannot exceed 0.3. Rules can be nested; for example, if the two bonds being compared are cyclic and in rings of different sizes (rule 9) but the hydrogen counts of both atoms are equal (rule 12), then the similarity coefficient cannot exceed 0.5. If, within the definition of the configuration file, there are no critical differences, the similarity coefficient may take a maximum value of 1.0.

Having thus established the possible range of the similarity coefficient (*min* to *max, min* is always zero in the current version of Mogul), the third and final step is to compute the exact value within that range from the formula

$$similarity = C\{min + [\Sigma_i\Sigma_j(wt_j\cdot match_{ij})/\Sigma_i\Sigma_j(wt_j)]\cdot[max - min]\} \quad (1)$$

The outer summations are over the pairs of peripheral atoms, *i*, derived in the first step in the algorithm; the inner summations are over atom properties, *j*. These properties are specified in the final section of the configuration file, i.e., in Figure 3, they are atomic number and the highest bond order of the bonds formed by the atom. wt_j is the weight of property *j* (e.g. 5 for atomic number); $match_{ij}$ is unity if the *j*th property has the same value in the *i*th pair of peripheral atoms; otherwise it is zero. The coefficient *C* is set to 0.99 unless the distribution for which the calculation is being performed is an exact match (in Mogul terms) of the query fragment, in which case it is set to 1. Its purpose is to ensure that only a distribution that is an exact match of the query fragment can have a similarity of 1.

It is important that no atom or bond information is used in similarity configuration files that is not explicitly or implicitly contained in the key values used for searching the Mogul tree. Only if this condition is fulfilled will all members of a given Mogul distribution have the same similarity value. In view of the discontinuous nature of the similarity coefficient and to distinguish it from more familiar constructs such as the Tanimoto coefficient, we refer to the quantity as a *relevance* rating in the Mogul interface. Our justification for such an ad hoc formalism is simply that it has been shown in practice to deliver satisfactory results.

Similarity Coefficient Tuning. The similarity configuration files were optimized against small test sets of bond, valence-angle, and torsion query fragments. For each, the CSD search program ConQuest[13] was used to find a selection of "hits" such as might have been found in a Mogul generalized search (maximum R-factor allowed = 7.5%). The similarity of each "hit" to its corresponding query was judged manually (we first established that experienced chemists show a high degree of concordance in such judgments). The configuration files were iteratively refined to produce good correlations between calculated similarity coefficients and the manually judged similarity estimates. In the case of bonds and valence angles, we then determined, for each "hit", the absolute value $|obs - mean|$, where *obs* is the bond length or angle in the "hit", as taken from the CSD structure in which it occurred, and *mean* is the average value obtained from an *exact* Mogul search for the corresponding query. These data were used to establish that there were only a small percentage (<4%) of "hits" with high ($\geq$ 0.7) similarities and large $|obs - mean|$ values (where large is > 0.03Å for bonds, > 3° for angles). The configuration files thus determined were used without further change in the full validation described in the Results section.

DATA COVERAGE

Cyclic torsions and fragments involving metal and hydrogen atoms are excluded from the libraries. Linear or near-linear torsions (W−X−Y−Z where one or both of the W−X−Y and X−Y−Z valence angles is close to 180°) are excluded since they are numerically poorly defined. Apart from that, Mogul contains every crystallographically independent bond, angle, and torsion in the CSD (and will be updated as the CSD grows) except for distributions corresponding to very common types of fragments. These are reduced to 10 000 arbitrarily chosen observations (though a better policy may have been to select 10 000 observations from a subset of the most precise structures in the CSD). This was done because of the impact on search speeds of downloading distributions containing, in some cases, over half a million observations, and in the belief that 10 000 observations are more than enough for most if not all practical applications of Mogul. Data are included irrespective of the experimental quality of the crystal structures whence they came. However, filtering of search results on crystallographic R-factor is possible.

Only the absolute values of torsion angles are stored. This is because a large majority of structures in the CSD either crystallize in non-Sohnke space groups (groups with mirror planes or inversion centers, hence containing both a molecule and its mirror image in the unit cell) or in Sohnke groups but with absolute configuration or absolute structure not determined. In the first case, the sign of any given torsion angle is effectively arbitrary; in the second case it is unreliable. However, by setting all torsion angles to their absolute values, we have lost information for the relatively small proportion of structures where the sign *is* meaningful (e.g. many peptide structures). Also, the relative signs of consecutive torsions along chains is lost.

QUERY INPUT AND PREPROCESSING

Graphical and Instruction-File Interfaces. When used via the graphical interface, query molecules (2D or 3D) can be read in a variety of formats or sketched. Searching is performed either by selecting an individual fragment in the query molecule or by invoking an option that searches for all bonds and/or valence angles and/or acyclic torsions. When used via the instruction-file interface, the query molecule file is submitted to Mogul together with an instruction file specifying which fragments are to be searched for, search control parameters, etc. If the input structure is a polymer, it is necessary to include more than one monomer unit to ensure that the chemical environment of each independent fragment is fully defined.

