TUTORIAL REVIEW

CSR
www.rsc.org/csr

Research applications of the Cambridge Structural Database (CSD)

Frank H. Allen and Robin Taylor
Cambridge Crystallographic Data Centre (CCDC), 12 Union Road, Cambridge, UK CB2 1EZ. E-mail: allen@ccdc.cam.ac.uk; taylor@ccdc.cam.ac.uk

Received 5th April 2004
First published as an Advance Article on the web 10th September 2004

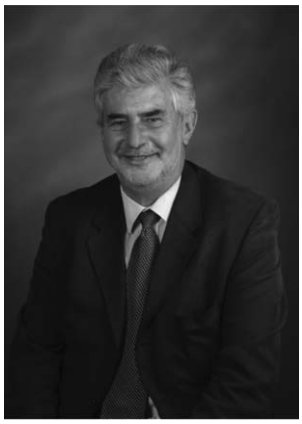
Crystal structure data are of fundamental importance in a wide spectrum of scientific activities. This *tutorial review* summarises the principal application areas, so far, for the data from more than 300,000 crystal structures of small organic and metal-organic compounds that are stored in the Cambridge Structural Database (CSD). Direct use of the accumulated data is valuable in establishing standard molecular dimensions, determining conformational preferences and in the study of intermolecular interactions, all of which are crucial in structural chemistry and rational drug design. More recently, information derived from the CSD has been used to construct two dynamic libraries of structural knowledge: Mogul, which stores intramolecular information, and IsoStar, which stores information about intermolecular interactions. These electronic libraries provide information "at the touch of a button". In their turn, the libraries also serve as sources of structural knowledge for applications software that address specific problems in small-molecule and biological chemistry.

1 Introduction

Crystal structure analyses are remarkable for the richness of information that they provide. Because this information yields both the geometric structure of a molecule and also characterises the nature and geometry of its interactions with other molecules and ions, crystal structure data are crucially important to a very wide range of scientific activities. Examples include: (a) structural and supramolecular chemistry, (b) conformational analysis and the prediction of protein–ligand interactions – both vital components of paradigms for rational drug design, and (c) crystal engineering, crystal growth, crystal structure prediction and polymorphism – all of which are important in drug development and materials design. Since the late 1960s, the results of published crystal structure analyses have been collected in five databases which together cover the complete spectrum of chemical compounds.

This review concentrates on the scientific applications of the Cambridge Structural Database (CSD) of small organic and metal-organic molecules,[1,2] and a principal purpose is to illustrate the scientific value of analysing the crystallographic results for many chemical structures or substructures taken together. Techniques for data visualisation and data analysis then permit, for example: the determination of mean values for geometrical parameters, the observation of preferred conformations or coordination sphere geometries, the mapping of their interconversion pathways, and the observation and analysis of the intermolecular interactions that are responsible for molecular aggregation and crystal growth. Thus, we describe (a) how the CSD can be used for basic research in some of the areas listed above, (b) how CSD data can be converted into rapidly accessible electronic libraries of structural knowledge, and (c) how these libraries can, in their turn, be used as knowledge engines that underpin further software applications designed to solve problems in structural chemistry, rational drug design and crystallography.

Frank Allen was born in Reading in 1944 and studied chemistry at Imperial College, London, receiving a BSc in 1965 and a PhD in 1968. Following postdoctoral work at the University of British Columbia, Vancouver, he joined the CCDC in 1970. He has been involved in most major CCDC developments since then, with a strong accent on research applications of the Cambridge Structural Database. He received the RSC Prize for Structural Chemistry in 1994 and the Herman Skolnik Award of the American Chemical Society Division of Chemical Information in 2003. He is now Executive Director of the CCDC and a Visiting Professor of Chemistry at the University of Bristol.



Frank Allen

Robin Taylor was born in Birmingham in 1951 and received a BA in chemistry from the University of Oxford in 1973 and a PhD in chemical crystallography from the University of Cambridge in 1976. Following postdoctoral research at York, Pittsburgh and the CCDC, he joined Zeneca Agrochemicals, becoming Group Leader in molecular modelling. He returned to the CCDC in 1994 as Development Director, and his principal responsibility is to lead the development of new products. His particular interest is the application of the Cambridge Structural Database to computer-aided molecular design.

Robin Taylor

DOI: 10.1039/b309040j

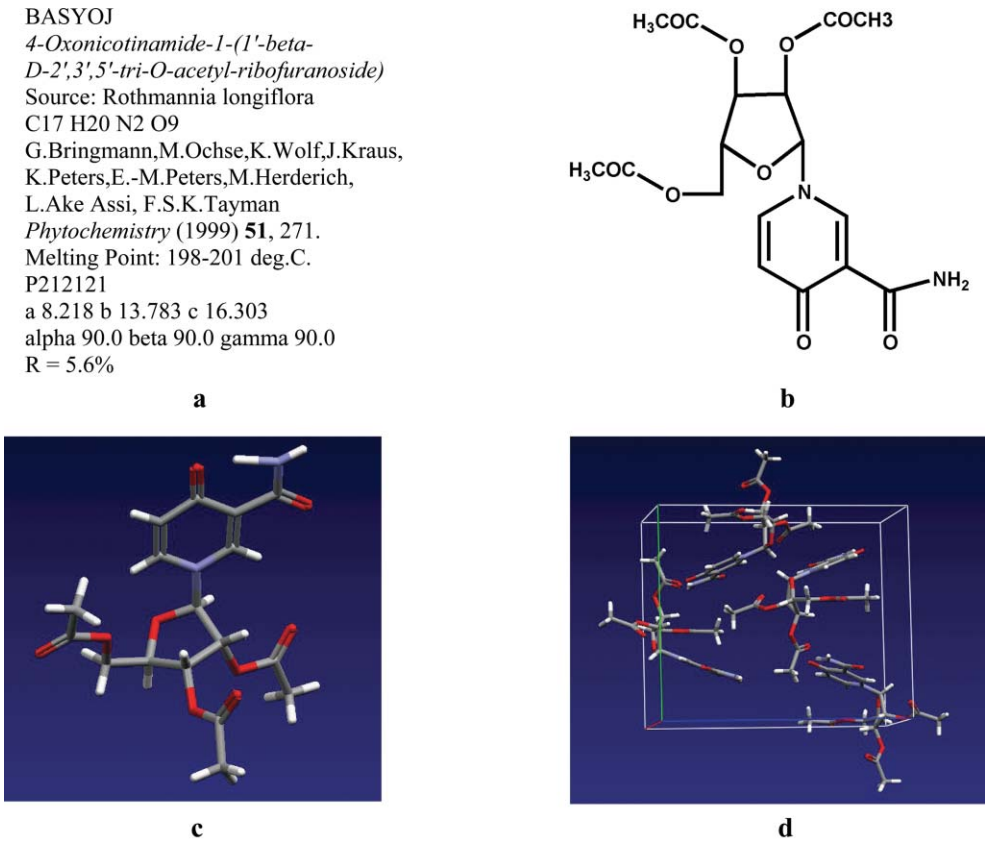


Fig. 1 Schematic view of the information content of a Cambridge Structural Database entry: (a) bibliographic, chemical and crystallographic text, (b) 2D chemical structural formula (chemical connectivity), together with (c) 3D molecular structure, and (d) 3D crystal structure derived from stored atomic coordinates and crystal data.

2 The Cambridge Structural Database (CSD) and the CSD System[1,2]

Compilation of the CSD began in 1965, at a time when less than 1000 structures were published annually worldwide. The aim was to record the experimental results of each analysis: cell dimensions, space group and atomic coordinate data, as well as bibliographic and chemical information. In 2003, the CSD archived its 300,000th structure, and its current growth rate is well in excess of 25,000 structures per year. The information content of a typical database entry (crystal structure) is illustrated in Fig. 1. Some 99% of CSD entries are abstracted from the published literature, with the remainder being deposited as *Private Communications* to the CSD. All data undergo checks for accuracy and internal consistency before being archived to the master database.

The complete CSD System comprises the database itself together with software tools for searching and visualising database entries, and for analysing structural information.

ConQuest searches the CSD *via* queries based on text, chemical and numerical fields. Substructure searches of the 2D chemical structural diagrams (Fig. 1b) are the most important search mechanism. Queries are entered graphically, and can be embellished with 3D geometrical constraints, *e.g.* to locate specific conformations or stereochemistries, or specific pharmacophoric patterns. ConQuest will also search for non-bonded interactions (intermolecular or intramolecular) using geometrical constraints on, *e.g.*, hydrogen-bond geometry *etc*. The software will tabulate user-specified geometrical and crystallographic data for each substructure located in a search. Fig. 2 shows a ConQuest search query that will locate hydrogen bonds involving QA–H donors (QA = N or O) and S atoms in (R_1,R_2)C=S substructures, using numerical criteria to define the limiting H-bond geometry. Geometrical descriptors for each substructure retrieved from the CSD can also be generated, and here the H···S distance and the C=S···H and [N or O]–H···S angles were output.

Mercury provides both general and advanced functionality for viewing 3D molecular and crystal structures, including the display of chemical bond types on 3D images (see *e.g.* Fig. 1c). Important features of Mercury are its ability to locate, build and display networks of interactions (Fig. 3a), *e.g.* hydrogen bonds, short non-bonded contacts or user-defined contact types, and to display slices through crystals as an aid to the rationalisation of crystal morphology and crystal growth (Fig. 3b).

Vista creates a spreadsheet of geometrical and crystallographic data generated by ConQuest. Vista will generate histograms and scattergrams (Cartesian and polar) for parameter distributions, and provides simple and advanced statistical functionality for data analysis. Fig. 4 shows (a) the histogram of H···S distances and (b) the scatterplot of H···S distance *vs.* [N or O]–H···S angle for the hits resulting from the ConQuest search of Fig. 2. These results confirm the H-bond acceptor ability of the thione-S (see Section 5.1 below) and show the commonly observed linear relationship between bond length (bond strength) and H-bond directionality at the donor-H.

