DOI: 10.1002/anie.201306438

Structural Databases

The Cambridge Structural Database in Retrospect and Prospect

Colin R. Groom and Frank H. Allen**

CCDC origins and development · crystal structure database · drug discovery · drug formulation · structure correlation

The Cambridge Crystallographic Data Centre (CCDC) was established in 1965 to record numerical, chemical and bibliographic data relating to published organic and metal–organic crystal structures. The Cambridge Structural Database (CSD) now stores data for nearly 700 000 structures and is a comprehensive and fully retrospective historical archive of small-molecule crystallography. Nearly 40 000 new structures are added each year. As X-ray crystallography celebrates its centenary as a subject, and the CCDC approaches its own 50th year, this article traces the origins of the CCDC as a publicly funded organization and its onward development into a self-financing charitable institution. Principally, however, we describe the growth of the CSD and its extensive associated software system, and summarize its impact and value as a basis for research in structural chemistry, materials science and the life sciences, including drug discovery and drug development. Finally, the article considers the CCDC's funding model in relation to open access and open data paradigms.

1. Origins of the Crystallographic Databases

Crystallography has been replete with commemorations recently, and particularly the centenary of the determination of the very first crystal structure (zinc blende) by W. L. and W. H. Bragg in 1912 and 1913,[1] and the designation of 2014 as the International Year of Crystallography (IYCr) by the United Nations. The IYCr provides an opportunity to look back at the successes of the subject and to examine how crystal structure information continues to inform and influence major scientific developments. The crystallographic databases are a historical record of the past hundred years and have played a major role in bringing crystal structure data to scientists in many disciplines. What is it, then, that makes crystal structure data so uniquely valuable?

Following the work of the Braggs, X-ray crystal structure analysis was quickly recognized as a very special analytical technique indeed: in 1929, just 16 years after its discovery, the output of its practitioners was already being compiled from their disparate original sources by *Strukturberichte* (Structure Reports),[2a] to provide readily accessible descriptions of newly determined crystal structures on a regular publication schedule. The *Strukturberichte* morphed seamlessly into *Structure Reports*[2b] as an official publication of the International Union of Crystallography (IUCr) until the 1990s. So, despite 1929 being the year of the Wall Street crash, it was a rather special year for the curating of crystal structure information. It was also special in heralding the enormous scientific value of that curated information, since it was in 1929 that Linus Pauling published his five rules for determining the structures of complex inorganic ionic crystals.[3] These rules were derived by analysis of data accumulated in the previous 16 years, almost certainly providing the first example of structure correlation or structural systematics.

Crystal structure analysis was also moving away from its inorganic origins and beginning to resolve a long list of uncertainties and unknowns in general structural chemistry. For organic compounds, X-ray crystallography soon confirmed or established fundamental information about the nature of chemical bonding, the planarity of benzene rings, the structures of an increasingly wide range of natural and

[*] Dr. C. R. Groom
Executive Director, Cambridge Crystallographic Data Centre
12 Union Road, Cambridge CB2 1EZ (United Kingdom)
E-mail: groom@ccdc.cam.ac.uk
Dr. F. H. Allen
Emeritus Research Fellow, Cambridge Crystallographic Data Centre
12 Union Road, Cambridge CB2 1EZ (United Kingdom)
E-mail: allen@ccdc.cam.ac.uk

synthetic molecules, and importantly, the nature of intermolecular interactions, principally hydrogen bonds. All of this knowledge would soon be vital in molecular biology, a huge endeavor that was clearly envisioned even when the determination of the smallest organic structure was time consuming and often fiendishly difficult. The rest, as they say, is history—a history in which crystallography has played a major role in the award of 28 Nobel Prizes.[4]

Philosophically then, 1929 heralds the era of the modern computerized crystal structure databases which began operations nearly four decades later: The Cambridge Structural Database (CSD: Cambridge, UK)[5] of organic and metal–organic structures, founded in 1965, was followed in the early 1970s by the Inorganic Crystal Structure Database (ICSD: Karlsruhe, Germany),[6a] Metals and Alloys Crystal Structures Database (CRYSTMET: Ottawa, Canada),[6b] the Protein Data Bank[6c] (PDB: Brookhaven National Laboratory, NY, USA, now operated as the Worldwide PDB),[6d] and the Nucleic Acid Database (NDB: Rutgers University, NJ, USA).[6e]

2. The Cambridge Crystallographic Data Centre

Between 1929 and the early 1960s, printed data compendia reigned supreme as reference sources in most sciences. In crystallography, the acknowledged leaders, *Strukturberichte* and *Structure Reports*, were joined by a number of more specialized compilations, for example, *Crystal Data*,[7] *Tables of Interatomic Distances and Configuration in Molecules and Ions*,[8] and others. During this period also, scientists of all disciplines were becoming concerned about keeping pace with the rapid growth of the scientific literature—the information explosion. International discussions were held, beginning with The Royal Society Scientific Information Conference held in London in 1948, which generated a 700 page report.[9] In all of this, the physicist and crystallographer J. D. Bernal (1901–1971) was a major figure, and the related notes draw on a historical memoir presented in the 1980s by Bernal's collaborator Olga Kennard.[10]

In the Royal Society Conference Report,[9] Bernal noted that "*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*". Ultimately, promptings by many scientists brought the scientific information explosion to the attention of national governments and led to the creation of various organizations and projects. The possibilities offered by rapidly developing computer technologies also began to be realized. A joint working party of the Royal Society and the Department for Education and Science was asked to plan the UK contribution to a global effort. In 1964, Olga Kennard, who had been working with Bernal at Birkbeck College, London on some of the printed compendia,[7,8] was invited to create a "Crystallographic Data Centre" with funding from the new UK Office for Scientific and Technical Information (OSTI).