Bond Type Assignment and Standardization; Hydrogen Addition. The results of a Mogul search will be erroneous if the query does not have correct bond types or is missing hydrogen atoms or is in a tautomeric form not seen in the CSD. These factors are important, since many structures read into Mogul are likely to be freshly determined crystal structures in formats that do not carry bond-type information; they may also lack hydrogens. Further, "correct" bond types means not only chemically correct but obeying the arbitrary conventions of the CSD.[14] For example, conjugated bidentate metal ligands such as acetylacetonato are coded in the CSD with *delocalized* bonds. Mogul therefore incorporates software for detecting where the bonds are in an input structure, guessing the bond types, applying CSD bond-type conventions, and detecting and adding missing hydrogens. Since this cannot be done with complete reliability, manual structure-editing facilities are also available.

The bond-typing algorithm is particularly complex, involving many empirically derived heuristics based on bond-length distributions determined from the CSD. Key steps are as folows:

1. Pi-bonds between metal atoms and poly-hapto-bound ligands are identified by looking for the characteristic triangle of bonds between a metal and two nonmetallic atoms.

2. A geometry-based guess is made at the hybridization states of nonmetallic atoms. This cannot be done completely reliably since the input structure may be poorly refined and lacking hydrogens. Those hybridization states that can be safely inferred are used in several of the subsequent heuristics.

3. Bonds connecting sequences of nonmetallic atoms that are all pi-bonded to the same metal ion but do not form a closed ring (e.g. a metal-coordinated butadiene) are assigned the delocalized bond type.

4. Easy bond-type assignments are made. These include the following: bonds between metal atoms and oxygen, sulfur, and phosphorus atoms, whose types are normally obvious from bond-length considerations; easily recognized, common nonconjugated bonds, such as single bonds to terminal halogens.

5. Common groups such as azide, carboxylate, and sulfonamide, and some less common groups such as seleninyl and selenonyl, are recognized by their patterns of bonded atoms and geometries. They are assigned appropriate bond types.

6. The structure is examined to see if any bond types can now be deduced from valency arguments. For example, if a carbon atom thought to be sp^2 hybridized has already had two of its bonds assigned as single, the remaining one must be double. This procedure is repeated after several of the other steps in the algorithm, too.

7. It may now be possible to assign some other bond types with confidence. For example, if a 4-coordinate selenium has not been found to be part of an $(O{=})Se({=}O)X_2$ group in step 5, all the bonds it forms may safely be assumed single.

8. Obvious aromatic rings are identified and the ring bonds given the aromatic bond type.

9. Bonds between CC or CN atom pairs that are pi-bonded to a metal are assigned as double or triple, depending on their geometries. This step is rather imperfect as there is a large overlap in the observed geometries of these systems.

10. The geometries of carbon atoms that are ostensibly two-coordinate, but bent, are examined to determine whether the atoms are likely to be sp^3 hybridized (and therefore, by implication, bonded to two hydrogen atoms that are missing from the input structure). If so, the bonds in which the atoms are involved are assigned as single bonds.

11. The most complex part of the algorithm then deals with parts of the molecule that appear to be conjugated patterns of single and double bonds. Each bond in such a system is assessed to determine whether it is more likely to be single or double, based on its bond length and the relative frequency with which these bond types occur for that particular atom pair. (It is hard to exploit valence-angle information because so many conjugated systems involve cyclic systems where ring-closure constraints have a dominant influence.) Those bonds whose types can thus be guessed with most confidence are tentatively assigned, and a complete pattern of alternating double and single bonds built up, using valency-based arguments (see step 6) where possible. The final assignment of bond types to the complete conjugated system is then examined and rejected, e.g., if it has led to a highly unlikely valency for one of the atoms. In this case, the algorithm backtracks to the point at which the valency error was introduced and reverses the bond-type assignment that caused the problem. Successive iterations of this procedure are performed until a satisfactory set of bond types is found or until no further options remain to be explored, in which case the least egregious solution is accepted.

12. Bond types are set for metal−nitrogen and metal−carbon bonds, both of which are difficult. No attempt is made to identify metal−metal multiple bonds: they are all assigned the single bond type.

13. A final search is undertaken for any nonconjugated multiple bonds that may have been missed. This uses a Bayesian classifier that takes into account both the bond-length distributions of single, double, and triple bonds between the various pairs of nonmetallic elements, and the relative frequencies with which these bonds occur in the CSD.

