Librarians, crystal structures and drug design

Frank H. Allen and Robin Taylor

DOI: 10.1039/b511106b

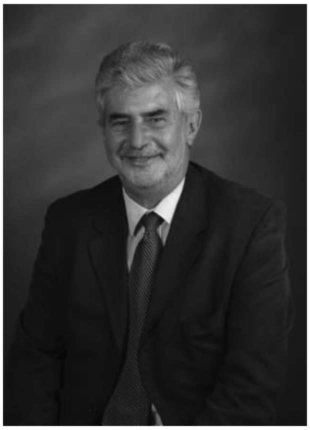
There are now 355,000 published crystal structures of organic and metal-organic compounds, all of which have been acquired, validated, chemically annotated and organised for searching in the Cambridge Structural Database (CSD). The CSD is used in rational drug design and is beginning to answer important questions relevant to the formulation of pharmaceutical active ingredients. The value and credibility of this research are ultimately dependent on the accuracy and completeness of the underlying crystal-structure data.

Science is built up with facts, as a house is with stones,[1] and it used to be that librarians were the sole custodians of scientific facts, recorded in printed journals, books and compendia. Nowadays, e-journals, electronic databases and the Web are major sources of information for all and, indeed, the Web is perceived as the *only* source of information by many. Reliance on the Web, in particular, raises questions about the underlying sources of information; if we are to perform top-quality science, our data sources must have top-quality accuracy and completeness. Witness the apocryphal story of the thermal conductivity of indium, which was copied from one compendium to another until one researcher became suspicious – the datum did not fit an expected trend – and found that the value had been mis-transcribed from the original paper. One wonders how many failed experiments or missed inventions resulted from this mistake, and how many more problems were avoided by catching the error before it was broadcast further.

In structural chemistry, crystal structures have a special importance. They are valuable because they provide precise geometrical information about the sizes and shapes of molecules, as well as being *the* major source of experimental data about intermolecular interactions. In fact, rather few of the data generated are normally of interest to the original researcher, who may only wish to know the geometry of a metal coordination sphere or the stereochemistry at a specific atom. However, other aspects of the structure may have significance for other scientists, *e.g.*, conformations about flexible bonds, or short non-bonded contacts and their geometries. This type of experimental information, which is important in rational drug design, is now readily available in a few button clicks from two electronic libraries[2,3] generated from the Cambridge Structural Database (CSD),[4] the definitive database of organic and metallo-organic small-molecule crystal structures.

Compilation of the CSD began at the newly formed Cambridge Crystallographic Data Centre (CCDC) in 1965, so the CCDC shares its 40th anniversary

Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, UK CB2 1EZ.
E-mail: allen@ccdc.cam.ac.uk; taylor@ccdc.cam.ac.uk



Frank Allen (b. 1944) 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.

Robin Taylor (b. 1951) studied chemistry at Oxford and Cambridge Universities, and followed that with short spells at York and Pittsburgh Universities and Westminster Medical School, where he first developed his interest in statistics and what is now called data mining. Having spent five years at the CCDC in the early 1980s, he left for almost a decade of industrial research at ICI/Zeneca Agrochemicals, before returning to the CCDC where he now leads the software development team.

with *Chemical Communications*. With computers in their infancy, early progress was slow, but from the outset the paramount guiding principles were to maximise accuracy and completeness in the developing database. Until the advent of the Crystallographic Information File (CIF),[5] all data – nearly 200,000 structures – had to be re-encoded from journal articles and their associated hard copy supplementary information. If we accept that the word that precedes 'mining' in normal English syntax is the commodity that we seek, then the CCDC were amongst the first practitioners of *real* 'data mining' – the location of relevant data in libraries, and often with the help of librarians.