3 Research applications, statistical methods and structure correlation

3.1 Overview, reviews and the CCDC WebCite database

The CSD System has been used for basic research since the first versions became available in the mid-1970s. These early research activities demonstrated the importance of statistical and graphical methods in the analysis of large volumes of numerical (geometrical) data,[3] and gave rise to the principle of

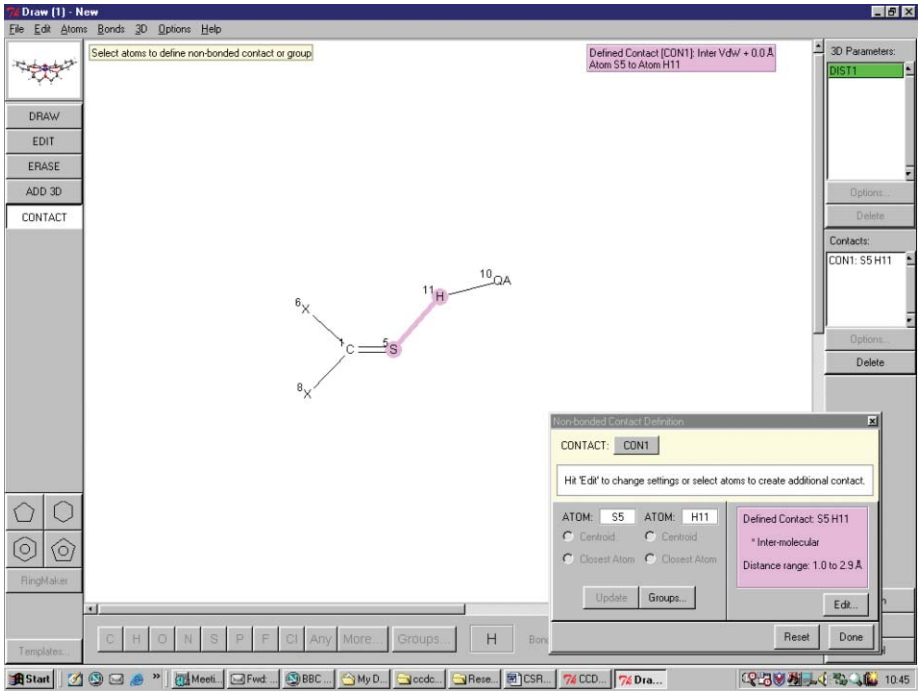


Fig. 2 Screenshot of ConQuest search query to locate C=S···H–QA [QA = N or O] hydrogen bonds within an S···H distance limit of 2.9 Å. Further graphical input permits the definition of other geometrical parameters for output or use as search constraints, as noted in the text.

structure correlation, enunciated by Bürgi and Dunitz.[4,5] Since that time, nearly 1000 CSD-based research papers have appeared in the literature, together with reviews of specific research areas and a number of monographs in which results obtained from the CSD have played a major role. Amongst these reviews, the two-volume book *Structure Correlation*[5] of 1994 is the most comprehensive. Other more recent material is cited in Refs. 6, 7 and 8. At the same time, the CCDC has maintained its own bibliographic database of CSD-based research publications, including brief synopses of each application. This database is kept as comprehensive as possible and searchable access is freely provided *via* the WebCite section of the CCDC Website at www.ccdc.cam.ac.uk.

3.2 Statistical and graphical methods[3]

A number of these techniques are embodied in the Vista program noted above, or are available in many external software packages. Of particular importance are: (a) descriptive statistics (mean, median and standard deviation) for a distribution of, *e.g.*, a specific bond length observed in many crystal structures; (b) parametric and non-parametric tests to assess the significance of differences between means; (c) the use of covariance, correlation and regression to determine the extent and nature of any relationship between pairs of parameters; and (d) multivariate methods, such as principal components analysis and cluster analysis. The latter are appropriate for structural problems which require the analysis of three or more parameters for each substructure retrieved from the CSD, *e.g.* the analysis of conformational preferences of *n*-membered rings, where each conformer is described by *n* torsion angles, or the analysis of metal coordination spheres, where each sphere is characterised by the $[n(n-1)]/2$ L–M–L valence angles in an ML_n species.

3.3 The principle of structure correlation

During the late 1970s and 1980s, Dunitz, Bürgi and co-workers enunciated the principle of structure–structure correlation[4,5] which underpinned their classic studies of reaction pathways in a series of aminoketones located using the CSD. In the Bürgi–Dunitz hypothesis, the static distortions exhibited by a specific molecular fragment in a wide variety of crystalline environments are assumed to map the distortions that the fragment would undergo along a reaction or interconversion pathway, *i.e.* the various static fragments are considered to form a series of structural 'snapshots' along the pathway, and the observed structures tend to concentrate in low lying regions of the potential energy hypersurface. This important principle is exemplified and illustrated elsewhere in this review.

4 Intramolecular aspects of CSD-based research

4.1 Mean molecular dimensions

Structural information derived from crystal structure data has played the key role in the systematisation of structural chemistry since the pioneering work of Pauling in the 1930s. A simple and obvious use of CSD data has been to generate mean values for standard geometrical parameters, such as bond lengths and valence angles, to act as benchmarks against which new data may be compared, or to act as restraints during the refinement of novel structures. Two major compilations of mean bond lengths[9] were generated during the late 1980s for organic molecules,[9a] and for metal-organic complexes of the *d*- and *f*-block metals.[9b] Bond length distributions for more than 1000 chemical bond types were generated from the CSD, outliers were examined, and the distributions were characterised *via* the descriptive statistics reviewed in Section 3.2. Another notable compilation is that of Engh and Huber,[10] who derived mean bond lengths and valence angles for peptidic structures in the CSD, basing their classification on 31 C, N, O atom types that are most appropriate to the protein environment. These data continue to be used extensively in the determination, refinement and validation of novel protein crystal structures, and are built into many key computer programs in structural biology.

4.2 Conformational analysis

The generation of torsion angle distributions to determine conformational preferences about single rotatable bonds, or

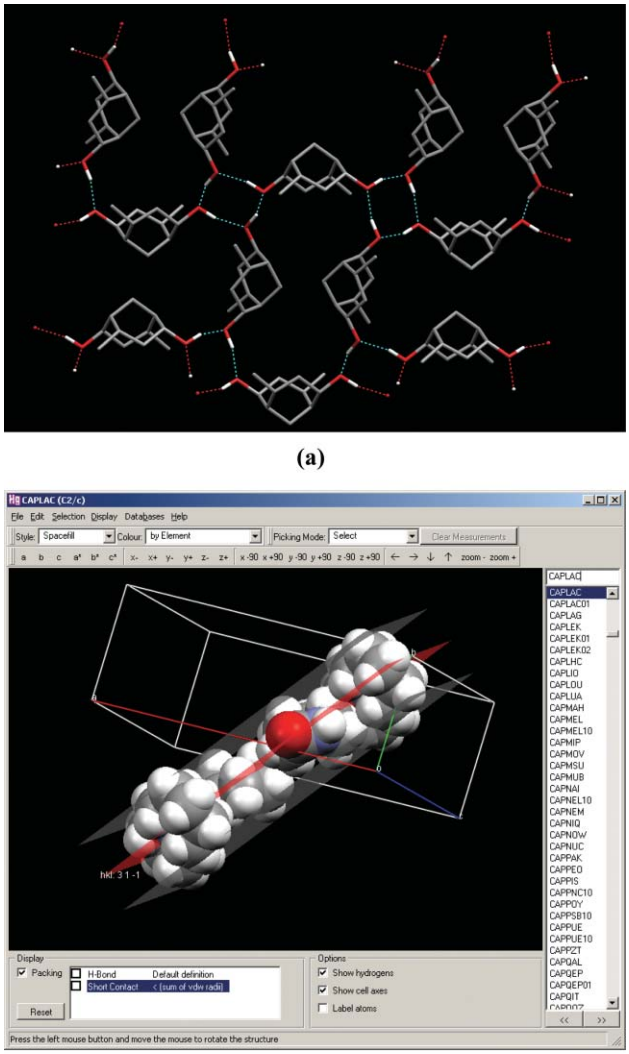


Fig. 3 Mercury plots of (a) an extended hydrogen bonded network, and (b) a slice through a crystal structure.

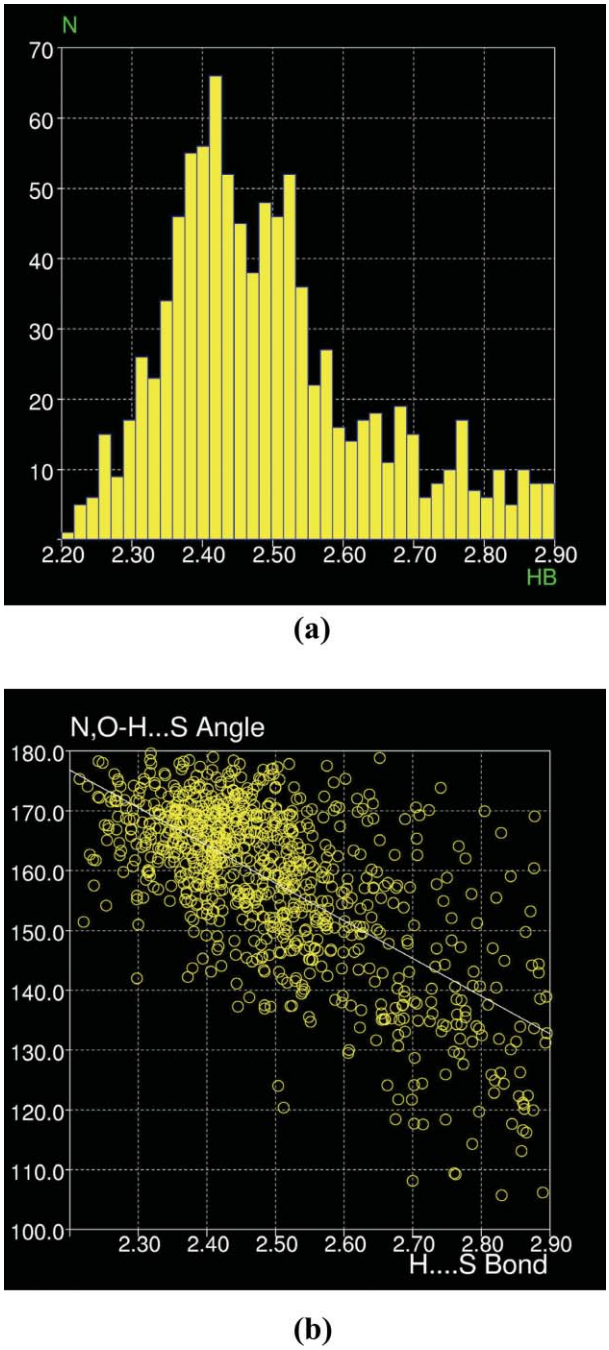


Fig. 4 Vista plots of geometrical data retrieved from the CSD by the ConQuest query of Fig. 2: (a) histogram of the S···H distance, and (b) scatterplot of the S···H distance *vs.* the [N or O]–H···S angle.