14. Any remaining, unset bonds are assigned the single bond type.

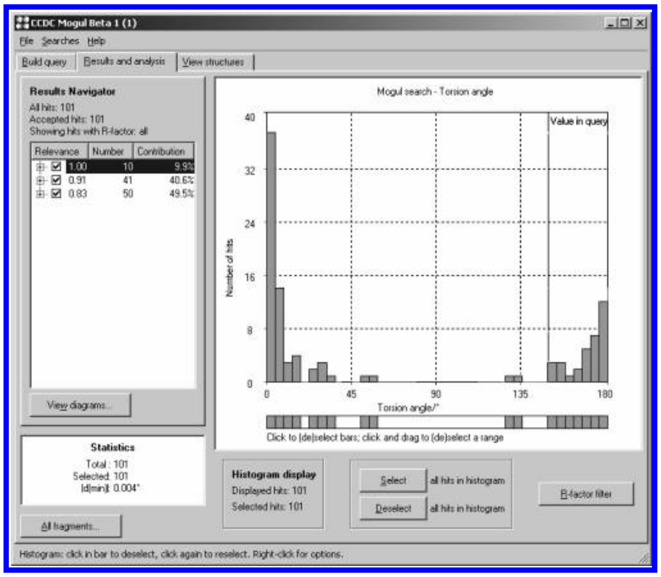
Once all bonds types have been assigned, they are standardized to CSD conventions (e.g. bidentate acetylacetonato ligands are given the delocalized bond type). Unfortunately, the conventions are not always rigorously adhered to in the CSD itself, so some chemical groups have to be set to the bond-type arrangement most commonly used in the CSD. The final step is to check for missing hydrogens using an algorithm taken from the ConQuest program.

Validation of Bond-Typing Algorithm. The algorithm was tested on a published set of 91 molecules taken from CSD crystal structures and used by other workers to test earlier algorithms.[15] (Five CSD reference codes listed in ref 15 do not exist in the CSD and are presumably erroneous or superseded. In each case, a suitable replacement reference code[16] can be identified, viz. AMOXCT10, not AMOXCT; ANFLCN10, not ANFLCN; PILLBA10, not PILLBL10; PRMESA10, not PRMESA; TBUCBD01, not TYBUCBD01). The algorithm was deemed successful on a molecule if all bonds were assigned the same bond types as in the CSD, or, occasionally (AMDMCN, OXTETK), if the bond types were different but still constituted a reasonable valence-bond representation. Our algorithm assigned correct bond types to 88 of the molecules, failing on 3. This success rate (97%) is the same as Hendlich et al. obtained with their program BALI[15] and higher than those achieved by the programs of Meng and Lewis and Baber and Hodgkin (80% and 82%, respectively).[17,18] One of our failures (BEVJER10) was also a BALI failure and results from assigning to a squarate ion bond types that correspond to an uncharged molecule (a difficult type of error to avoid). Our other two failures were OTETCB (assigned as the -one rather than the enol form) and PROMYC10 (a picrate ion with poor geometry).

Although the success rate is good, the test set is undemanding, consisting as it does of purely organic molecules (7 of the molecules are from structures containing metal ions, but it is clear from ref 15 that these ions were removed for the BALI tests). We therefore performed a more challenging validation on 1104 structures taken randomly from the CSD. These include many difficult metallo-organic compounds. In contrast to the earlier test, the algorithm was only deemed successful if it assigned correct bond types to all bonds in all crystallographically independent molecules and ions of the crystal structure (the earlier test set consisted of individual molecules rather than the complete contents of the asymmetric unit). Again, a success was counted if the algorithm produced a reasonable valence-bond assignment even if it differed from the CSD bond types (given the large size of the test set, we will inevitably have made occasional misjudgments and, in particular, may sometimes have been generous in our judgments on structures containing redox-active ligands). Successful results were obtained for 954 (86.4%) of the 1104 structures (Table S1, Supporting Information). The 150 failures were due to a variety of causes, e.g. the algorithm usually fails on porphyrins and on solvates with very poor geometries. Potential solutions (e.g. the use of templates) can be seen for many of the failures, and we anticipate that a success rate of slightly over 90% will eventually be achieved. 1−2% of structures are insoluble, no matter how much the algorithm is enhanced. These include structures that have missing hydrogen atoms and short C−C bond lengths in, for example, the middle of long chains. Such bonds could correspond either to multiple bonds or librationally shortened single bonds and the atomic-coordinate data are insufficient to resolve the ambiguity. (Whatever the bond type is set to, the length of the bond is likely to fall in a sparsely populated region of the resulting Mogul distribution.)

PROGRAM OUTPUT

Using the Mogul graphical interface to search, e.g., for all the bonds of a molecule results in a spreadsheet containing summary statistics for each bond (mean, sample standard deviation, etc.) derived from the experimentally determined geometries of the matching fragments found in the CSD. Selection of an individual bond in the spreadsheet causes further details of the search to be displayed, including a histogram, and details of matching CSD structures. When the input structure is 3D, a *z*-score is given for each bond. This is the deviation between the bond length in the query and the mean of the bond lengths of the hits found by the Mogul search, normalized by sample standard deviation.



Results Navigator
All hits: 101
Accepted hits: 101
Showing hits with R-factor: all

Relevance	Number	Contribution
☑ 1.00	10	9.9%
☑ 0.91	41	40.6%
☑ 0.83	50	49.5%

Click to (de)select bars; click and drag to (de)select a range

Statistics
Total : 101
Selected: 101
|d(min)|: 0.004°

Histogram display
Displayed hits: 101
Selected hits: 101

Select all hits in histogram
Deselect all hits in histogram

Histogram: click in bar to deselect, click again to reselect. Right-click for options.

Figure 4. Torsion-angle histogram as presented in the Mogul graphical interface.

Valence angle results are presented similarly. For torsions, the search results can be displayed as a histogram (e.g. Figure 4), but most of the summary statistics are not listed as they are meaningless for multimodal distributions. In place of a *z*-score, the value d_{min} is given, being the difference between the absolute torsion angle in the query and the nearest observation to it among the absolute torsion angles of the hits from the Mogul search. Summary statistics and histograms can be written out so that they are available to client applications.