3. The Cambridge Structural Database

The Cambridge Crystallographic Data Centre (CCDC) was established in the Department of Organic Chemistry, University of Cambridge in 1965, where Olga Kennard had been invited to form an X-ray crystallography group. The CCDC's remit was to create a comprehensive and fully retrospective computerized database of organic and metal–organic structures determined by diffraction methods (X-ray and neutron). The database was to include bibliographic, chemical, and crystallographic information, and most importantly, the 3D atomic coordinate data generated by each analysis. Thus, while computer-based bibliographic-text databases were still in their infancy, the CCDC was charged with creating one of the world's first fully electronic numerical data compilations. Importantly also, the embryo CSD would be growing up within a scientific department with the close involvement of active researchers, an involvement that has continued to be a guiding principle.

All information in the developing CSD had to be abstracted from printed journals, and at its inception the CCDC was faced with a backlog of about 4000 structures while assimilating all current publications. This required

Colin R. Groom has been Executive Director of the Cambridge Crystallographic Data Centre since 2008. He has a B.Sc. in biotechnology and a Ph.D. in protein crystallography from the University of Leeds and held postdoctoral appointments at Leeds and at Massey University, New Zealand. He joined Pfizer UK in 1994, establishing a protein crystallography facility and then establishing and leading molecular informatics units in both the UK and the USA. He joined UCB (Celltech) in 2002, leading computational and investigational chemistry teams and projects. He has been at the forefront of applying structural information to drug design and has a number of publications and patents in this area. He is a Fellow of the Royal Society of Chemistry and is on the Editorial Boards of several journals.

Frank H. Allen is an Emeritus Research Fellow at the CCDC, where he was Executive Director until his retirement in 2008. He has worked at the CCDC since 1970, following undergraduate and graduate studies in chemistry and crystallography at Imperial College, London and postdoctoral work at the University of British Columbia, Vancouver. At the CCDC he has been involved in creating the Cambridge Structural Database (CSD), in software development, and in research applications of CSD data; he has authored more than 230 publications. He is a Fellow of the Royal Society of Chemistry and was awarded the RSC Silver Medal and Prize for Structural Chemistry in 1994 and the Herman Skolnik Award of the American Chemical Society Division of Chemical Information in 2003. He was Editor of Acta Crystallographica, Section B from 1993-2002.

systems for identifying those publications that contained crystal structures, abstracting bibliographic and chemical information, and re-keyboarding numerical data tables. It also required novel software for checking data integrity and internal consistency, performing corrections (with authors' involvement), and organizing and disseminating the evaluated information.[11] It will be no surprise that *Strukturberichte* and *Structure Reports* were of immense value in dealing with the backlog.

Early dissemination of the accumulated information was a priority to reassure funding agencies and engage the user community. From 1970, this took the form of annual Bibliographic Volumes in the *Molecular Structures and Dimensions* (MSD) series,[12a] published in collaboration with the IUCr. The volumes were classified into 86 chemical "chapters", with a variety of indexes and, from Volume 12 onwards, the inclusion of 2D chemical diagrams. The MSD series was augmented in 1972 by Volume A1: *Interatomic Distances 1960–1965*. All MSD volumes were produced using the (then) new computer typesetting technology. One reviewer,[12b] having noted the size and weight of one individual volume (2.4 kg, $32 \times 23 \times 5$ cm), suggested various alternative uses (as a doorstop, flower press, etc.), but also complimented the series as "…. *essential for all scientists concerned with organic or organometallic molecular structures in the crystalline state. It removes every excuse for ignorance concerning the literature in this field*".

Despite these kind words, the MSD books, useful as they were for manual literature searches, resonated back to the era of printed compilations. The full value of the CSD could only be realized through specialist software for searching the database, analyzing its contents, and displaying structures and data. Early software was completed by 1978[5a] and has since undergone continuous development.[5b,c] The ability to export the complete CSD System was initially problematic, given the number of computer operating systems then in vogue, and users usually had to carry out local implementations. Punched cards and 2400-foot-long magnetic tapes were the order of the day! These problems began to recede with the advent of the DEC Vax and cartridge tapes, heralding the modern era of CDs and relatively few universal operating systems.

4. Development of CCDC Funding Models

Building a comprehensive data-oriented system such as the CSD System is not free: it requires experienced staff with diverse scientific and technical talents, together with a continually upgraded technology infrastructure. All of this requires financial support, and CCDC funding can be divided into two eras. First, UK Government funding was provided by OSTI and then by the UK Science and Engineering Research Council (SERC) during the 1970s, with the University of Cambridge contributing accommodation and computing facilities. Interest in the CSD System grew rapidly from the late 1970s, principally from universities worldwide, and a network of National Affiliated Centres (NACs) was established, each contributing agreed additional funding based on local CSD usage and local economic factors. Each NAC was responsible for local distribution, but some smaller countries formed regional groupings or were supplied directly from Cambridge. For academics, variants of this system still remain as active and valuable collaborations.