for complete ring systems, is one of the most common applications of the CSD, particularly in molecular modelling[6] and in structure determination from powder diffraction data.[11] Conformations adopted in individual crystal structures will be affected by different crystal field environments, and cannot be assumed to represent either the global minimum energy geometry or the geometry adopted when a molecule binds to a protein. However, if a specific molecular substructure containing a rotatable bond is observed in a series of crystal structures, it is likely that the more strained higher energy conformations will be observed less often than relatively unstrained lower energy geometries, in accordance with the structure correlation principle. Thus, the observed distribution of torsion angles around a rotatable bond should reflect the potential energy curve for rotation about that bond.

This hypothesis was tested[12] for twelve common substructural fragments by comparing torsion angle distributions from the CSD with those obtained from *ab initio* molecular orbital calculations. Each substructure was able to adopt two conformers, *anti* and *gauche*, and the qualitative complementarity of the experimental and calculated profiles was striking, with the natural logarithm of the relative frequencies of the two conformers in crystal structures being linearly related to their *ab initio* calculated energy differences. While systematic packing effects can cause substantial deviations from this Boltzmann-like result,[6] the overall conclusion[12] was that (a) torsion angles with strain energies of >1 kcal mol^{-1} are rarely observed in crystal structures, and (b) crystal structure conformations are indeed good guides to conformational preferences in solution. Indeed, Taylor[6] presents arguments based on CSD observations which indicate that crystal structure conformations are better guides to *in vivo* solution conformations than those derived from *in vacuo ab initio* calculations. This conclusion is borne out by a study of the conformations of synthetic ligands[13] in the CSD and in protein-bound complexes retrieved from the Protein Data Bank.

Further evidence supporting the basic tenet of the structure correlation principle – that crystal conformations tend to cluster in low energy regions of the potential energy hypersurface – arises from conformational studies on a variety of multivariate

systems. Thus, the CSD has been used[14] to study the conformational variations in benzophenones by generating a scatterplot (Fig. 5) of the two torsion angles (TOR1 and

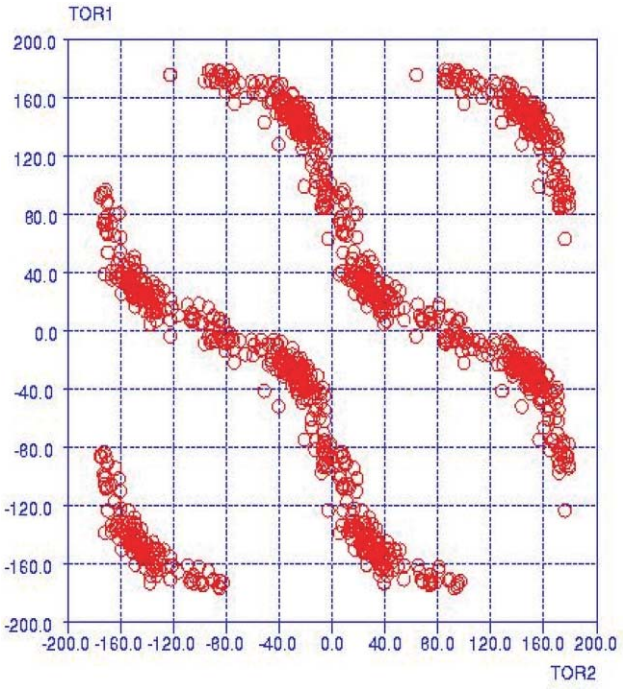


Fig. 5 Symmetry-expanded Ramachandran-like plot (after Ref. 14) of the O=C–C_{ar}–C_{ar} torsion angles in benzophenone substructures retrieved from the CSD using ConQuest.

TOR2) that quantify the conformations of the two independent phenyl rings with respect to the >C=O group. In the original paper,[14] Fig. 5 is superimposed on the contoured calculated potential energy hypersurface which has energy minima at TOR1, TOR2 = +30, −30° and its symmetry equivalents, and where these global minima are connected by low energy valleys that correspond to the conformational interconversion pathways depicted so clearly in Fig. 5. A considerable body of work, reviewed elsewhere,[7] has also been carried out on the detection of conformational preferences for ring systems using the multivariate analysis techniques of principal components analysis and cluster analysis. Again, there is close complementarity between the conformational mappings and clusterings generated from crystal structure data and those generated by minimum energy calculations.

4.3 Structure–property relationships

From the mid-1980s, the principle of structure–structure correlation was extended to structure–property relationships, through the work of Kirby and others on structure–reactivity correlations. These crystallographic approaches to transition state structure have been extensively reviewed by Kirby,[15] and are exemplified by the study of C–O bond length variations in a series of axial tetrahydropyranyl acetals (I, Fig. 6), in which C–OR bond breaking should be highly dependent on the orientations of the O lone pairs: optimum nσ-σ* (C–OR) overlap stabilises both the ground state axial conformation and the oxocarbocation (II, Fig. 6). This orbital overlap also leads to hydrolysis of these compounds, and given the high degree of stereochemical control (I → II, Fig. 6), a relationship was sought between the length of the ground state C–OR bond and the rate at which it was broken. Since chemical evidence had shown that the pK_a of the conjugate acid of the leaving group

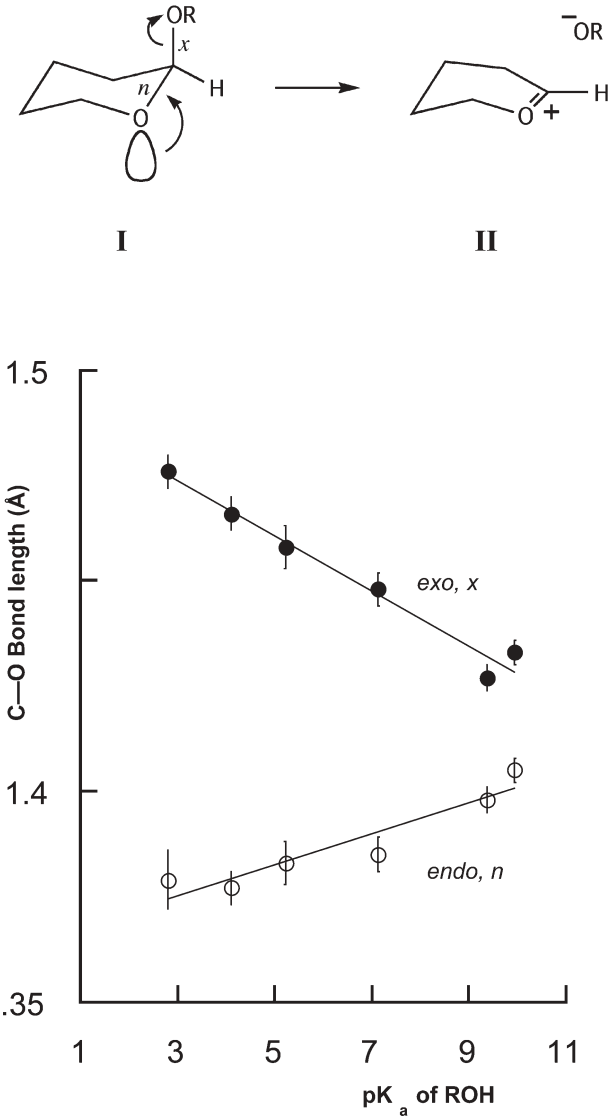


Fig. 6 Relationship between the bond lengths of the endocyclic and exocyclic C–O bonds at the acetal centres of axial tetrahydropyran acetals (I), and the pK_a of the conjugate acid (ROH) of the leaving group (II).

(ROH) was related to reactivity, plots were made of the *exo* and *endo* C–O bond lengths in a variety of axial tetrahydropyranyls. Fig. 6 shows the respective negative and positive linear correlations, which show an increasing divergence between the two C–O distances for the better leaving groups, as might be expected from the reaction.

4.4 Metal coordination sphere geometries and their interconversions

Crystallography is the method of choice for the characterisation of novel metal-organic species, and over 50% of CSD entries contain a transition metal. The CSD therefore contains a wealth of data that are relevant to systematic studies in molecular inorganic chemistry, and these have recently been reviewed.[8] A number of these applications, *e.g.* the determination of typical molecular dimensions, conformational analyses, and studies of reaction pathways and intermolecular interactions, have their organic parallels. However, studies of metal–ligand bonding, catalytic systems, secondary bonding, and aurophilic and agostic interactions are specifically metal-organic in nature.