RESULTS

The quality of results from Mogul must ultimately be judged by their ability to predict the dimensions of query molecules. Two questions are pertinent: are results from Mogul searches unbiased and are they sufficiently precise?

Statistical Considerations. Ideally, the mean of (for example) a Mogul bond-length distribution should be an unbiased estimate of the bond length of the corresponding query fragment. (Here, and for the remainder of the article, the word “distribution” is now used to refer to all the observations found by a Mogul search, whether that search was exact or generalized.) This will be so if the Mogul key schemes and similarity calculations adequately represent fragment environments, so that CSD fragments that are similar to the query, but differ in some respect that has important and systematic geometrical consequences, are not found as hits. For bond lengths and angles, this can be investigated by examining the Δ values obtained when Mogul searches are performed on a test set of query fragments whose dimensions have been determined experimentally to good precision. Δ is defined as

$$\Delta = |obs - mean| \qquad (2)$$

where *obs* is the value of the bond length or valence angle in the query fragment and *mean* is the average of the corresponding distribution found by Mogul. The Δ statistic is inappropriate for torsion angles, for which d_{min} (see above) must be used.

Table 5. Summary of Validation Results[a]

	precision		accuracy	
search type	range	number	range	number
bonds, exact	$\sigma \le 0.02$	25	$\Delta \le 0.02$	25
	$0.02 < \sigma \le 0.04$	8	$0.02 < \Delta \le 0.04$	6
	$0.04 < \sigma$	2	$0.04 < \Delta$	4
bonds, generalized	$\sigma \le 0.02$	15	$\Delta \le 0.02$	14
	$0.02 < \sigma \le 0.04$	3	$0.02 < \Delta \le 0.04$	4
	$0.04 < \sigma$	2	$0.04 < \Delta$	2
bonds, exact, in comparison of exact versus generalized	$\sigma \le 0.02$	44	$\Delta \le 0.02$	45
	$0.02 < \sigma \le 0.04$	3	$0.02 < \Delta \le 0.04$	5
	$0.04 < \sigma$	3	$0.04 < \Delta$	0
bonds, generalized, in comparison of exact versus generalized	$\sigma \le 0.02$	39	$\Delta \le 0.02$	45
	$0.02 < \sigma \le 0.04$	8	$0.02 < \Delta \le 0.04$	5
	$0.04 < \sigma$	3	$0.04 < \Delta$	0
angles, exact	$\sigma \le 2$	36	$\Delta \le 2$	41
	$2 < \sigma \le 3$	7	$2 < \Delta \le 3$	3
	$3 < \sigma$	5	$3 < \Delta$	4
angles, generalized	$\sigma \le 2$	8	$\Delta \le 2$	12
	$2 < \sigma \le 3$	9	$2 < \Delta \le 3$	3
	$3 < \sigma$	3	$3 < \Delta$	5
angles, exact, in comparison of exact versus generalized	$\sigma \le 2$	37	$\Delta \le 2$	36
	$2 < \sigma \le 3$	8	$2 < \Delta \le 3$	9
	$3 < \sigma$	5	$3 < \Delta$	5
angles, generalized, in comparison of exact versus generalized	$\sigma \le 2$	29	$\Delta \le 2$	36
	$2 < \sigma \le 3$	16	$2 < \Delta \le 3$	8
	$3 < \sigma$	5	$3 < \Delta$	6
torsion, generalized			$d_{min} \le 1$	36
			$1 < d_{min} \le 3$	9
			$3 < d_{min}$	5

[a] Units are Å for bonds, ° for angles, torsions; table gives number of test fragments falling in given range, e.g. in the exact bond searches, 25 of the 35 (= 25 + 8 + 2) test fragments had $\sigma < 0.02$ Å.

The precision with which Mogul can predict fragment dimensions will be determined by the dispersion of the geometry distributions it produces. This, in turn, will partly depend on whether an exact or generalized search has been performed. The total sample variance of a bond length or valence angle distribution, σ^2, arises from three sources: experimental errors in the individual observations; variations in the intramolecular environments of the individual fragments; and variations in their crystal-field (i.e. intermolecular) environments:[19]

$$\sigma^2 = \sigma^2(experimental) + \sigma^2(intramolecular) + \sigma^2(intermolecular) \quad (3)$$

As bond lengths and valence angles are relatively hard parameters, σ^2*(intermolecular)* can be presumed negligible. Since generalizing a search causes fragments to be taken from a wider range of substructural environments than in an exact search, it will tend to increase the value of σ^2*(intramolecular)*. This increase will be modest if the similarity calculation is effective at allowing through only fragments that are closely related to the query. This can be tested by comparing the σ values of distributions found by exact and generalized searches.

Test Set. A random selection was made of good-quality structures (crystallographic R-factor < 5%) that were added to the CSD between November 2003 and January 2004, inclusive. The Mogul data libraries were built from a version of the CSD (viz. 5.25) that preceded this period. Hence, none of the test structures was used in the creation of the Mogul libraries. Random bond, angle, and torsion fragments were chosen from the test molecules and used to investigate the quality of Mogul results. Results of the investigation are summarized in Table 5 and discussed below. All searches were performed with Mogul Version 1.0.