Systems were devised for data validation – some 10% of typeset structure data contained at least one numerical error – and for adding bibliographic, structural and chemical information, such as compound names and, most importantly, the formal bond types that provide full 2D and 3D chemical searchability. Even though the CIF now avoids most of the pitfalls of transcription, many of the 150,000 CIFs received in recent years are themselves not without fault – format errors, mis-edited data items, *etc.* – and validation still remains just as crucial for electronic data, as it does for the 3–4% of structures for which data are still re-typed from hard-copy documents.

There have, of course, been major improvements in the automated building of CSD entries, including the algorithmic assignment of chemical information such as bond types. Here, success rates are currently at about the 85% level, despite the extreme range and novelty of organic and, particularly, metal-organic chemistry recorded in the CSD. However, manual editing is still necessary to achieve the highest possible standards. Thus, the CCDC's scientific editors – skilled electronic librarians – continue to add value in many ways. The CSD of 2005 currently records 355,000 structures, and this 40-year growth suggests that, had operations not started in 1965, a complete, accurate archive might not have been possible: at 2005 values, the cost of assembling the CSD would exceed £15 million. It is interesting, also, to speculate on how many years of human endeavour are recorded within the CSD. Even at an average over time of one month per structure (conservative, given that someone spent time synthesising the substance, in addition to the crystallographer's efforts), this figure is close to 30,000 years.

The CCDC provides software[6] to search the CSD for 2D and 3D substructures – including non-bonded fragment searches – and for visualising and analysing the geometries of retrieved structures. The complete CSD System is widely used in the pharmaceutical industry, and in academic institutions in 65 countries. Over 1,200 papers based on CSD information have been published.[4,7] Most of these describe knowledge mining experiments, and, in recent years, the CCDC has developed two extensive electronic libraries of fundamental structural knowledge derived from the CSD: Mogul[2] and IsoStar.[3] These libraries can provide almost instantaneous answers to common questions posed by structural chemists, and have particular relevance in drug design.

The earliest, and still the most common, use of the CSD in rational drug design is as a source of reliable, experimentally-determined molecular geometries. Substructure searching can be used to investigate the conformational preferences of any molecular fragment, provided, of course, that it is common enough to appear in a few CSD crystal structures. This method of conformational analysis has become somewhat less important as theoretical molecular-geometry calculations have improved, but is still very useful, particularly for difficult systems such as metal complexes, hypervalent species and molecules capable of forming intramolecular hydrogen-bonds. In the latter case, *in vacuo* calculations tend to find intramolecular hydrogen-bonds that may not occur in a condensed phase. Conversely, the CSD gives reliable information on which intramolecular hydrogen-bonded motifs are likely in non-gaseous states.