A further impetus to CSD usage and finances began in the early 1980s with interest from major pharmaceutical companies. Computational modeling, as a basis for what was then termed rational drug design, required extensive structural knowledge, particularly about the conformational preferences of molecules and the intermolecular interactions made by chemical functional groups. Information was also needed to parameterize force fields and to provide the experimental underpinning for automated conformer generation and protein-ligand docking. Information in the CSD was ideal for such applications and is now established as an integral part of modern drug discovery.[13] Other companies, for example in the fine chemicals and petrochemical industries, also became CSD System subscribers.

The advent of commercial revenues prompted discussion with the UK SERC and the University, and in 1987 the CCDC began to break away from public funding, so that it would not be competing directly for funding with the very scientists it was established to support. The CCDC became fully independent by 1989 as a self-financing, self-administering UK Charity—heralding the second financial era. The new CCDC remained close to the University and is now a University Partner Institute, accepted as a suitable institute for the training of postgraduate researchers. As a Charity, the affairs of the CCDC are overseen by an international Board of Governors (Trustees) who represent the beneficiaries, namely depositors and users, and also provide strategic and scientific input to the development of the Centre.

5. The Current CSD System

The CSD now contains information on nearly 700000 crystal structures, with its annual growth shown in Figure 1. Data from around 250000 structures were re-keyboarded directly from journals or supplementary documents, and all of the data were carefully checked for typographical errors and

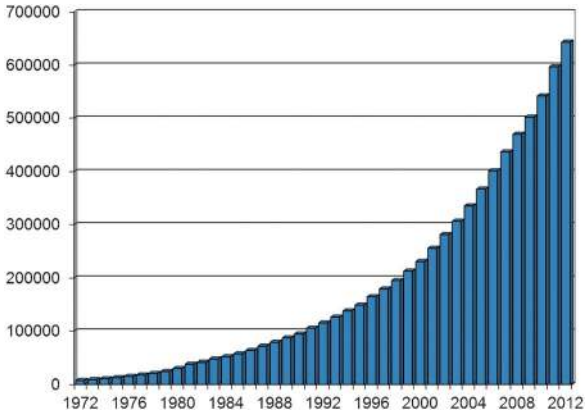

Figure 1. Annual growth of the Cambridge Structural Database from 1970 to 2012.

for scientific integrity. A turning point in data acquisition came in the early 1990s with the development, in collaboration with the IUCr and its Commissions, of the Crystallographic Information File (CIF)[14] and its subsequent universal adoption for the electronic transmission of crystal structure data. The IUCr[15] and the CCDC[16] now provide software systems for CIF checking by authors, and the incidence of errors has fallen dramatically. Early in 2013 new internal software, making full use of modern technology, now accelerates the movement of incoming raw CIF data to final CSD entries, sweeping away a number of outmoded clerical procedures. This software incorporates scientific modules, for example deCIFer,[17] that automate data evaluation by codifying the experience gained during 40 years of manual operations. A brief statistical overview of CSD information content is given in Table 1.

Over the last 15 years, the software provided to CSD users has also undergone significant development. *ConQuest*[18a] performs searches of all CSD information fields and can combine 2D substructure searches with 3D geometrical constraints to locate hydrogen bonds or other nonbonded interactions. A Web implementation, *WebCSD*, is also available[18b] as illustrated in Figure 2. *Mercury*,[18a,20] the CCDC's structure visualizer, has now developed into a comprehensive analysis suite for both structures and geometrical data. Apart from standard options, *Mercury* (Figure 3) will display intermolecular interactions, H-bonded synthons,[23] extended networks and graph-set descriptors[24] and computed powder patterns. Facilities for analyzing geometrical data[25] retrieved by *ConQuest* are also now integrated within *Mercury*.

The CSD System also contains two extensive knowledge bases. *Mogul*[26a] contains more than 20 million bond lengths, valence angles, and torsions organized into more than 1.5 million chemically searchable distributions, each relating to a specific chemical environment, as shown in Figure 4. A recent extension[26b] incorporates searchable conformation data for chemical rings. *Mogul* will also generate torsional distributions for all rotatable bonds in an input molecule, or check all geometry against mean values from the CSD, a feature that is useful in solving and refining novel crystal structures of both small molecules,[27] and of ligands bound to

Table 1: Summary statistics for the Cambridge Structural Database on July 1, 2013.

	Structures	[%] of CSD
total no. of structures	658059	100.0
no. of different compounds	601308	–
no. of literature sources	1518	–
organic structures	280809	42.6
transition metal present	353201	53.7
Li–Fr or Be–Ra present	33011	5.0
main-group metal present	40166	6.1
3D coordinates present	614824	93.4
error-free coordinates	604539	98.3[a]
neutron studies	1583	0.2
powder diffraction studies	2721	0.4
low/high temp. studies	288213	43.9
absolute configuration determined	13510	2.1
disorder present in structure	149994	22.8
polymorphic structures	19990	3.0
R-factor <0.100	618027	93.9
R-factor <0.075	560089	85.1
R-factor <0.050	361367	54.9
R-factor <0.030	74501	11.3
no. of atoms with 3D coordinates	50821771	–

[a] Taken as a percentage of structures for which coordinates are present in the CSD.