A natural focus of attention is the geometry of metal coordination spheres, and of their interconversions from one archetypal geometry to another. An example is provided by a recent study of 3-coordinate transition metal (Tr) species,[16] in which the TrL_3 sphere was characterised using the three L–Tr–L valence angles. Principal components analysis showed that the vast majority of such species had a trigonal planar (*tp*) geometry, with some deviations towards a Y-shaped geometry (one angle significantly smaller than 120° due to small-ring formation involving a pair of ligand atoms). More rare was a T-shaped geometry with two angles of *ca.* 90° and one close to 180°, and it was observed that many of these geometries occurred in Hg(II) complexes, as illustrated in the principal components plot of Fig. 7a. This plot also shows a number of data points (TrL_3 substructures) which connect the cluster of T-shaped geometries with the central *tp*-cluster. Given the predominant linear 2-coordination adopted by Hg(II), it is tempting to speculate that the T-shaped cluster of Fig. 7a, together with the intermediate geometries that link this cluster with the *tp*-cluster, represents snapshots along the minimum energy pathway for addition of ligand L_3 to a 2-coordinate L_1–Hg–L_2 centre. This pathway is indeed realised in Fig. 7b by plotting DNORM against AOPP. DNORM is the normalised Hg–L_3 distance, *i.e.* d(Hg–L_3) – *rad*(Hg) – *rad*(L_3), in which the *rad* are covalent radii (the normalisation is necessary since we do not know *a priori* the chemical identity of the generic ligand atom L_3 located in the CSD search). AOPP is the valence angle L_1–Hg–L_2 which is opposite to the point of attachment of the incoming ligand L_3. As expected, AOPP is seen to decrease from 180° (T-shape) to 120° (*tp*) as DNORM decreases from the longer Hg–L_3 bond distances exhibited in the T-shaped species.

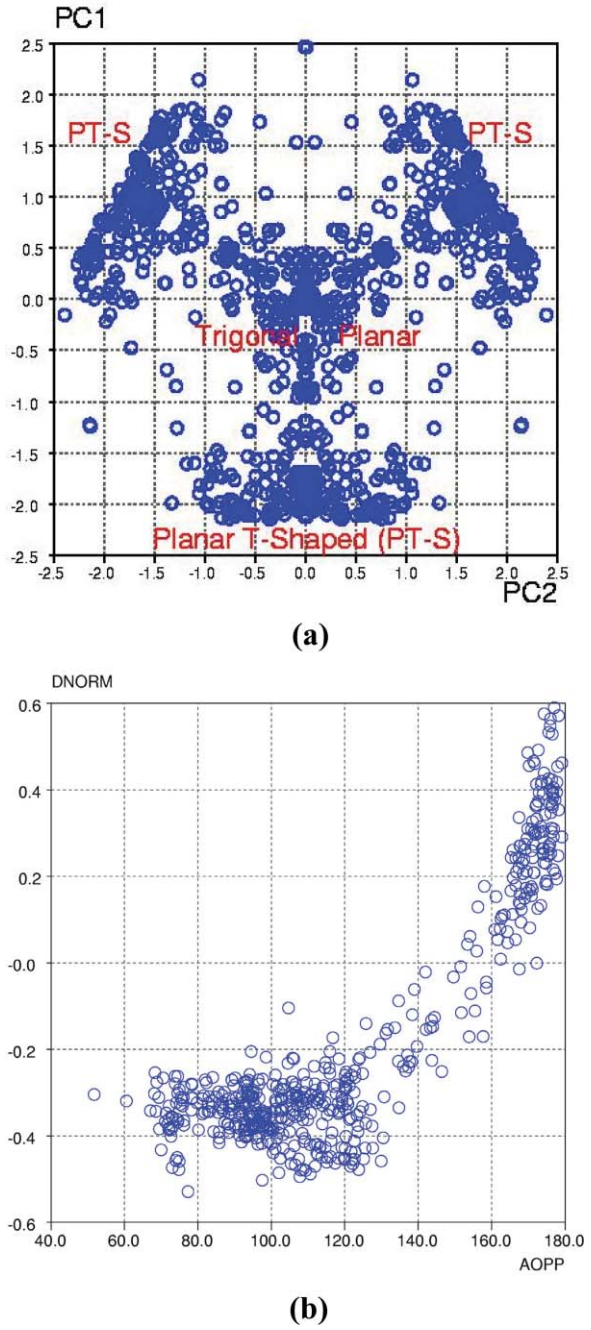


Fig. 7 Analysis of 3-coordinate Hg(II) complexes: (a) principal components map based on the three L–Hg(II)–L valence angles, and (b) reaction pathway for addition of L_3 to 2-coordinate L_1–Hg(II)–L_2 species. The parameters DNORM and AOPP are defined in Section 4.4.

5 Intermolecular aspects of CSD-based research

A crystal structure is the archetypal supermolecule and crystal structure analysis is the only experimental technique that routinely permits direct observation of the intermolecular interactions that control the formation of supramolecular entities. Thus, the technique reveals the types of interactions that occur, their geometrical characteristics and their directional preferences. Such knowledge makes vital contributions to our understanding and development of, *e.g.*, supramolecular synthesis, crystal engineering, protein–ligand interactions, crystal growth, crystal structure prediction and, of course, in the solution and validation of novel crystal structures.

The CSD is therefore a major source of knowledge on intermolecular interactions of all types. The CSD software (ConQuest) permits graphical encoding (Fig. 2) of search queries that involve chemical substructures linked *via* non-bonded 'connections' that are defined in terms of distances (in Å, or relative to sums of van der Waals radii). Other geometrical limitations, if known, may also be used in query definitions. However, crystal structure data can only provide rather general information about the relative strengths of non-bonded interactions, and it is now common to combine CSD studies with calculations of interaction energies carried out using a variety of *ab initio* methods. Here, CSD analyses are used to identify highly populated regions of interaction space, and the (often computer-intensive) computational exploration of the potential energy hypersurface is then restricted to these regions.

5.1 Hydrogen bonding

As documented elsewhere[6,7] the CSD has been used extensively in studies of strong hydrogen bonds, particularly those having N or O as donors and acceptors. Apart from the determination of H-bond distances and angles, studies have examined (a) H-bond lone-pair directionality at the acceptor, (b) resonance-assisted and resonance-induced H-bonds, (c) intramolecular H-bonds, (d) competition effects in systems having a number of acceptors and donors, (e) H-bonded patterns and their relative probabilities of formation, and (f) the role of H-bonds in polymorphic systems.

CSD analysis is exemplified by a study[17] of the resonance-induced hydrogen bonding of N–H or O–H donors to sulfur acceptors in $(R_1R_2)C{=}S$ systems. The >C=S bond is not a natural dipole due to the almost equal electronegativities of C and S, by contrast to the situation in >C=O bonds where the electronegativity of O makes it a strong acceptor. Nevertheless, the structure of thiourea is dominated by N–H···S=C(R_1R_2) bonding. The CSD analysis[17] of all [N or O]–H···S=C(R_1R_2)

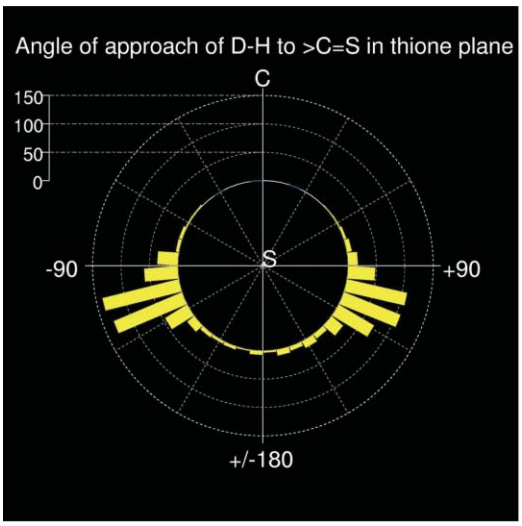
substructures showed that only those systems in which one or both of R_1, R_2 were electron-donating substituents, *e.g.* the amine groups of thiourea, formed H-bonds. Here, the effective electronegativity of S is significantly increased by resonance effects (III, Fig. 8), so that it now becomes an effective acceptor, and the geometrical distributions shown in Fig. 4a,b show typical H-bonding behaviour. Importantly also, further analysis revealed (a) a significant preference for the H-donor to approach the S acceptor in the >C=S plane, and at a C=S···H angle (~105°, Fig. 8a) which clearly shows S-lone pair directionality; and (b) an interaction energy of −20 kJ mol^{-1} computed using intermolecular perturbation theory (IMPT)[18] with an O–H donor in the >C=S plane at *d*(S···H) = 2.40 Å and with a C=S···H angle of 95° (Fig. 8b). This value is somewhat less attractive than interaction energies computed for >C=O···H–O bonds (*ca.* −28 kJ mol^{-1}).

One of the major contributions of CSD analysis to H-bond research has been to establish the existence of a wide range of weaker hydrogen bonds[19] involving: (a) weak donors and strong acceptors, *e.g.* C–H···O, C–H···N *etc.*; (b) strong donors and weak acceptors, *e.g.* O,N–H···Cl, O,N–H···π; and (c) weak donors and weak acceptors, *e.g.* C–H···Cl, C–H···π. Of particular importance was the clear identification of short C–H···O and C–H···N contacts as true hydrogen bonds in the early 1980s.[20] This much cited paper, based on neutron diffraction studies retrieved from the CSD, finally ended all speculation as to the nature of these short interactions involving acidic C–H hydrogens, and put an end to the 'dark ages'[19] that had existed since the late 1960s in which such interactions had to be described in the literature using the most circumspect (and often contorted) phraseology that did not include the words 'hydrogen bond'!

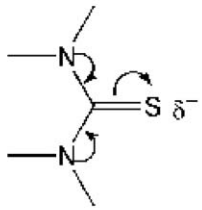
5.2 Interactions not mediated by hydrogen

A review of supramolecular synthons[21] illustrates the structural importance of a wide range of attractive non-bonded interactions that are not mediated by hydrogen, and notes the value of CSD analyses in identifying and characterising these interactions. In practice, the combination of CSD analysis and *ab initio* calculations has again proved valuable, so that the relative robustness of these interactions can be compared with one another and with the more well understood hydrogen bonded interactions.