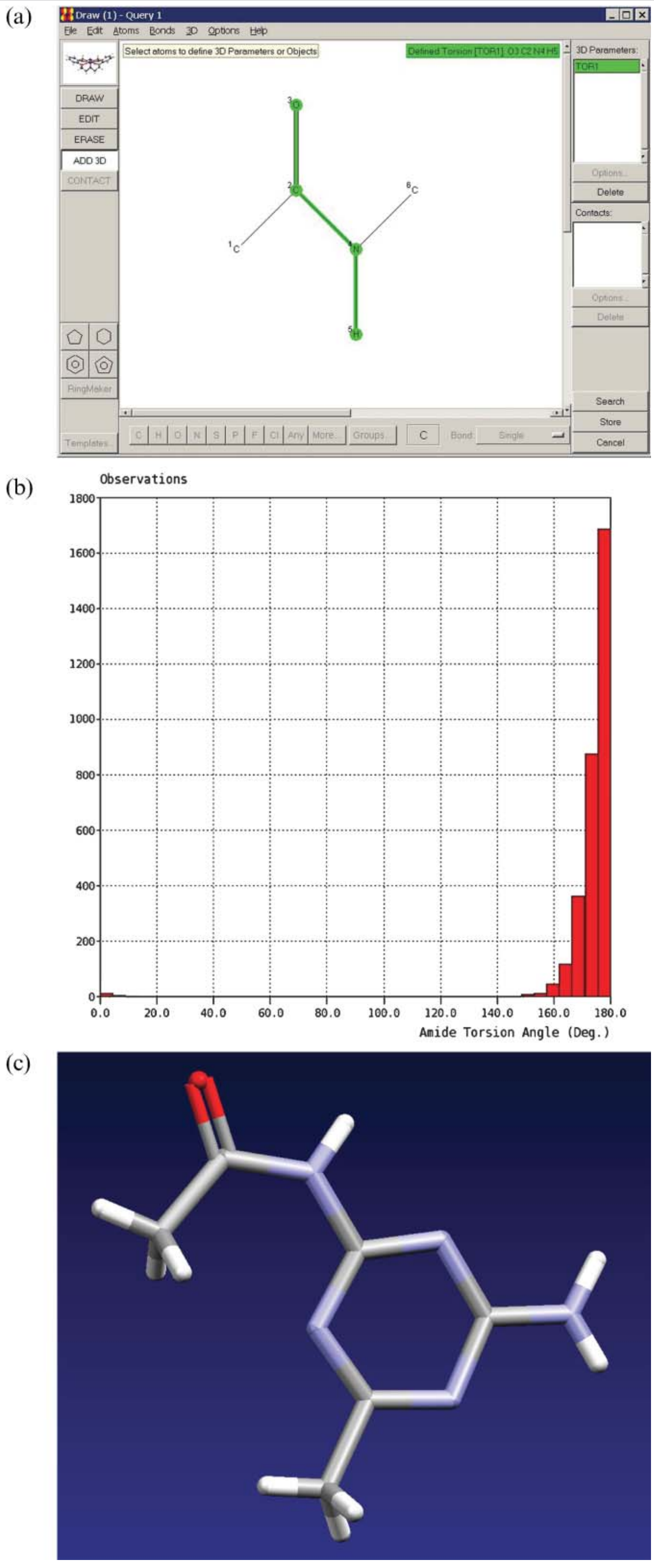
A key advantage of using the CSD for conformational analysis is that it often gives clues as to *how* a particular conformation may be achieved. Suppose, for example, that we want to design a molecule containing a *cis*-amide linkage, suspecting that it will interact well with a particular arrangement of atoms in an enzyme active site. Lactams are obvious, but can we find acyclic amides with a *cis* preference? A search of the CSD for acyclic amides (Fig. 1a) shows, unsurprisingly, that almost all are *trans*. A few, however, are *cis* (Fig. 1b: torsion angle $< 20°$). When they are examined, a surprising number have the amide nitrogen bonded to the 2-position of a pyrimidine ring, or to a related aromatic ring in which both *ortho* ring atoms are N. In fact, almost half the hits from a substructure search for an amide in this environment have the *cis* geometry (Fig. 1c), suggesting this system as a promising candidate.

Drug design groups now subject very large numbers of molecules to virtual high-throughput screening using protein–ligand docking programs.[8] Potentially, the conformational preferences obtainable from the CSD can be used to reject molecules whose docked conformations contain unlikely geometrical features. For this purpose, however, it is essential that the CSD searches are performed automatically, not by a user drawing a substructure and searching for it manually. Very fast searching is also required, since millions of docking solutions may need to be examined. These considerations have led to the development of Mogul[2] to retrieve automatically the CSD torsion-angle data required to "validate" a given molecular conformation. This is done by classifying all the acyclic torsion angles in the CSD into types, based on a series of keys that capture features of the chemical environment of each torsion. By use of a tree indexed on these keys, it is possible to extract very quickly those torsional fragments from the CSD that match any given torsion angle in a query molecule. It is similarly possible to retrieve bond-length and valence-angle distributions from the > 20 million data items indexed in Mogul, allowing the values of these parameters to be checked in any molecule of interest. The crystal-structure solution program CRYSTALS,[9] for example, uses Mogul both to highlight unlikely geometrical features in a partially-refined crystal structure, and, if required, to set target values for use in restrained least-squares refinement. The possibility of using an analogous procedure to set up ligand dictionaries for refinement of protein–ligand crystal structures is obvious. Some workers have gone one step further and used CSD information in programs, *e.g.* MIMUMBA[10] and *et*,[11] that will not just validate, but actually predict, low-energy conformations.