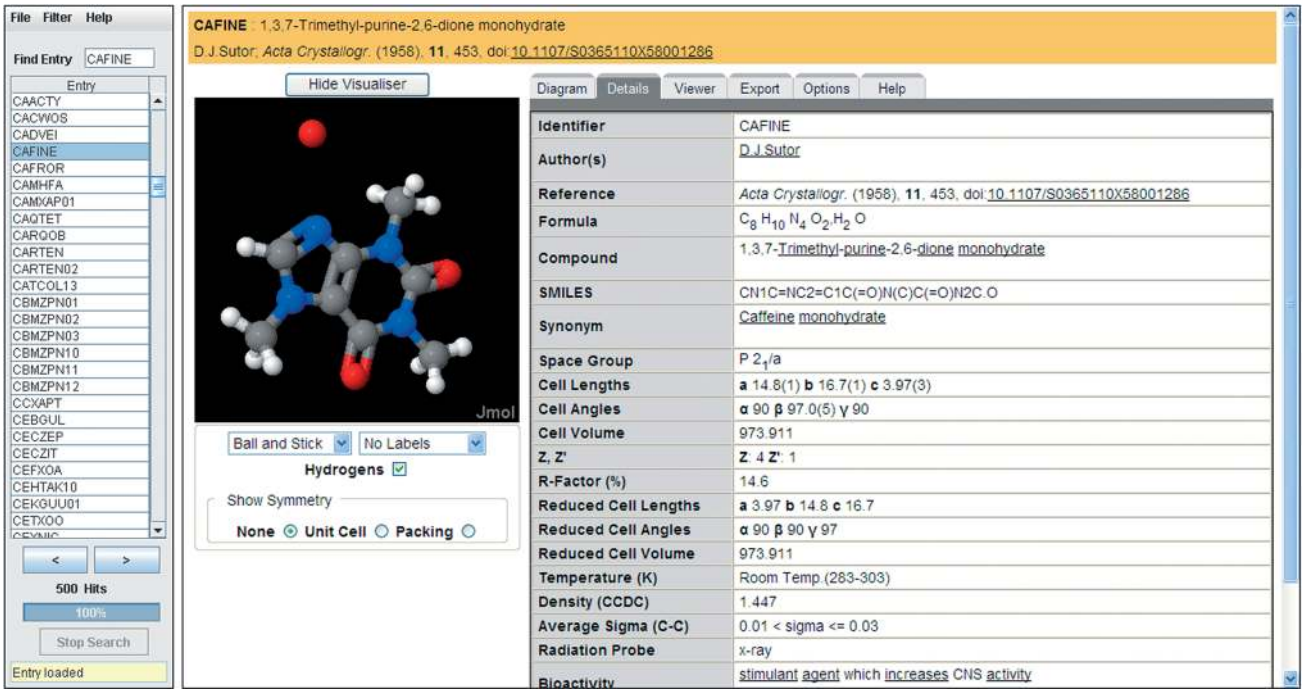

Figure 2. Information pane for caffeine monohydrate (CSD code CAFINE[19]) from the *WebCSD* application.