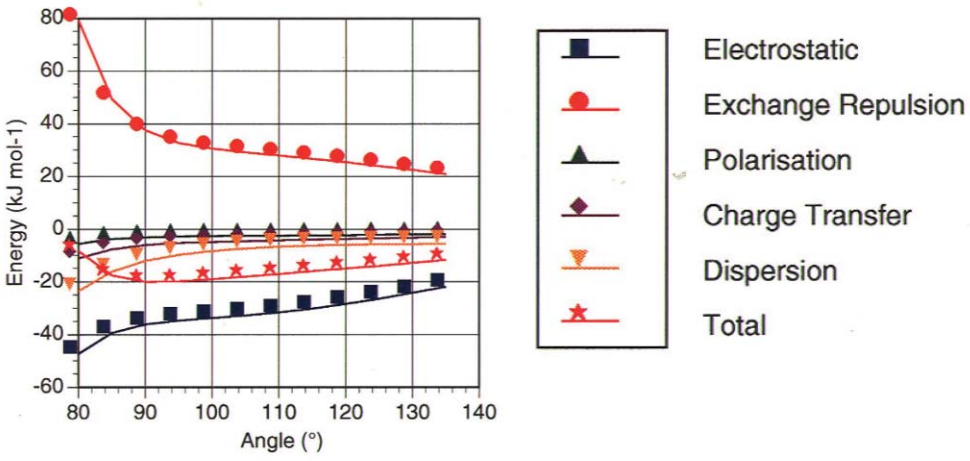
The marked tendency of the halogens X = Cl, Br, I to form short contacts to each other and to electronegative N and O atoms is well known. A combined CSD/IMPT analysis of C–X···O=C< systems[22] showed a marked preference for the



(a)



III



(b)

Fig. 8 Directionality of [N or O]–H···S=C(R_1,R_2) hydrogen bonds at the S-acceptor in thiones: (a) polar histogram of the H···S=C angle, and (b) interaction energies calculated using the IMPT procedure[18] using thiourea and methanol as model molecules, restricting the donor-H to lie in the thione plane, and varying the H···S=C angle.

shortest X···O interactions to form along the extension of the C–X bond and with interaction energies ranging from −7 to −10 kJ mol^{-1} depending on the nature of X and the bonding environment of O=C<. These energies are comparable to the strengths of C–H···O hydrogen bonds. By contrast[23] C–Cl···Cl–C interactions tend to form with C–Cl···Cl angles close to 90°, and these two studies together[22,23] provide a picture of Cl as presenting a quadrupolar electrostatic environment, with an area of electropositive potential pointing outwards along the extension of the C–Cl bond, and an area of electronegative potential in directions perpendicular to the bond. This model is further supported[24] by combined CSD and *ab initio* results for C–Cl···H–N,O hydrogen bonds and their much stronger metal-Cl···H–N,O analogues, which are characterised by H···Cl–C and H···Cl-metal angles of *ca.* 90°–100°.

The importance of group–group interactions has also been highlighted[25a] during a CSD analysis designed to locate isosteric replacements in modelling protein–ligand interactions. A later paper[25b] presented an in-depth study of carbonyl–carbonyl interactions, and showed that dipolar >C=O···O=C< interactions most commonly form in a slightly sheared antiparallel arrangement having interaction energies of about −20 kJ mol^{-1}, comparable to the energies exhibited by medium-strength hydrogen bonds, for example the (N,O)–H···S bonds discussed above. Carbonyl–carbonyl interactions have also been shown to be significant in stabilising certain protein secondary structure motifs[25c] and in stabilising the partially allowed Ramachandran conformations of asparagine and aspartic acid.[25d]

6 Knowledge-based structural libraries derived from the CSD

The undoubted value of crystallographic data does not guarantee their widespread use. The advent of the Internet has revolutionised how people think about information provision, so that "answers at the touch of a button" is now the expectation of the day. For a crystallographic database, this is not so easy. Enthusiasts of the CSD have exploited programs such as ConQuest to great effect, as the above sections illustrate, but these programs cannot be mastered without a certain investment of time and effort: reasonably complex interfaces have to be learnt, search substructures drawn, results subjected to statistical analysis, and so on. The problem is exacerbated by the development of "high throughput" computational chemistry, in which, for example, a molecular modeller engaged in drug discovery might use a protein–ligand docking program to predict the binding conformations and affinities of hundreds of thousands of computer-built molecules. In principle, the CSD contains information to identify and reject any molecule whose docked conformation is physically unrealistic, but it is clearly impossible for the necessary torsion-angle distributions to be generated manually.

These factors have led several groups to derive from the CSD and other crystallographic databases a number of structural libraries that capture key information in a form that is easy and quick to use, either by a human or a client computer program. These libraries may be divided into two types: those providing information about intramolecular parameters and those providing intermolecular data.

6.1 Libraries of intramolecular geometry

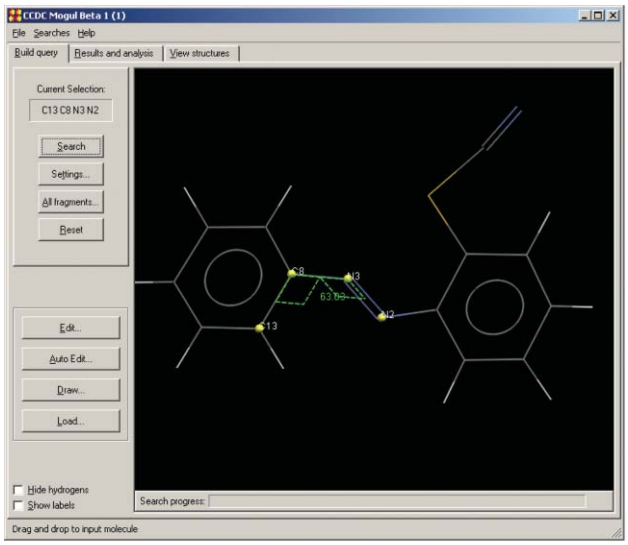
One of the best-known structural libraries to be derived from a crystallographic database is *MIMUMBA*,[26] a collection of 216 torsion-angle distributions obtained by searching the CSD for common molecular fragments, each containing a single rotatable bond. The library can be used in conformational analysis by identifying the rotatable bonds in a molecule of interest and assigning to them torsion-angle values on the basis of a best match to the MIMUMBA torsional distributions (since some of the MIMUMBA molecular fragments are highly generic, it is invariably possible to find a match). This produces a list of possible conformations for the molecule as a whole which can be empirically ranked and subjected to energy minimisation. Part of the procedure involves converting the observed torsional distributions into pseudo-energy curves, a procedure first described by Murray-Rust.[27] MIMUMBA was shown to be successful in finding the experimentally observed conformations of eight out of nine protein-bound molecules (*ligands*) taken from the PDB.

Another library, *et*[28] differs from MIMUMBA in taking better account of correlations between the torsion angles of adjacent rotatable bonds. These correlations are often very strong – witness the large regions of unoccupied space in the classic Ramachandran plot of amino-acid residue (ϕ,ψ) values – and are therefore often effective at identifying combinations of torsion-angle values that are energetically out of reach. *et* contains about 800 substructural fragments, each typically containing from 1 to 3 variable torsions. A subset of about 18,000 diverse organic molecules from the CSD was used to identify the conformational "bins" into which each fragment can fall. For a given bin, information was stored about the average torsion angle of each rotatable bond in the fragment, together with its standard deviation. Conformational analysis for a molecule proceeds by finding all the fragments in the library that match onto the molecule. Of these, the largest fragments are chosen, since they will be the ones that capture best the correlations between adjacent torsions and therefore restrict conformational space most effectively. Once the complete molecule is matched and possible torsion angles assigned, the molecular conformations that remain possible are subjected to bump checking and other tests, leading to a final list of likely geometries. The methodology was tested against 113 molecules whose protein-bound conformations have been determined and deposited in the Protein Data Bank (PDB).[29] When *et* was used to generate 25 conformations for each ligand, a conformation within 1.5 Å RMSD of the observed ligand geometry was found in about 90 of the 113 cases.

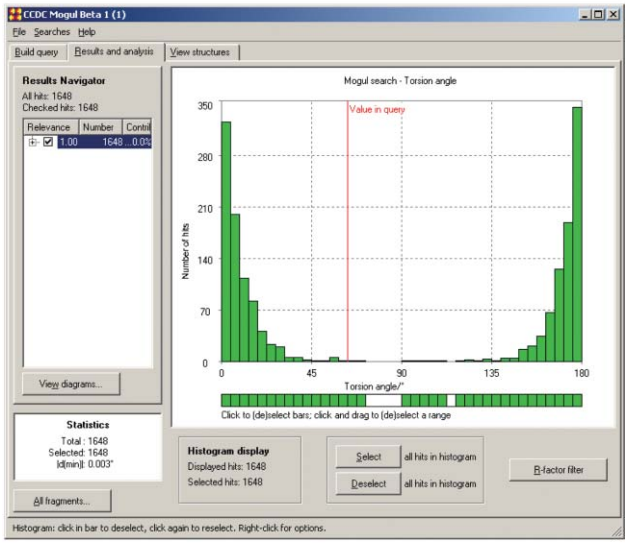
Both MIMUMBA and *et* suffer from the disadvantage that they are based on "snapshots" of the CSD as it was when the libraries were developed. A similar complaint can be laid, of course, against the printed bond-length compilations[9] referred to earlier. With this in mind, the most recently developed component of the distributed CSD System is a molecular geometry library, *Mogul*, that will be continuously updated as the CSD grows. Mogul is able to read a molecule – submitted either manually or by another computer program *via* an instruction-file interface – and *automatically* perform substructure searches of the CSD for all the bonds, angles and acyclic torsions. The crystallographically-determined distributions of the geometries of these fragments are then returned to the user or client program. They may then be used, *e.g.*, to check the dimensions of a computer-generated molecular model.