One of the strengths of crystallography is the detailed information that it gives on intermolecular non-bonded contacts. By superimposing crystallographically-observed contacts between two groups A and B so that the A moieties are overlaid, a three-dimensional scatterplot can be produced (Fig. 2a) showing the experimental distribution of B (the "contact" or "probe" group) around A (the "central group"). The scatterplot can be converted to a contoured surface (Fig. 2b) showing the density of contact groups around the central group. Moreover, a normalisation procedure based on the stoichiometries and unit-cell volumes of the contributing crystal structures can be used to convert the contoured surface into a "propensity" plot. A propensity of p indicates that the density of contacts is p times what would be expected at random, so regions where $p > 1$ are, by implication, energetically favourable, and *vice versa*.

The IsoStar library[3] contains over 25,000 of these scatterplots, each one giving information on a particular A···B pair. This compendium is useful as a basic source of knowledge and as an ideas generator. For example, drug designers frequently need answers to simple questions, such as whether a fluorine atom is likely to form a hydrogen bond, or whether a thiazole sulfur is likely to form a non-bonded contact to a carbonyl oxygen atom. Visual examination of the appropriate IsoStar scatterplots is a quick and reliable way of answering such queries. Because the scatterplots are hyperlinked to the underlying crystallographic database, any individual contact can be examined. This often gives ideas about how a particular interaction can be stabilised; for example, hyperlinking from the scatterplot of aromatic carbon atoms around indole suggests different types of electron-deficient ring systems that might, if incorporated into a ligand, form a stabilising, parallel-stacked interaction with a tryptophan side-chain.

Apart from its use as an encyclopaedia of non-bonded interactions, IsoStar is used by the SuperStar program[12] to


Draw (1) - Query 1

Select atoms to define 3D Parameters or Objects

Defined Torsion [TOR1] O3 C2 N4 H5

Fig. 1 (a) ConQuest search query for amides in which the amide C–N bond has been constrained to be acyclic; (b) Torsion angle distribution from the results of the search of Fig. 1a; (c) Example of the N_{ar} – C_{ar}(–N_{amide}) – N_{ar} system containing acyclic amides that are most likely to exist in a *cis*-conformation (see text).