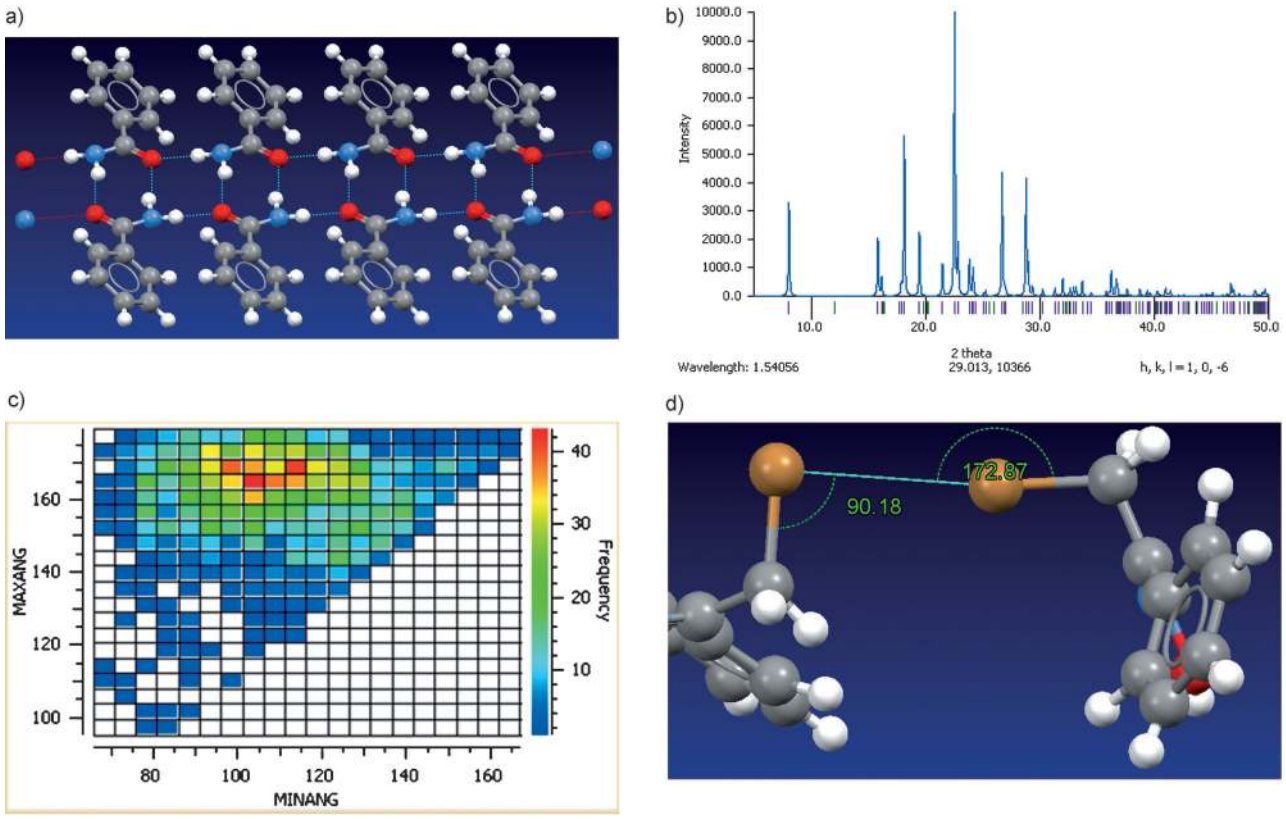

Figure 3. The many faces of *Mercury*: a) extended hydrogen-bonded structure of benzamide (CSD code BZAMID[21]), b) the computed powder pattern for BZAMID,[21] c) heat map of the two C-Hal-Hal angles in halogen–halogen interactions, showing the preference for one angle to be close to 90° and the other close to 180°, as indicated in the example (CSD code ABACOX10)[22] in part (d).

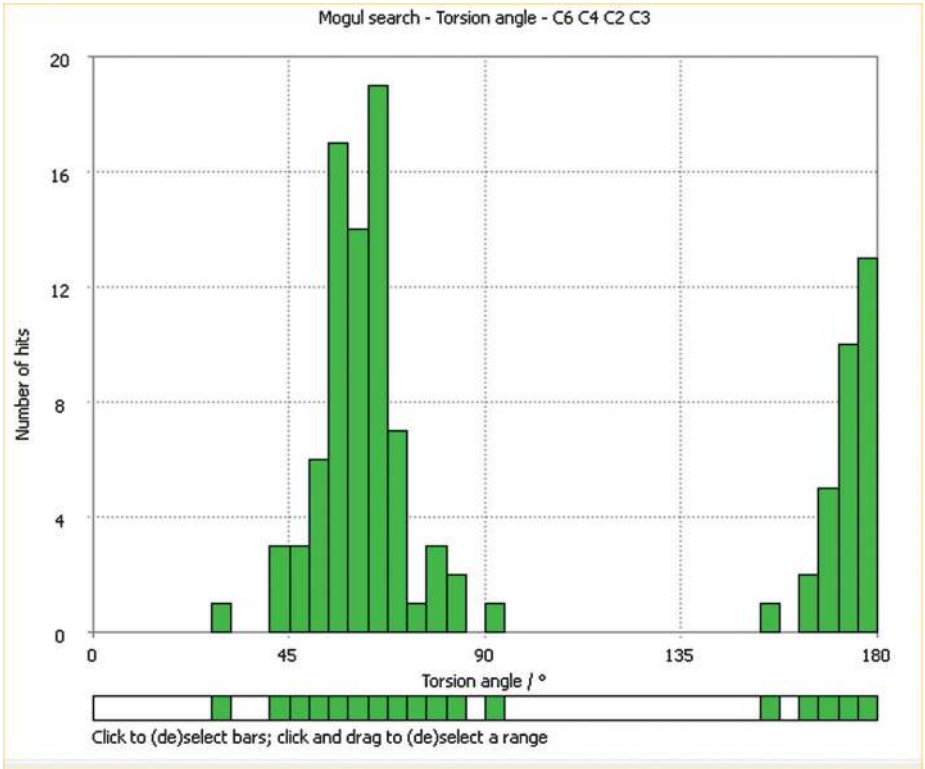

Figure 4. Distribution of C-C-C-C torsion angles in 2,3-dimethylbutane-like fragments, $(C_{sp^3})_2$-CH-CH-$(C_{sp^3})_2$, in the CSD from the *Mogul* knowledge base.