Rather than doing conventional substructure searching, which involves atom-by-atom and bond-by-bond comparison (graph matching) of the query substructure with database molecules, Mogul works by using "chemical keys". Each fragment (*i.e.* bond, angle or acyclic torsion) in the query molecule is assigned a set of key values that collectively describe the substructural environment of the fragment. For example, two of the key values used to describe the C–C bond between the methyl and carboxylic acid groups of ethanoic acid would be C.4.3 and C.3.0. The first key describes the methyl carbon (a C atom, bonded to 4 other atoms, of which 3 are hydrogens) and the second the carboxylic carbon (a C atom bonded to 3 other atoms, none of which are hydrogens). Other keys would capture the nature of the atoms in the next bonding shell, and

the types of the various bonds between the atoms. A search tree is then traversed to find all fragments in CSD structures that have identical key values; this is approximately equivalent to performing a substructure graph-match but is much faster. Should insufficient hits be obtained, a 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). The experimentally-determined 3D coordinates of the hits – CSD structures containing fragments with the same or similar key values to that of the query – are used to output summary statistics (mean, sample standard deviation, histograms, *etc.*) of the dimension of interest (length if the fragment is a bond; angle if it is a valence angle or torsion). These may be viewed graphically or read by an external application. An example Mogul search display is shown in Fig. 9.



(a)



(b)

Fig. 9 Results of a Mogul search for the C(ring)–C(ring)–N=N torsion fragment in azobenzenes that lack substituents in the *ortho* positions. The search was performed (a) by selecting the four atoms defining the torsion in the query molecule, which came from a crystal structure. The resulting histogram (b) is produced in less than a second and shows that the value of the torsion angle in the query molecule (indicated by the red line) is unusual.

6.2 IsoStar: a library of intermolecular interactions

As mentioned earlier, programs such as ConQuest allow searches to be performed for intermolecular contacts between any given pair of groups A and B. By superimposing the A···B contacts thus found so that the A moieties are overlaid in a least-squares sense, a three-dimensional scatterplot can be produced showing the experimental distribution of B (the 'contact group') around an average A group (the 'central group'). Examples of three such scatterplots are shown in Fig. 10. The *IsoStar* library,[30] compiled and regularly updated by CCDC for the last six years, contains over 25,000 different scatterplots like these, most derived by searching for contacts in CSD structures but a substantial minority based on protein–ligand interactions in the PDB.

An IsoStar scatterplot provides two basic types of information: where the contact group tends to be positioned around the central group and how frequently interactions between the two groups are observed. Fig. 10a, for example, shows that the carboxylate ion in small-molecule crystal structures from the CSD prefers to form hydrogen bonds along the oxygen sp^2 lone-pair directions (and forms many of them). Fig. 10b shows that the same geometrical tendency is seen when ligand carboxylate groups form hydrogen bonds to protein residues in PDB structures. Fig. 10c shows that, of the two oxygen atoms in an ester linkage, the carbonyl oxygen commonly accepts hydrogen bonds but the bridging oxygen almost never does.

Fig. 10c illustrates a common problem with the "raw" scatterplots from IsoStar: there may be so many observations that it becomes difficult to see the wood for the trees. In particular, the large number of O–H···O=C contacts on this plot makes it impossible to assess whether this type of acceptor oxygen atom shows the same lone-pair directionality as observed, for example, with the carboxylate oxygens in Fig. 10a, b. When using the IsoStar interface, the complexity of the image can be reduced somewhat by altering distance limits – for example, showing only those contacts that are much shorter than the sum of the van der Waals radii of the interacting atoms. This often makes it easier to identify directional preferences. A neater answer to the problem, however, involves embedding the plot in a regular three-dimensional grid and counting the number of contact groups falling in each grid cube. Contouring on these counts then permits calculation and display of a surface showing the density distribution of hydroxyl contacts around the central ester group (Fig. 10d). This much simpler representation of the data shows clearly that ester carbonyl oxygens do have the same preference as carboxylate oxygens for forming hydrogen bonds along lone-pair directions. Moreover, it now becomes apparent that one lone pair – *anti* to the linking oxygen – is favoured over the other, presumably for steric reasons.

Importantly, the scatterplot data are hyperlinked to the CSD (or PDB) structures from which the plot was derived. Fig. 11a shows the scatterplot of O–H contacts around the ethynyl group. By clicking on the shortest O–H···π contact, the user is presented with the crystal structure in which that contact was found (CSD entry BETXAZ,[31] Fig. 11b), a structure comprising a tetrameric motif connected by C–H···O and O–H···π hydrogen bonds. Hyperlinking in this way enables users to gain insight into the circumstances in which particular interactions are likely to form. In the present example, the highly hindered nature of the tertiary alcohol almost certainly prevents the formation of the ···OH···OH··· rings or chains that might normally be expected in an alcohol crystal structure. The weaker CH···O and OH···π hydrogen bonds may therefore be presumed to be the best of the options that remain once ···OH···OH··· is excluded – and made more attractive than

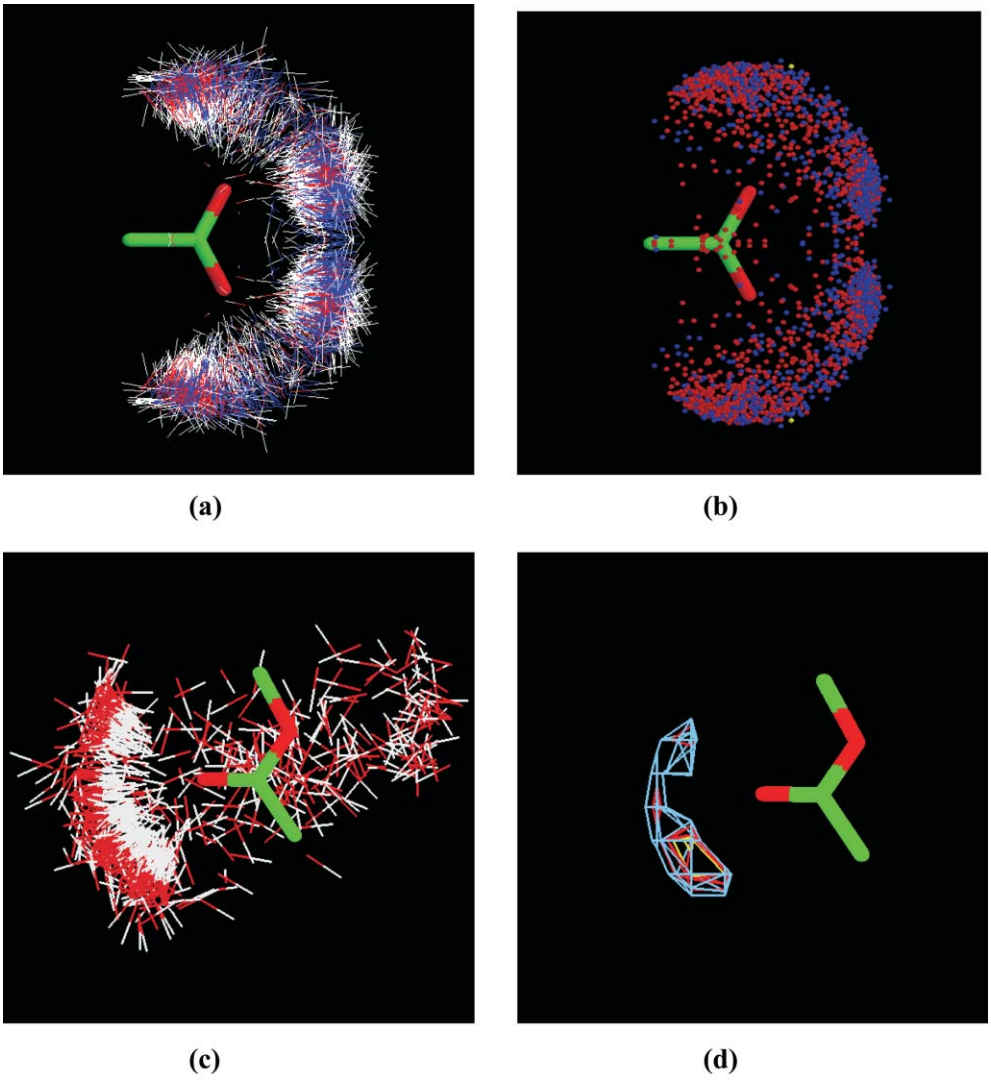


Fig. 10 The IsoStar library of intermolecular interactions: (a,b) distribution of N–H and O–H (donor) contact groups around carboxylate central groups as observed in (a) the CSD and (b) the PDB, and (c,d) the distribution of O–H (donor) contact groups around ester central groups in the CSD presented as a standard scatterplot (c) and as a contoured density distribution (d).

otherwise might have been expected by the relatively acidic nature of the acetylenic CH proton.