Bond Lengths. Three experiments were performed, one looking at the results of exact searches, one at results from generalized searches, and one comparing exact and generalized results. For the former, three structures were taken from the test set. One (AHONAM) is a rhenium triphenylphosphine derivative and also contains an acetone solvate molecule. The second (AHOPIW) is organic but contains second-row atoms (phosphorus, sulfur). The third (BASKAI) contains only carbon, nitrogen, oxygen, and hydrogen. Exact searches were performed for all bonds in the structures. Filtering was applied (here, and in all other bond-length and valence-angle searches described below) so that only hits with R-factors below 5% were accepted. If two different query bonds found the same Mogul distribution (a frequent occurrence—molecules often contain topologically equivalent bonds) only the first was used in the analysis. Exact searches finding fewer than 5 hits (after R-factor filtering) were rejected. The results (σ, Δ and number of hits, N) from the 35 surviving searches were collated (Table S2, Supporting Information).

Next, 20 bonds from test-set structures were found that returned no hits when an exact search was performed. The results of generalized searches on these bonds were collated (Table S3, Supporting Information). A similarity threshold of 0.75 was used, i.e., no hit was accepted whose similarity to the query was below this value (the same threshold was used in all other generalized searches described below).

Finally, 50 bonds were found that produced a moderate number of hits (between 9 and 35) when exact searches were performed. The results of the exact searches were collated, together with results from generalized searches on the same bonds (Table S4, Supporting Information). Parameters for the generalized searches were chosen so that the number of hits obtained was at least twice that obtained in the corresponding exact search.

Both the precision and accuracy of the exact searches, as judged by σ and Δ respectively, are generally satisfactory (Tables 5, S2). The average value of σ is 0.020 Å, which is only slightly higher than typical standard deviations reported in manually compiled bond-length tabulations,[9] where each distribution was "hand-crafted", e.g. by removal of outliers and laborious refinement of search substructures. The average Δ is satisfactory at 0.018 Å. Both for σ and Δ, 25 of the 35 values fall below 0.020 Å. There are, however, some severe outliers for bonds in AHONAM, viz. σ for C26−C27 (0.069 Å), σ and Δ for O3−C27 (0.064 Å and 0.112 Å, respectively), and Δ for C10−C11 (0.085 Å). The former two are bonds in the acetone solvate molecule, which is both poorly refined in the query structure and in many of the hits contributing to the Mogul distribution. The latter has an unrealistic value in the query structure. The implicit problem is that use of only an R-factor filter is sometimes inadequate. Structures containing heavy metals may have good R-factors yet poorly refined light-atom positions, especially for atoms belonging to thermally mobile or disordered solvate molecules.

The accuracy and precision of the generalized searches (Tables 5, S3) compare well with those of exact searches. Thus, both σ and Δ have average values of 0.018 Å; 15 of the 20 σ values and 14 of the Δ values fall below 0.020 Å. The two Δ outliers (JAFSOZ N1−N2, 0.065 Å; IHIYED B1−F1, 0.059 Å) both occur in query structures that contain heavy metals where, again, R-factor may be a rather poor guide to experimental precision, particularly for an atom as light as boron.

The effect of generalization can best be seen by the paired comparisons (Tables 5, S4). Because of the R-factor filter, σ^2*(experimental)* in eq 3 may be assumed small, at least in most cases. Thus, a trend for generalized searches to produce appreciably larger values of σ^2*(intramolecular)* than exact searches should be apparent in the σ values. In fact, any such trend is small. Average σ is 0.016 Å for exact searches and 0.019 Å for generalized; 29 exact searches have σ values at least 0.001 Å smaller than the corresponding generalized searches, but there are 10 search pairs where the opposite is the case. Comparison of Δ values gives similar results. We conclude that the similarity calculation performs well in finding fragments that, while not an exact match for the query, are close enough to be good for bond-length prediction.

In summary, pooling the 35 searches in Table S2, the 50 exact searches in Table S4 and the 20 searches in Table S3, some 62 (59%) have $\Delta \leq 0.010$ Å, 84 (80%) have $\Delta \leq 0.020$ Å, and 94 (90%) have $\Delta \leq 0.030$ Å. The four searches delivering $\Delta > 0.050$ Å can all be ascribed to poor query geometries and/or inclusion in the Mogul libraries of poorly refined solvate fragments (which has an obvious implication for future versions of Mogul).

Valence Angles. The accuracy and precision of valence angle searches were investigated using an analogous protocol to that described above for bond lengths. Results are given in Tables S5−S7 (Supporting Information). Average values of σ and Δ for the 48 exact searches (Tables 5, S5) are 1.8° and 1.2°, respectively. 36 σ values and 41 Δ values fall below 2°. Large outliers ($\sigma = 6.3, 6.4°$, $\Delta = 8.3°$) occur for two angles of the acetone solvate molecule in AHONAM and may be ascribed to the same cause as the corresponding bond-length outliers discussed earlier. Other σ values exceeding 3° occur for C10−N1−C13 in AHOPIW and C18−C16−C19 and C20−C17−C21 in BASKAI. In each case, the σ value is large because of a single outlier in the Mogul distribution. Two are due to disordered groups in otherwise well-refined CSD entries; the other occurs in a CSD structure containing a polyiodide chain, where light-atom positions may be of relatively low precision despite an R-factor below 5%.