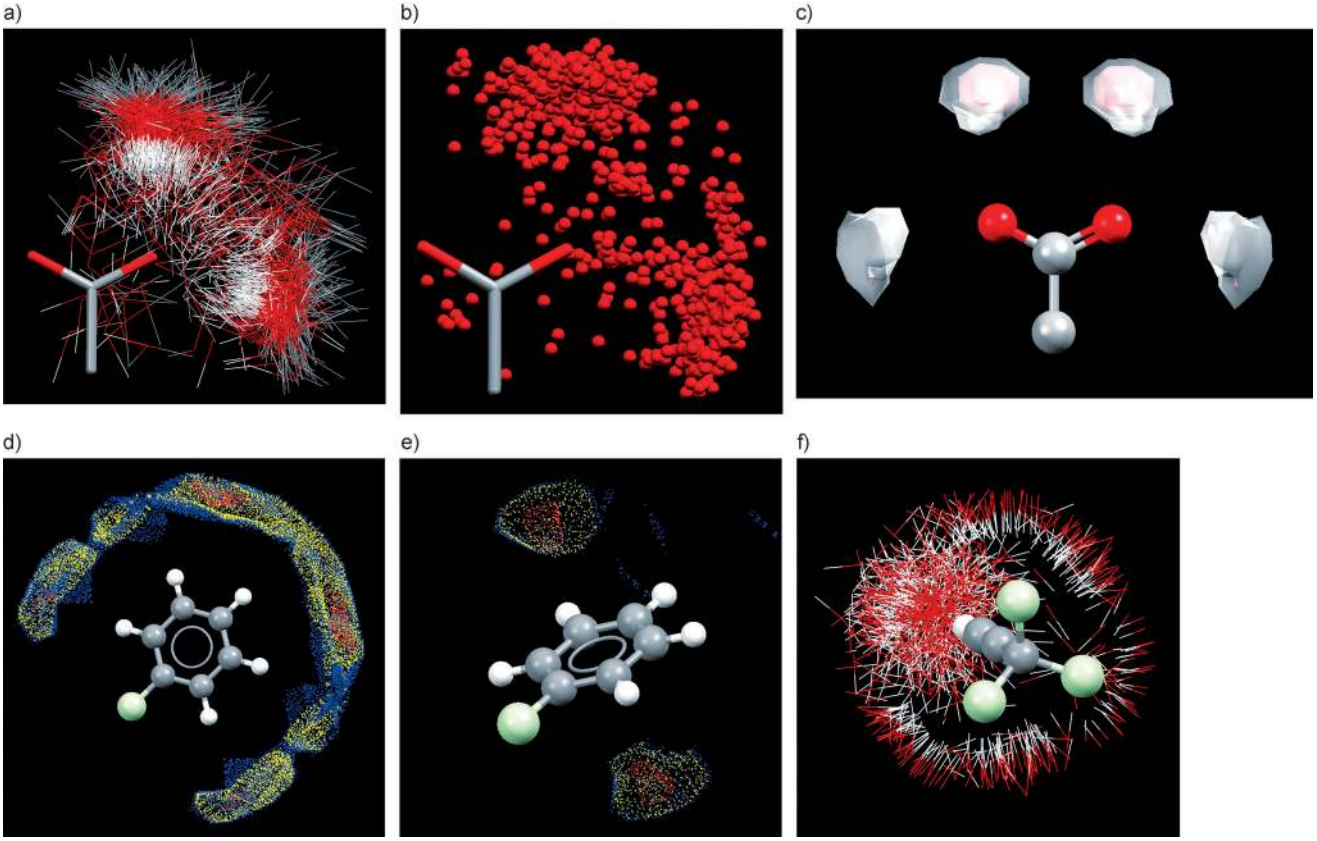

Figure 5. The *IsoStar* knowledge base. Distribution of O–H donors around one of the equivalent oxygen atoms of a charged carboxylate group using a) CSD data and b) PDB data; c) a contoured plot of the full symmetrized distribution around the carboxylate. Distribution of d) carbonyl oxygen O atoms and e) aliphatic C–H atoms around a phenyl ring, and f) the distribution of O–H groups around an ethynyl group C_{sp^3}–C≡C–H. In (f), note the formation of C≡C–H···O bonds along the direction of the ethynyl bond, and the formation of a ring of O–H···π(C≡C) bonds perpendicular to that bond.

proteins.[28] *IsoStar*[29] is a library of graphical and numerical information about nonbonded interactions, providing interactive 3D scatterplots showing the distribution of one of 48 contact groups, for example, an H-bond donor, around a central group, as shown in Figure 5. The 300 central groups cover a wide range of chemical functionality. More than 25500 scatterplots are available, with over 20000 derived from the CSD and more than 5500 from protein–ligand complexes in the PDB. About 1500 ab initio potential energy minima[30] are also included in *IsoStar*.

6. Diversification of CCDC Software

To better engage with the drug discovery and crystallographic user communities, the CCDC has diversified its distributed software, while preserving the essential link to crystal structure data. This software now includes: *GOLD*,[31] a protein–ligand docking program that uses CSD conformations and H-bond information in parameterization and scoring functions; *SuperStar*,[32] which uses experimental knowledge from *IsoStar* to generate maps of interaction sites in protein binding cavities for a selection of functional group probes (Figure 6); *Relibase*,[33] a database system derived from the PDB[6c,d] that permits a wide variety of searches for proteins, ligands, and their interactions; and *DASH*,[34] which uses direct space methods to solve structures from X-ray powder-diffraction data and has direct links to *Mogul*[26] to provide conformational knowledge during model building for more complex structures,[35] thus reducing the search space.