The IsoStar interface provides browsing facilities by which users may navigate through the huge array of interactions contained within the library. While all the examples discussed above are hydrogen bonds, the library contains many other

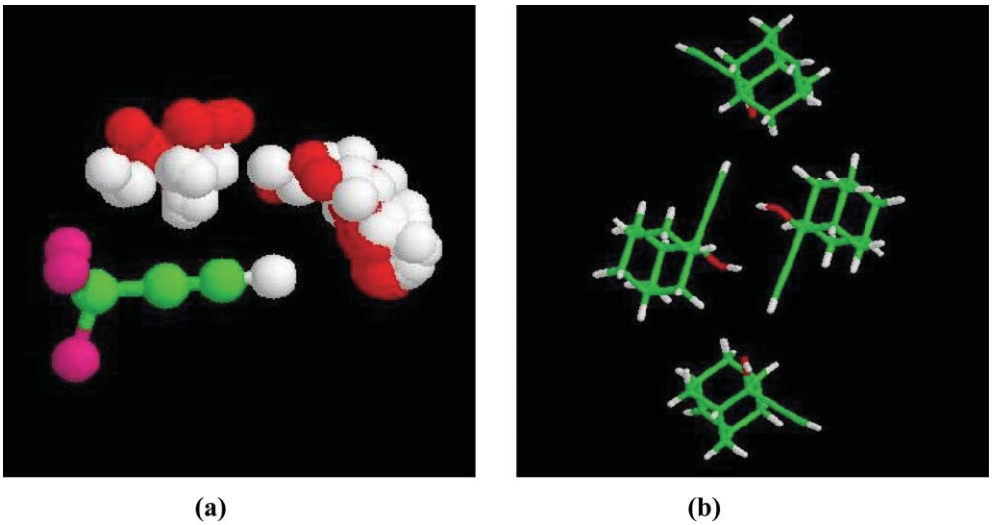


Fig. 11 The IsoStar library of intermolecular interactions: (a) scatterplot of O–H donors around an acetylenic central group (shorter contacts shown), and (b) hyperlinking from IsoStar to the CSD for the shortest O–H···π interaction (CSD Reference Code BETXAZ[31]).

types of interactions, including hydrophobic contacts and attractive electrostatic interactions. The strength of the library is in providing very quick and reliable answers to important but relatively straightforward questions. Here are some typical examples. Do organically-bound fluorine atoms accept hydrogen bonds? Answer: only rarely. Which is the better acceptor in an oxazole ring, the nitrogen or oxygen atom? Answer: the nitrogen. What types of contacts do aromatic sulfur atoms form? Answer: it depends on the heterocycle – for example, thiophene sulfurs show somewhat different preferences to thiadiazole sulfurs. Is there a precedent for a protein tryptophan residue accepting an NH$\cdots\pi$ hydrogen bond from a bound ligand? Answer: yes. All of these questions could also be answered by using a program such as ConQuest; but IsoStar gives answers in seconds rather than minutes.

7 Knowledge-based applications software

So far, this review has focused on the knowledge that may be gleaned from the CSD by a capable chemist or crystallographer sitting in front of a computer terminal using software such as ConQuest, Vista, Mogul or IsoStar. We rely on the CSD to provide the data, the program to analyse and present it in useful ways, and the human to interpret it. The theme of this last section is how crystallographic data can be coupled directly with a client computer program, so that the human user of the client program utilises the crystallographic data at second-hand, without seeing it directly.

7.1 Integration of Mogul into crystallographic software

One of the most obvious uses of Mogul is to check the dimensions of a newly determined or partially refined crystal structure. Significant discrepancies between the observed bond lengths and angles of the new structure and the mean values of the corresponding geometrical distributions in the Mogul library can then be reviewed. They will indicate either experimental errors or genuine differences that are chemically noteworthy. The *CRYSTALS* package for X-ray structure refinement has been modified by its authors to use Mogul in this way.[32] The two programs interact directly with each other, CRYSTALS submitting a crystal structure to Mogul, along with an instruction file. Mogul performs the required searches and returns the results to CRYSTALS. Here, they are converted to *z*-scores, defined in CRYSTALS as $(obs - median)/sd$, where *obs* is the observed value of a dimension in the crystal structure, *median* is the median of the matching Mogul distribution, and *sd* is the sample standard deviation of that distribution. Extreme *z*-scores indicate suspect molecular dimensions. CRYSTALS will also compute a *z*-score for each atom, being the mean of all the bond-length and valence-angle *z*-scores in which that atom is involved. Atoms with high *z*-scores are thus highlighted as having some problem – for example, they might have been assigned an incorrect element type by the crystallographer. The average bond lengths and angles may also be used by CRYSTALS as chemical restraints during further refinement of the structure.

7.2 SuperStar: exploring protein–ligand interactions using IsoStar data

One of the most important uses of crystallographically-derived intermolecular information is in structure-based drug design, *i.e.* the rational design of drugs given the 3D structure of the target protein binding site. This is because most small molecules that bind to proteins do so non-covalently. Hence, an ability to predict the non-covalent interactions that are likely to be favoured in a binding site is essential. Many different approaches have been taken to this problem. A very well-known example is the program *GRID*.[33] This uses an empirical force-field to calculate the energy of interaction between the protein and a probe (a small functional grouping such as methyl or carbonyl) positioned at various points in the binding site. Display of the results as a contoured surface highlights the ''hot-spots'' – regions where the protein–probe interaction is particularly favourable. These, in turn, can be used to hypothesise pharmacophores (3D arrangements of functional groups that should interact well with the binding site and therefore confer binding affinity) for guiding drug design.

Programs like GRID rely heavily on the quality of the empirical energy expressions used to estimate non-bonded interaction energies. An alternative knowledge-based approach is to estimate the probability of an interaction based on how often it has been observed in crystal structures. IsoStar, of course, is an ideal source of this type of information. The program *SuperStar*[34] uses IsoStar to achieve the same end as GRID, but without the need for empirical energy calculations.

Central to the SuperStar approach is a method for placing a density surface such as the one shown in Fig. 10d on a meaningful scale. This can done by considering the parent IsoStar scatterplot (Fig. 10c) and estimating the density of contacts that would be expected in the plot if the spatial distribution of central (ester) and contact (hydroxyl) groups was entirely random in CSD crystal structures containing both groups. The expectation density, d_e, can be computed from:

$$d_e = \Sigma n_i(central) \,.\, n_i(contact)/V_i$$

where the summation is over all the CSD structures containing both the central and contact groups, V_i is the unit-cell volume of the i^{th} such structure, and $n_i(central)$ and $n_i(contact)$ are the number of crystallographically-independent central groups and the total number of contact groups, respectively, in the unit cell of structure *i*. The actual density of contact groups in any region of the scatterplot can be divided by d_e to yield a *propensity* (p). This indicates whether the contact density in that region is greater ($p > 1$) or less ($p < 1$) than would be expected by chance, and by how much. By implication, regions of the plot with $p > 1$ correspond to energetically favourable positions for the contact group around the central group, and the greater the propensity, the more favourable the position is likely to be.

The importance of this scaling process is that it makes it meaningful to *combine* information from different IsoStar scatterplots. This is the foundation of the SuperStar methodology. The program partitions the protein binding site into its constituent chemical groupings, the partitioning been done in such a way that each grouping corresponds to one of the central groups in the IsoStar library. Given a particular probe group, it is thus possible to retrieve the scatterplots of the distributions of the probe groups around all of the chemical groupings present in the protein binding site. Each IsoStar scatterplot is overlaid on all parts of the protein binding site that it matches, converted to a contoured density surface, and then normalised to a user-selected propensity value. The separate surfaces can then be combined into an overall propensity map indicating the regions of the binding cavity most likely to be favourable for the probe group. Where two propensity surfaces from different IsoStar scatterplots overlap they are combined by multiplication. Fig. 12 shows an example SuperStar surface indicating where OH groups are favoured in the binding site of L-arabinose binding protein. The structure is taken from PDB entry 1ABE,[35] which contains a bound L-arabinose ligand. The observed position of the arabinose is shown in Fig. 12 to illustrate that there is a good correspondence between the observed positions of ligand OH groups and high-propensity regions of the SuperStar map.

An important question is whether it is meaningful to use

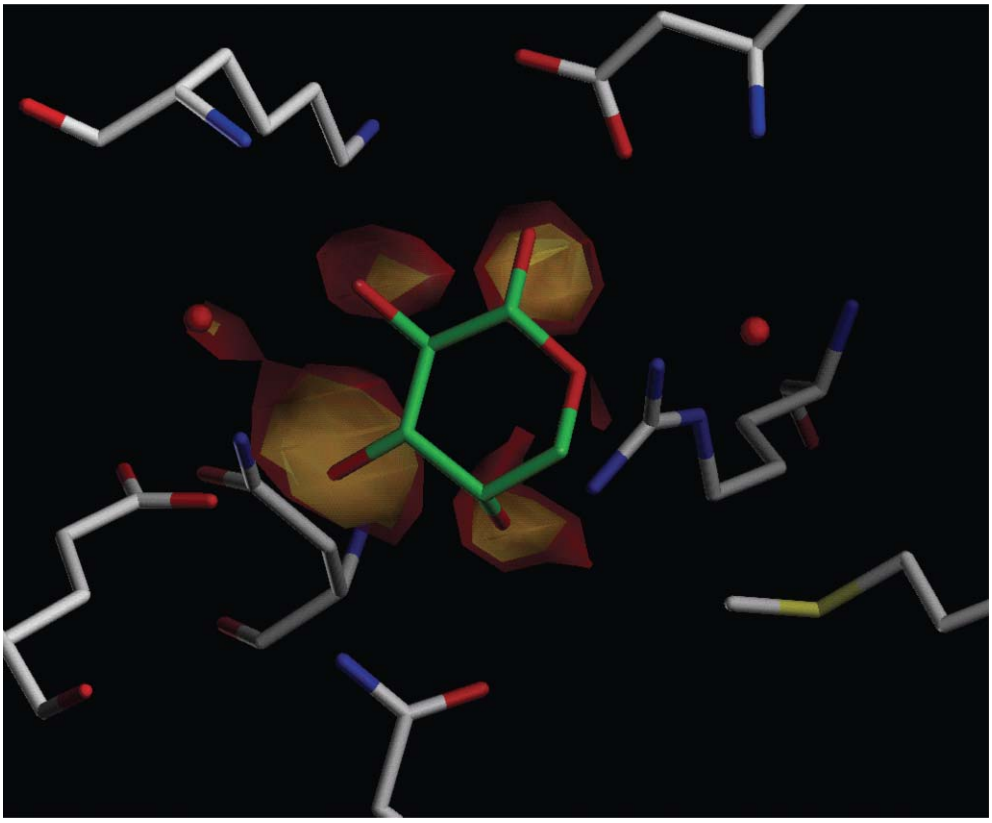


Fig. 12 SuperStar propensity map (OH probe) indicating where OH groups are most likely to bind to L-arabinose binding protein (PDB entry 1ABE[35]). The observed position of arabinose in the protein–ligand crystal structure is also shown.