Average σ and Δ values for the generalized searches (Tables 5, S6) are somewhat higher than those from the exact searches at 2.3° and 2.8°, respectively; 8 of the 20 σ values and 12 of the Δ values fall below 2°. The most noteworthy results are huge Δ outliers of 8.6° (C1−B1−C25 in BATCUV) and 16.2° (C13−C10−C17 in AHOTIA), together with the somewhat less outrageous Δ value of 4.8° for C4−C5−C10 of BASZUR. In BATCUV, the query is an unusual $(C_6F_5)_3B.CO.R$ species. ConQuest searches of the CSD for systems related to this, such as $(C_6F_5)_4B^-$, show that the angles at boron are extremely variable and bimodally distributed. In BATCUV itself, the three topologically equivalent C(ar)−B−C(=O) angles are 98.3, 108.0, and 116.9°. We conclude that the high inherent variability of the boron valence angles, coupled with the low precision with which boron atoms may often be located in X-ray structures, make BATCUV a pathological example, not just for Mogul but for other geometry-prediction algorithms.

The AHOTIA query angle is of the type $CH_3-C=C$, where the C=C bond is pi-bonded to osmium and part of a fused cyclobutene ring. Most of the Mogul hits are pi-bonded tetramethylcyclobutadiene metal complexes. It is unclear whether the high Δ value is due to experimental inaccuracy in the query structure or whether it indicates that the hits found by Mogul are insufficiently chemically similar. If the latter, this is the first example discussed so far that indicates deficiencies in the way that Mogul represents fragment chemical environments. It is worth noting, however, that the generalized search producing this outlier required backtracking further up the tree than the default settings of the program allow. Thus, anyone using the program with default settings would not get the distribution found in our work; rather, they would fail to get any hits at all.

The last outlier, in BASZUR, definitely exposes weaknesses in Mogul's chemical-fragment environment representation. The query structure is a 9,10-ethano-9,10-dihydroanthracene derivative (Figure 5). The experimentally measured value of the query angle is 113.3°. There are three other angles in the structure that are topologically equivalent, with measured values of 112.8, 112.8, and 112.9°. Thus, there can be little doubt that the experimental value for the query is reasonably accurate, and the large discrepancy between it and the mean (118.0°) of the Mogul distribution is a result of the generalized Mogul search retrieving inappropriate hits; in particular, some of the hits are not bridged ring systems. If the R-factor filter is removed and an exact search done, a Mogul distribution containing 8 observations is found, and the mean (112.9°) is now satisfactorily close to the experimental value in the query. This establishes that the poor performance of the original search is a consequence of the generalization and hence points to a deficiency in the similarity calculation. In passing, we have found other examples where an exact search finding a very small number

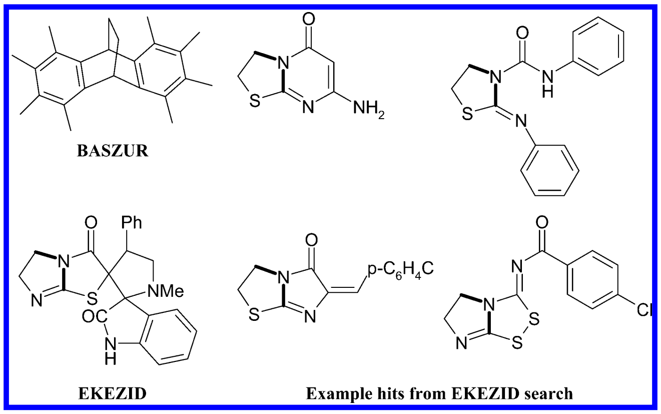


Figure 5. The major molecular components in the CSD entries BASZUR and EKEZID (left); some hits (middle, right) from a Mogul search for the angle shown in bold in EKEZID.

of observations (<4) produces a better mean valence angle than a corresponding generalized search. This is sometimes because the query molecule contains a bulky group that distorts nearby valence angles; the similarity calculation has little scope for recognizing steric bulk.