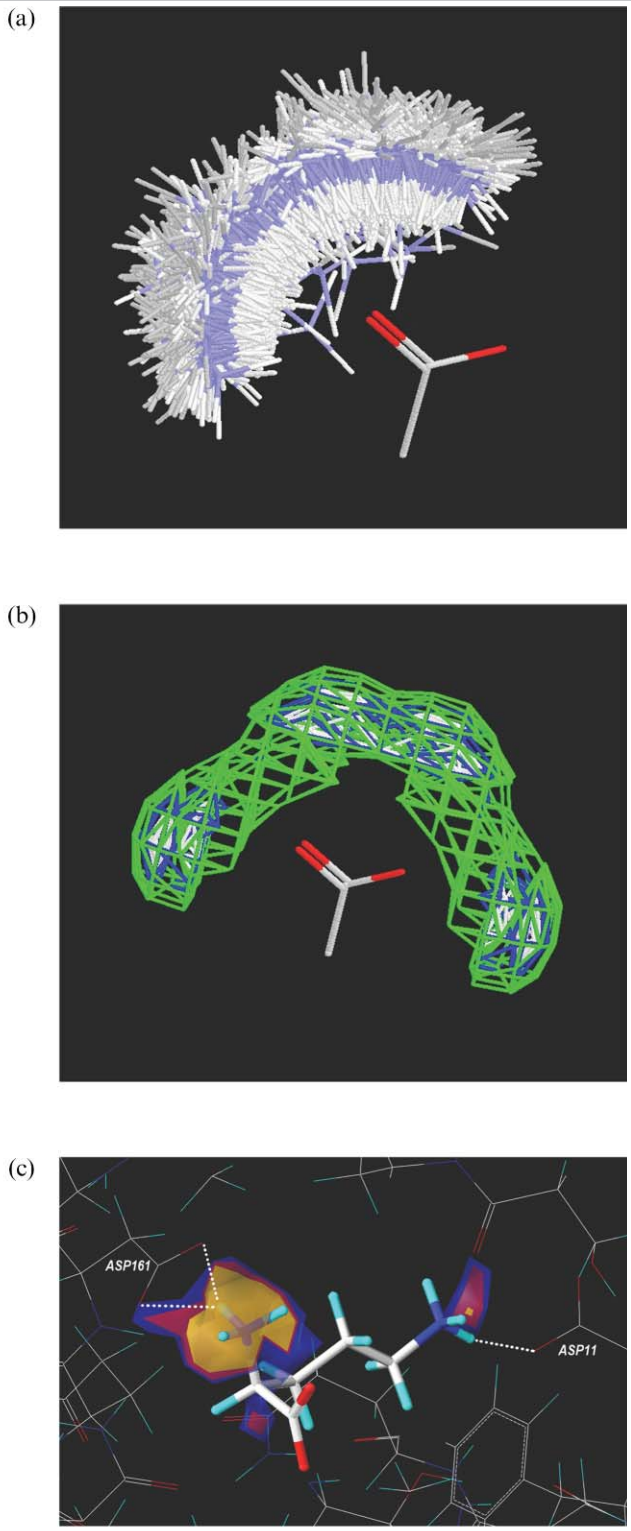


Fig. 2 (a) IsoStar scatterplot showing the distribution of NH_3^+ "contact" groups around an ionised carboxylate "central" group; because of the *mm*-symmetry of the carboxylate, all contacts are reflected into a single quadrant of the interaction space; (b) Contour plot of the distribution of Fig. 2a showing the density of NH_3^+….carboxylate contacts; highest density contours are drawn in white, followed by blue and then green; (c) SuperStar plot for ornithine binding protein (PDB entry 1LAH), showing the two NH_3^+ groups in the ligand occupying positions predicted for that group by the SuperStar algorithm.

identify binding "hot spots" in enzyme active sites. This program first divides the residues in an active site into their constituent functional groups. The crystallographically-observed propensity distribution of a chosen probe around each functional group is retrieved from IsoStar and overlaid onto the group in the active site. The final result is a composite map for the entire site, showing the likely points at which the probe group will bind (Fig. 2c). The reliability of SuperStar can be estimated by comparing its predictions with the observed positions of ligand groups in experimental protein–ligand crystal structures taken from the Protein Data Bank (PDB[13]). This shows the program to be reliable, but only when corrections are made for two important differences between small-molecule crystal structures and protein–ligand complexes. First, hydrophobic contacts are more frequent in the latter, perhaps because many small-molecule crystals are grown from non-aqueous solvents. This problem can be corrected by use of water–octanol π coefficients, suitably attenuated to allow for the fact that the average crystal-field is not as hydrophobic as octanol. Secondly, because small-molecule crystal structures are tightly packed, relatively long non-bonded contacts are under-represented compared with protein–ligand complexes. This is treated by setting propensities to 1 beyond a suitable contact distance.