small-molecule crystal-structure data to predict nonbonded interactions in protein–ligand complexes. Side-by-side comparisons of IsoStar scatterplots based on CSD and PDB data indicate that the geometries of nonbonded interactions are similar in small-molecule and protein–ligand crystal structures. This is supported by calculation of similarity coefficients for 72 pairs of scatterplots, which show that only rarely does a PDB-based scatterplot differ significantly in shape from its CSD-based analogue.[36] However, early experiments with SuperStar did highlight an important difference in the frequencies with which different types of contacts occur in small-molecule and protein–ligand crystal structures.[34] Specifically, hydrophobic contacts such as $CH_3\cdots CH_3$ occur relatively less often in the CSD than in the PDB, taking stoichiometric factors into account. Conversely, contacts between a polar group and a hydrophobe, such as $CH_3\cdots O{=}C$, are relatively more common in the CSD. A possible explanation is that a major driving force for protein–ligand binding is the entropic gain resulting from the displacement of water molecules from protein hydrophobic cavities by ligand hydrophobic groups. Most of the crystals used to obtain structures in the CSD, however, were probably not grown from aqueous solvent. Thus, it may well be that there is a smaller free-energy gain from the formation of hydrophobic contacts in the majority of small-molecule crystallisations. Whatever the reason, the difference is real and has to be corrected in order to obtain reasonable results from SuperStar. In current releases of the program, the correction is very crude: propensities for hydrophobic contacts are artificially increased by a constant factor chosen to get the best predictive reliability. However, the most recent in-house development version of the program incorporates a more theoretically satisfying correction based on substituent octanol–water π values.

SuperStar was validated on 122 protein–ligand complexes from the PDB (later validations have used many more complexes but produced comparable success rates). Each complex was prepared by removal of the ligand and placement of hydrogen atoms on the protein residues. Four propensity maps were then generated for each binding site, using the probe groups: carbonyl oxygen, $-NH_3$ nitrogen, methyl carbon and hydroxyl oxygen. Ligands were then placed back into the binding sites in their experimentally-observed positions. At any point in space occupied by a ligand atom of the same type as one of the four probe groups, it was ascertained which of the probes had the highest SuperStar propensity. If this was the probe matching the ligand atom, SuperStar was held to have made a correct prediction. Otherwise, it was incorrect. Having used four probe atoms, a 25% success rate could be expected at random; however the actual success rate varied between 68% and 82%, depending on the solvent accessibility of the ligand atom (it being harder to predict highly solvent-accessible positions).

To our knowledge, no systematic comparison has been done between results obtained using energy-based programs such as GRID and knowledge-based programs like SuperStar. Anecdotal reports indicate that both methodologies give comparable overall prediction success rates but may vary in reliability on any given system. This is to be expected: the programs take very different approaches and are unlikely to make the same errors. The use of both energy-based and knowledge-based methods together is therefore likely to give more robust results than either on its own.

8 Conclusion

This review has summarised the wide range of scientific applications of the CSD, involving fundamental direct applications of the stored data, through the generation of knowledge-based systems such as Mogul and IsoStar, to the integration of this knowledge within software applications that

are designed to solve significant problems in structural chemistry and biology. These CSD-related developments have been catalysed by the very rapid evolution of computer technology over recent decades, and also by the highly significant continuous growth of the CSD itself. While small-molecule crystal structure analysis may now be considered to be a mature science, itself fuelled by technological advances, the results of the technique remain a fundamental and lasting resource. Every crystal structure is valuable, and it is vital to capture as many structures as possible within the publicly available databases. It is a major concern[37] that the work involved in placing structural information into the public domain *via* traditional methods is now relatively time consuming by comparison with the time taken for crystal structure analysis, thus limiting the number of structures that can ultimately be archived to the databases. It is appropriate that the crystallographic community is seeking ways to improve this situation.

References

1 F. H. Allen, *Acta Crystallogr., Sect. B*, 2002, **58**, 380–388.
2 I. J. Bruno, J. C. Cole, P. R. Edgington, M. Kessler, C. F. Macrae, P. McCabe, J. Pearson and R. Taylor, *Acta Crystallogr., Sect. B*, 2002, **58**, 389–397.
3 R. Taylor and F. H. Allen, Statistical and Numerical Methods of Data Analysis, in *Structure Correlation* (ed. H.-B. Bürgi and J. D. Dunitz), VCH, Weinheim, Germany, 1994, pp. 111–162.
4 H.-B. Bürgi and J. D. Dunitz, *Acc. Chem. Res.*, 1983, **16**, 153–161.
5 H.-B. Bürgi and J. D. Dunitz, *Structure Correlation*, VCH, Weinheim, Germany, 1994.
6 R. Taylor, *Acta Crystallogr., Sect. D*, 2002, **58**, 879–888.
7 F. H. Allen and W. D. S. Motherwell, *Acta Crystallogr., Sect. B*, 2002, **58**, 407–422.
8 A. G. Orpen, *Acta Crystallogr., Sect. B*, 2002, **58**, 398–406.
9 (*a*) F. H. Allen, O. Kennard, D. G. Watson, L. Brammer, A. G. Orpen and R. Taylor, *J. Chem. Soc., Perkin Trans. 2*, 1987, S1–S19; (*b*) A. G. Orpen, L. Brammer, F. H. Allen, O. Kennard, D. G. Watson and R. Taylor, *J. Chem. Soc., Dalton Trans.*, 1989, S1–S83.
10 R. A. Engh and R. Huber, *Acta Crystallogr., Sect. A*, 1991, **47**, 392–398.
11 W. I. F. David, K. Shankland and N. Shankland, *Chem. Commun.*, 1998, 931–932.
12 F. H. Allen, S. E. Harris and R. Taylor, *J. Computer-Aided Mol. Des.*, 1996, **10**, 247–254.
13 H.-J. Böhm and G. Klebe, *Angew. Chem., Int. Ed. Engl.*, 1996, **35**, 2588–2614.
14 Z. Rappoport, S. E. Biali and M. Kaftory, *J. Am. Chem. Soc.*, 1990, **112**, 7742–7750.
15 A. J. Kirby, *Adv. Phys. Org. Chem.*, 1994, **29**, 87–183.
16 F. H. Allen, R. Mondal, N. A. Pitchford and J. A. K. Howard, *Helv. Chim. Acta*, 2003, **86**, 1129–1139.
17 F. H. Allen, C. M. Bird, R. S. Rowland and P. R. Raithby, *Acta Crystallogr., Sect. B*, 1997, **53**, 696–701.
18 I. C. Hayes and A. J. Stone, *J. Mol. Phys.*, 1984, **53**, 83–105.
19 G. R. Desiraju and T. Steiner, *The Weak Hydrogen Bond in Structural Chemistry and Biology*, Oxford University Press, Oxford, 1999.
20 R. Taylor and O. Kennard, *J. Am. Chem. Soc.*, 1982, **104**, 5063–5070.
21 G. R. Desiraju, *Angew. Chem., Int. Ed. Engl.*, 1995, **34**, 2311–2327.
22 J. P. M. Lommerse, A. J. Stone, R. Taylor and F. H. Allen, *J. Am. Chem. Soc.*, 1996, **118**, 3108–3116.
23 S. L. Price, A. J. Stone, J. Lucas, R. S. Rowland and A. E. Thornley, *J. Am. Chem. Soc.*, 1994, **116**, 4910–4918.
24 L. Brammer, E. A. Bruton and P. Sherwood, *Cryst. Growth Des.*, 2001, **1**, 277–290.
25 (*a*) R. Taylor, A. Mullaley and G. W. Mullier, *Pestic. Sci.*, 1990, **29**, 197–213; (*b*) F. H. Allen, C. A. Baalham, J. P. M. Lommerse and P. R. Raithby, *Acta Crystallogr., Sect. B*, 1998, **54**, 320–329; (*c*) P. H. Maccallum, R. Poet and E. J. Milner-White, *J. Mol. Biol.*, 1995, **248**, 374–384; (*d*) C. M. Deane, F. H. Allen, R. Taylor and T. L. Blundell, *Protein Eng.*, 1999, **12**, 1025–1028.
26 G. Klebe and T. Mietzner, *J. Computer-Aided Mol. Des.*, 1994, **8**, 583–606.
27 P. Murray-Rust, in *Molecular Structure and Biological Activity*, (ed. J. F. Griffin and W. L. Duax), Elsevier, New York, 1982, pp. 117–133.
28 B. P. Feuston, M. D. Miller, J. C. Culberson, R. B. Nachbar and S. K. Kearsley, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 754–763.
29 H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, V. Ravichandran, B. Schneider, N. Thanki, D. Padilla, H. Weissig, J. D. Westbrook and C. Zardecki, *Acta Crystallogr., Sect. D*, 2002, **58**, 899–907.
30 I. J. Bruno, J. C. Cole, J. P. M. Lommerse, R. S. Rowland, R. Taylor and M. L. Verdonk, *J. Computer-Aided Mol. Des.*, 1997, **11**, 525–537.
31 S. Y. Lin, Y. Okaya, D. M. Chiou and W. J. Le Noble, *Acta Crystallogr., Sect. B*, 1982, **38**, 1669–1673.
32 D. J. Watkin and R. J. Cooper, see http://www.xtl.ox.ac.uk/crystals.html.
33 P. J. Goodford, *J. Med. Chem.*, 1985, **28**, 849–857.
34 M. L. Verdonk, J. C. Cole and R. Taylor, *J. Mol. Biol.*, 1999, **289**, 1093–1108.
35 F. A. Quiocho and N. K. Vyas, *Nature*, 1984, **310**, 381–386.
36 M. L. Verdonk, Private Communication.
37 F. H. Allen, *Cryst. Rev.*, 2004, **10**, 3–15.