The paired comparisons of exact and generalized valence-angle searches (Tables 5, S7) show that generalization lowers accuracy and precision but not by much. Average σ values are 1.9° and 2.1°, and average Δ values are 1.6° and 1.7°, for exact and generalized searches, respectively. In 32 cases, σ from the exact search is at least 0.1° lower than σ from the corresponding generalized search; 10 search pairs show the opposite trend. Corresponding figures for Δ are 25 and 20, respectively. Overall, we conclude that the valence-angle similarity calculation is performing well. There are, however, a small number of gross outliers. They occur for the following: N1–C1–O1 of JAFKOR (exact $\sigma = 11.2°$, generalized $\sigma = 7.7°$); O1–C11–O2 of EKERIV (exact $\sigma = 8.5°$, generalized $\sigma = 5.8°$); C2–N1–C4 of EKEZID (exact $\Delta = 6.5°$, generalized $\Delta = 5.8°$); C19–B1–C25 of BATCUV (exact $\Delta = 6.0°$, generalized $\sigma = 5.7°$, generalized $\Delta = 5.3°$); and C6–C7–C8 of IHOQUR (generalized $\Delta = 5.1°$). The problems in JAFKOR and EKERIV are again due to extreme outliers in the Mogul distributions, attributable to poorly refined, possibly disordered groups in CSD structures. BATCUV has already been discussed above. The IHOQUR outlier represents another failure of the similarity calculation: some of the hits are highly bridged ring systems—which the query molecule is not—whose valence angles are distorted by ring-closure constraints. The exact search finds only a small number of bridged-ring molecules and gives a much better mean valence angle.

The EKEZID outlier is significant because it provides the only clear example in our survey of an *exact* Mogul search failing to find hits that are adequate models for the query. The structure of EKEZID is shown in Figure 5. Some of the Mogul hits are also shown. The hit that is arguably the most chemically similar to EKEZID (bottom right) is the only hit that has a similar valence angle (108.6°; the query angle is the same). It appears that the extent of the chemical-environment information held in the Mogul valence-angle key scheme is simply inadequate in this case.

In summary, pooling the 48 searches in Table S5, the 50 exact searches in Table S7 and the 20 searches in Table S6, some 58 (49%) have $\Delta \leq 1°$, 89 (75%) have $\Delta \leq 2°$ and 104 (88%) have $\Delta \leq 3°$. Two of the six searches delivering $\Delta > 4°$ can be ascribed, at least in part, to deficiencies in the way Mogul represents the chemical environments of angles, either in the key scheme or the similarity calculation. One other may be due to this or to experimental errors in the query structure or CSD entries. Two relate to a borate system that represents an example of extreme difficulty. The remaining one is due to poor experimental geometries, either in the query structure or in CSD entries. We conclude that the Mogul representation of angle environments, although usually adequate, is not quite as reliable as for bonds.

A problem exists for ions such as hexafluorophosphate that give rise to distributions (in this case of F–P–F angles) that are bimodal because some relate to cis and some to trans atoms. Nothing is done in Mogul to deconvolute this type of distribution. Bimodal valence-angle distributions occasionally occur for other reasons, e.g. the distribution of C(ring)–C(ring)–N(=N-) angles in azobenzenes has two peaks, one corresponding to ring carbons that are syn to the azo group, where C...N repulsion tends to increase the angle, and the other corresponding to carbons anti to the azo linkage.

Torsion Angles. Since torsion-angle distributions are generally multimodal, the mean is inadequate as an estimate of central location and both σ and Δ are, in consequence, of no use. Also, torsion angles are highly sensitive to crystal-field environments so σ^2(*experimental*) and σ^2(*intramolecular*) in eq 3 are likely to be swamped by σ^2(*intermolecular*). Thus, the importance of R-factor filtering and the difference between exact and generalized searching are likely to be of far less consequence for torsions than for bonds and valence angles. These factors indicate that a different protocol is necessary to assess the predictive value of Mogul torsion-angle distributions. The procedure used was as follows. Random torsion angles were chosen from test-set structures. A generalized Mogul search was performed on each, no R-factor filter being applied. Each resulting distribution was presented as a histogram, using 40 equally spaced bins over the full 0–180° range. Any test torsion angle that gave rise to a Mogul histogram in which more than half the bins were occupied was rejected. The experimentally observed value of each query torsion angle was then compared with the corresponding Mogul histogram and the following were noted: d_{min}; whether the query torsion angle fell in an occupied bin of the histogram; whether the query torsion angle fell in a peak of the histogram (defined as a histogram bin containing more observations than either of the immediately adjacent bins). Results are collated in Table S8 (Supporting Information). 41 of the 50 query torsion angles fall in occupied bins, although only 19 fall in bins that represent local maxima of the histogram. The average value of d_{min} is 1.1°. There are 36 d_{min} values below 1° and only 5 above 3° (Table 5). Of the 9 query torsion angles that do not fall in occupied bins, all but two have d_{min} values ≤ 4°.

If the torsion distributions were approximately uniform, these statistics would not be impressive. For example, given a distribution of 30 torsion angles evenly spaced through the range 0–180°, the maximum possible value of d_{min} for any query torsion angle would be only 2.9°. However, the distributions used in this experiment are very far from uniform: on average, they have only 35% of bars occupied, corresponding to almost two-thirds of the 0–180° range being empty. At random, therefore, we would only expect about 16 or 17 of the query torsions to fall in occupied bins. Random expectation values of d_{min} cannot be estimated easily but would certainly far exceed most if not all of the observed values. The results are therefore encouraging and suggest that Mogul torsion distributions will be of considerable value in algorithms for generating low-energy molecular conformations. For example, such an algorithm might make effective use of Mogul by allowing only torsion angles that lie within occupied regions of the corresponding Mogul distribution, permitting 4° of leeway either side of each such region.