Recently, interest has arisen in applying the CSD to the problems of drug development (as opposed to its established use in drug *discovery*). The focus is on questions such as whether a given active ingredient is likely to crystallise in more than one polymorphic form, whether it is prone to form hydrated crystal structures, and what the best choice of counter-ion or co-crystal partner might be. These are challenging questions, and the required methodologies are only just emerging, but recent publications suggest that the CSD may be of value. For example, the occurrence of hydrates in the CSD may be correlated with the presence or absence of particular groups.[14] For ionic active ingredients, searches of the CSD for robust hydrogen-bonded motifs may help in selecting counter-ions that are likely to give stable crystal structures (Fig. 3).[15] In

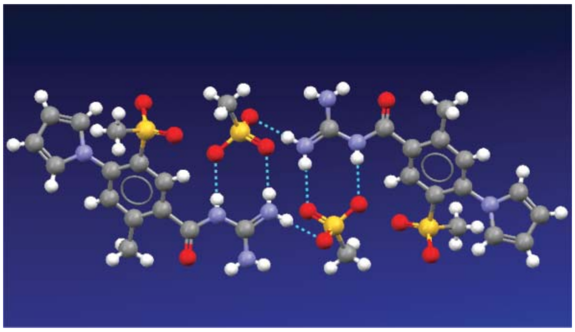


Fig. 3 Many drug molecules are ionised and must therefore be formulated with counter-ions. Searching the CSD for robust hydrogen-bond networks, such as those involving the sulfonate ion shown (CSD entry NEGTIC), may help pharmaceutical development chemists select counter-ions that will produce stable crystal forms, leading to longer shelf-life.[15a]

crystal-structure prediction, ranking of candidate structures on calculated lattice energy is of restricted value, partly because of limitations in the energy calculations, and also because, presumably, crystals sometimes form under kinetic control. The CSD, being a record of what has occurred in practice, may yield information that can be used to re-rank predicted structures[16] (*e.g.*, depending on whether a predicted structure contains a motif that is found commonly in the crystal structures of similar molecules).

As a complement to these cutting-edge problems, staff at CCDC face challenging, if more mundane, difficulties in ensuring that the CSD remains accurate, up to date and as comprehensive as possible. The key issue is one of numbers: leading crystallography groups have the capacity to solve literally thousands of structures a year. Current input to the CSD is about 30,000 structures per annum with a historical doubling period of 8–9 years; this period may well decrease given the advances that have occurred in experimental crystallography. Two problems are thus created. First, improved software is needed for database creation. The aim is to deal automatically with as many structures as possible (*e.g.* using algorithms to assign bond types and generate clear chemical diagrams). This will leave a manageable number of difficult structures for editors to process manually (*e.g.* cluster compounds, complexes with metal–metal multiple bonds, metal complexes involving redox-active ligands). Secondly, crystallographers and chemists do not have time to publish all the crystal structures that are solved.[17] Ensuring that as many as possible of these unpublished structures find their way into the CSD is a logistical and political problem that is receiving attention by the CCDC, the International Union of Crystallography, and other organisations.

Finally, let us turn back to the librarian. The knowledge-based methods described above are only as good as the data on which they are based. How is this essential foundation of "trusted data" to be maintained and paid for? Crystallographic databases have different funding mechanisms. For example, the PDB is funded by US National Agencies and is free to the end-user. The CSD gets little or no public funding, so end-users pay directly. Unsurprisingly, the former model is the more popular with scientists. Indeed, it is sometimes argued that published crystal structure data should be free to scientists as of right. Even if this argument were to be accepted, it misses a point: the natural state of published crystal structures is to be scattered in a highly inconvenient manner.[18] The user of the CSD is paying, not for the results themselves, but for the convenience of having them collected together, thoroughly checked, and distributed in a highly searchable form. The CCDC also works with major journal publishers to facilitate the flow of data prior to publication (for example, ensuring referees get access to structures associated with submitted papers, and that each of these structures is novel). CCDC staff will often track a structure through a sequence of submissions to several journals, possibly in different revisions, until its eventual publication and accession to the CSD.