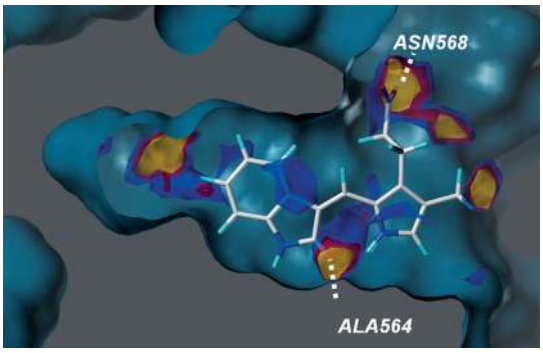

Figure 6. Interaction "hot spots" (shown in gold) generated by *SuperStar* for a C=O group within the binding site of tyrosine kinase, with the experimentally determined position of an oxindole inhibitor superimposed. The correspondence of the two inhibitor C=O groups with two of the *SuperStar* hot spot predictions is clearly observed.

7. The CSD as a Catalyst for Scientific Research

From its earliest days, CSD information has been cited routinely for comparison purposes in crystal structure reports. More fundamentally, the CSD has underpinned much original research in those disciplines where knowledge of chemical structure is critical, and some 3000 papers of this type have appeared in the literature so far[36a] and have been the subject of a recent citation analysis.[36b]

A number of early CSD-based studies concerned reaction pathway analysis and structure correlation,[37a] following on from the original work of Bürgi and Dunitz[37b] who mapped the reaction pathway for attack on a carbonyl center by a nitrogen nucleophile. However, CSD-based research rapidly broadened to incorporate systematic studies of both intra- and intermolecular systems.

An early paper proved the ability of C–H donors to form hydrogen bonds with O, N, and Cl acceptors.[38] This paper was followed by a flood of CSD-based studies of intermolecular interactions, both strong and weak, and conventional and unconventional,[39] all of which which contributed to the formulation of supramolecular synthons[23] and graph-set descriptors[24] of H-bond networks. Also during the early period, existing tables of bond lengths[8] were replaced by definitive standards,[40] and research at the intramolecular level then extended to substituent effects, detailed studies of cyclic and acyclic conformational preferences, and in-depth studies of metal coordination environments. CSD-based research continues to make significant contributions in organic and metal–organic chemistry, crystal engineering, crystal structure prediction, protein–ligand interactions, drug discovery and drug development, and in materials science; these contributions have been well reviewed and exemplified elsewhere.[13,39,41] As an indication of the broad research appeal of the CSD, the American Chemical Society recently announced[42] that the current standard reference to the CSD System[5c] was the most highly cited of all chemistry papers that had been published in 2002.

8. The CCDC as a Scientific Institution

Initially directed towards crystal structure determination, research at the CCDC moved naturally towards applications of the growing database, and the number of papers by CCDC authors is now approaching 800. Some of these papers describe developments in the exported CSD System, for example, the introduction of *Mogul* and *IsoStar* and the validation of *SuperStar* and *GOLD*. Other papers result from the work of associated doctoral students and academic visitors, but the core of CCDC research arises from its own established staff. Current in-house research involves close interactions with industrial partners, most recently addressing specific areas of drug discovery and the solid-state formulation of active pharmaceutical ingredients.[43]

About 90% of small-molecule drug formulations are crystalline and are often delivered as salts or cocrystals with permitted excipients. The CSD is the ultimate library of solid-state forms, and recent research has led to software that enables a wide spectrum of scientists to better understand the formation and stability of crystalline solids. Thus, *Mercury* can now search for supramolecular synthons formed by specified functional groups, generalized packing features, where the user selects a feature of interest in an extended CSD structure, and crystal-packing similarity searches.[20]

Polymorphism can be critical in crystalline drug formulation, and knowledge-based H-bond propensity analysis[44] using CSD data is complementary to experimental polymorph screening. Starting with just a 2D chemical diagram of a target, CSD information from related molecules is used to predict the likelihood of formation of each potential H-bond in its crystal structure. Highly likely or unlikely H-bonds are quickly revealed, pointing to potential stability issues, as exemplified by a study of ritonavir.[45] H-bond propensities can also be used to assess multicomponent crystals, for example, cocrystals or solvates, where the second component often introduces alternative donor–acceptor possibilities.[46]

Ongoing approaches to multicomponent systems also make use of some of the CCDC's drug discovery tools. A recent approach[47] regards supramolecular design as a docking problem, analogous to protein–ligand docking, by treating known stable pure-drug frameworks as hosts into which a variety of potential cocrystal formers are docked using the *GOLD* methodology. In another approach,[48] *IsoStar*,[29] *SuperStar*,[32] and *Mercury*[20] are used to generate full interaction maps of drug molecules with respect to functional group probes, and this method is now being used to study polymorph stability, as shown in Figure 7.

Research to further the aims of drug discovery chemists remains important. Ongoing work addresses a number of challenges relating to the optimization of molecular conformations, always with reference to knowledge extracted from the CSD. Problems such a protein–ligand docking, molecular superposition, pharmacophore determination, and conformer generation all rely on optimizing the fit of a molecule to a target function, whilst ensuring a plausible molecular conformation.