CONCLUSIONS

Mogul allows the CSD to be interrogated easily for bond-length, valence-angle, and torsion-angle distributions and statistics. The starting point is a molecule, not a substructure, and all bonds, angles, and acyclic torsions, excluding those involving metal or hydrogen atoms, can be searched for by a single instruction. The software can be invoked by a third-party program via a text file interface. Validation results suggest that averages of Mogul bond-length and valence-angle distributions provide unbiased estimates of the corresponding dimensions of the query molecule in a large majority of cases and to good precision (typically ≤ 0.02 Å for bonds and $\leq 2°$ for valence angles). In over two hundred test searches, only one clear example was found where the Mogul key scheme inadequately represented the chemical environment of a molecular fragment (specifically, a valence angle); two or three cases were found where the similarity calculation allowed through hit structures that were not good models for the query fragment (again, valence angles). Searches on a random sample of fifty torsion fragments suggested that Mogul torsion-angle distributions are likely to be reliable and useful guides to molecular conformational preferences.

There is scope for improving the filtering options in Mogul. Only filtering on R-factor is possible at present and this is sometimes inadequate; for example, it may allow through heavy-metal-containing structures where light-atom positions are poorly determined or structures containing disordered groups or solvate molecules. Nonetheless, it is comparatively unusual that these have significant effects on summary statistics from Mogul. Further limitations in the program, to be addressed in future versions, include the following: conformational preferences of ring systems are not included; results cannot be obtained for fragments involving metal atoms; and while individual torsion-angle histograms are available, correlations between the values of the torsion angles around adjacent rotatable bonds (Ramachandran plots) are not. These limitations notwithstanding, we believe that Mogul will make CSD data much more accessible and therefore enhance their usefulness. For example, we have already shown that Mogul can be successfully interfaced to the CRYSTALS structure solution package[20] where it can be used both for validating newly determined crystal structures and for furnishing average molecular dimensions for use in restrained least-squares refinement. Mogul will be included in future releases of the CSD software system.

ACKNOWLEDGMENT

Ms. Clare Macrae and Dr. Greg Shields are thanked for advice and software contributions.

Supporting Information Available: List of CSD reference codes of successes and failures of bond-typing algorithm (Table S1); results from exact (Table S2) and generalized (Table S3) searches on bond fragments; comparison of exact and generalized searches on bond fragments (Table S4); results from exact (Table S5) and generalized (Table S6) searches on valence angle fragments; comparison of exact and generalized searches on valence angle fragments (Table S7); and results of searches on torsion fragments (Table S8). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

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

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

(3) Feuston, M. P.; Miller, M. D.; Culberson, J. C.; Nachbar, R. B.; Kearsley, S. K. Comparison of knowledge-based and distance-geometry approaches for generation of molecular conformations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 754−763.

(4) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.

(5) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.

(6) Lemmen, C.; Lengauer, T. Time-efficient flexible superposition of medium-sized molecules. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 357−368.

(7) Allen, F. H.; Harris, S. E.; Taylor, R. Comparison of conformer distributions in the crystalline state with conformational energies calculated by ab initio techniques. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 247−254.

(8) Boehm, H.-J.; Klebe, G. What can we learn from molecular recognition in protein−ligand complexes for the design of new drugs? *Angew. Chem., Int. Ed. Engl.* **1996**, *35*, 2588−2614.

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

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

(11) Redman, J.; Willett, P.; Allen, F. H.; Taylor, R. A citation analysis of the Cambridge Crystallographic Data Centre. *J. Appl. Crystallogr.* **2001**, *34*, 375−380.

(12) Willett, P. *Similarity and clustering methods in chemical information systems*; Research Studies Press: Letchworth, 1987.

(13) Bruno, I. J.; Cole, J. C.; Edgington, P. R.; Kessler, M.; Macrae, C. F.; McCabe, P.; Pearson, J.; Taylor, R. New software for searching the Cambridge Structural Database and visualizing crystal structures. *Acta Crystallogr., Sect. B* **2002**, *58*, 389−397.

(14) ConQuest 1.6 User Guide. Cambridge Crystallographic Data Centre, 2003.

(15) Hendlich, M.; Rippmann, F.; Barnickel, G. BALI: Automatic assignment of bond and atom types for protein ligands in the Brookhaven Protein Databank. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 774−778.

(16) CSD reference codes (six upper-case letters followed sometimes by two digits) are unique identifiers of crystal structures in the CSD.

(17) Meng, E. C.; Lewis, R. A. Determination of molecular topology and atomic hybridisation states from heavy atom coordinates. *J. Chem. Comput.* **1991**, *12*, 891−898.

(18) Baber, J. C.; Hodgkin, E. E. Automatic assignment of chemical connectivity in organic molecules in the Cambridge Structure Database. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 401−406.

(19) Taylor, R.; Kennard, O. The estimation of average molecular dimensions from crystallographic data. *Acta Crystallogr., Sect. B* **1983**, *39*, 517−525.

(20) Betteridge, P. W.; Carruthers, J. R.; Cooper, R. I.; Prout, K.; Watkin, D. J. CRYSTALS version 12: software for guided crystal structure analysis. *J. Appl. Crystallogr.* **2003**, *36*, 1487.

CI049780B