Whether the costs of database building should be borne directly by the tax payer or the end user is open to question, but someone has to pay. Or do they? The advent of the Web raises the possibility that collections of crystal structures could be created and placed on the Internet by the cooperative endeavours of crystallographers. Why not move to an informal, collective approach like this? The answer, above all, is that someone has to take responsibility for keeping crystallographic databases as comprehensive as possible, for checking and enhancing the data, and for putting things right when they are found to be wrong (typographical errors, space-group mis-assignments, missing solvents, incorrect bond types, *etc.*). This is exacting work, and, like the efforts of the librarian who keeps the right books on the right shelves, tends to go unnoticed. But if we drop our guard and allow our key sources of crystal structures to become fragmented, corrupted and untrustworthy, we will have irretrievably lost something unique. In general life, it may not matter that our sources of information fail to meet the highest possible standards, which is why we read newspapers, and search an Internet flooded with information of unknown provenance. But in science we must have higher aspirations.

Acknowledgements

We thank Dr Gary Battle (CCDC) for valuable assistance in preparing the Figures.

Notes and references

1 J. H. Poincaré, *La Science et l'hypothèse*, Flammarion, Paris, 1902.

2 I. J. Bruno, J. C. Cole, M. Kessler, J. Luo, W. D. S. Motherwell, L. H. Purkis, B. R. Smith, R. Taylor, R. I. Cooper, S. E. Harris and A. G. Orpen, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 2133–2144.

3 I. J. Bruno, J. C. Cole, J. P. M. Lommerse, R. S. Rowland, R. Taylor and M. L. Verdonk, *J. Comput. Aided Mol. Des.*, 1997, **11**, 525–537.

4 F. H. Allen, *Acta Crystallogr., Sect. B*, 2002, **58**, 380–388.

5 S. R. Hall, F. H. Allen and I. D. Brown, *Acta Crystallogr., Sect. A*, 1991, **47**, 655–685.

6 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.

7 See: www.ccdc.cam.ac.uk/free_services/webcite.

8 B. Shoichet and J. Alvarez, *Virtual Screening in Drug Discovery*, Taylor & Francis CRC Press, Boca Raton, Florida, USA, 2005.

9 R. I. Cooper and D. J. Watkin, *Acta Crystallogr., Sect. A*, 2002, **58** (Supplement), C58. See also: http://www.xtl.ox.ac.uk/crystals.html.

10 G. Klebe and T. Mietzner, *J. Comput. Aided Mol. Des.*, 1994, **8**, 583–606.

11 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.

12 (*a*) M. L. Verdonk, J. C. Cole and R. Taylor, *J. Mol. Biol.*, 1999, **289**, 1093–1108; (*b*) J. W. M. Nissink, C. W. Murray, M. J. Hartshorn, M. L. Verdonk, J. C. Cole and R. Taylor, *Proteins*, 2002, **49**, 457–471; (*c*) J. W. M. Nissink and R. Taylor, *Org. Biomol. Chem.*, 2004, **2**, 3238–3249.

13 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235–242.

14 (*a*) L. Infantes and W. D. S. Motherwell, *CrystEngComm*, 2002, **4**, 454–461; (*b*) L. Infantes, J. Chisholm and S. Motherwell, *CrystEngComm*, 2003, **5**, 480–486.

15 (*a*) D. A. Haynes, J. A. Chisholm, W. Jones and W. D. S. Motherwell, *CrystEngComm*, 2004, **6**, 584–588; (*b*) D. A. Haynes, W. Jones and W. D. S. Motherwell, *J. Pharm. Sci.*, in press.

16 J. A. Chisholm and W. D. S. Motherwell, in preparation.

17 F. H. Allen, *Cryst. Rev.*, 2004, **10**, 3–15.

18 *"The growing abundance of primary scientific publications and the confusion with which it is set out acts as a brake, as an element of friction, to the progress of science."* J. D. Bernal, *Royal Society Scientific Information Conference, Introduction to the Report and Papers*, The Royal Society, London, 1948.