9. The Future

The CSD is the central core of our activities, and the CCDC is now well-equipped to assimilate continued worldwide increases in crystal structure productivity but without loss of data quality. New data processing systems, noted earlier, open up new horizons. Freed from historical format restrictions, the information content of the CSD can now expand to incorporate atomic displacement parameters, improved descriptions of disorder, metal oxidation states, and additional data items now routinely available in deposited CIFs. Together with relevant technological advances, this opens up new possibilities for CSD System software, with significant extensions to search capabilities and improvements to structure visualization and data analysis. Software will also benefit from the ongoing research into the solid form, which is relevant not only in drug development but also across a broad spectrum of solid-state studies, while the CCDC's work on knowledge-based software solutions in drug

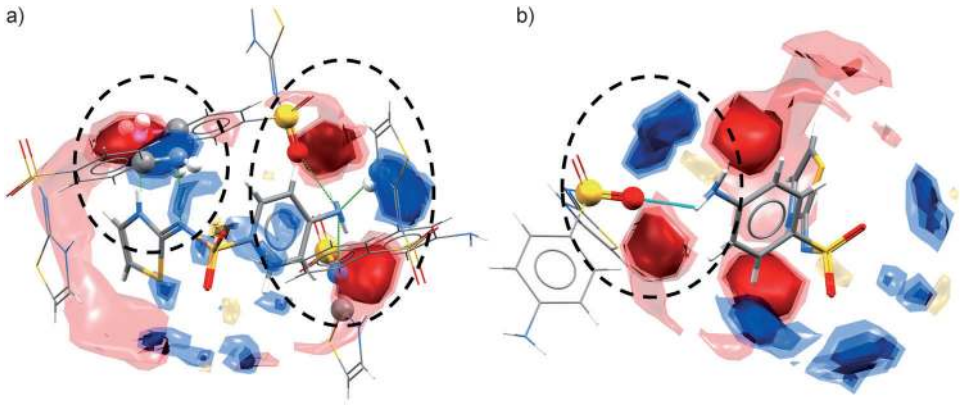

Figure 7. Full interaction maps for sulfathiazole shown within the packing diagrams of a) Polymorphic Form V (SUTHAZ19[49]) and b) Polymorphic Form I (SUTHAZ16[49]). Preferred acceptor positions are in red, donor positions are in blue. In (a), each of the strongest interaction map peaks have a matching donor or acceptor atom within their contours, but in (b) one of the acceptors is outside the closest interaction map peak. An unsatisfied donor or acceptor is a likely sign of metastability, evidenced here in Form I. For full details, see reference [48].

discovery is also expanding, through provision of a new conformer generator and the optimization of pharmacophoric pattern recognition.

One of the most significant issues of the future concerns calls for free and open access (OA) to crystal structure data. Over the last 15–20 years, the raw supplementary CIFs associated with crystal structure publications have become ever more freely available, either through the relevant journal or through the CCDC's free "request a structure" CIF service.[50] This has encouraged the creation of two CIF collections: the Crystallography Open Database (COD)[51a,b] containing donated and downloaded CIFs, and CrystalEye[51c,d] containing CIFs harvested automatically from the Web (although this collection appears not to have been updated since mid-2011). Both of these collections have received external funding, often time-limited, from various agencies, so they are not "free" in the absolute sense. In all OA paradigms somebody pays, usually a funding agency, an institution, or the authors themselves, so as to make the output "free at the point of access" for the reader or user.

The CIFs in these collections are a subset of the CSD, but onward conversion into a fully retrospective, fully comprehensive, and scientifically curated database, with high-quality access software and support services for depositors and users alike, requires a financially stable and permanent organization. Funding from a single national government, or group of governments as in the EU, relies on the "generosity" of a specific set of taxpayers and poses risks during economic downturns, or when it is deemed that available funds are better directed to other projects of national or international importance. These are very real risks, and while spreading the funding net as wide as possible does not eliminate risk, it does reduce it considerably. In the CCDC's case, its non-profit constitution, with international financial contributions from the user community in both academia and industry, may still represent the most viable solution for the permanent maintenance of a specialist scientific resource. The word "permanent" is important here: even a short-term loss of funding can seriously impair the core objectives of a data center such as the CCDC. Those objectives must be to maintain and disseminate a database system which is as complete and accurate as possible so that the scientific integrity of the historical record and its value to the community are maximized.[52] These may be high, and perhaps unattainable ideals, but we should at least aspire to them.

10. Conclusions

We conclude with another anniversary: in 2015 the CCDC will celebrate 50 years of service to the scientific community, almost exactly half of the century since the determination of that first crystal structure.[1] During those 50 years, the CCDC has addressed the scientific, technical, and financial issues set out in the original 1964 invitation to establish the Data Centre. During those 50 years, the choice of small-molecule crystallography has been more than vindicated: as was realized by 1929, crystallography is indeed a special analytical technique, and the value of its data to the scientific community has broadened in every decade since the 1960s. Not only was the invitation scientifically visionary, it also came at exactly the right moment: if the CCDC had not started work within a rather small window of opportunity in the mid-1960s, it may not have started at all. The number of published crystal structures was such that a database was of immediate value, but the number was not large enough to present an insurmountable backlog of existing structures to process. However this would soon have been the case given the rapid increase in the number of structures determined (see Figure 2). Thus, start-up workload and expenses would have increased significantly for every year of delay beyond 1970.[52]

Most importantly the CSD has benefited from the long-term support of the crystallographic community—to have the coordinates of every organic and metal–organic compound for which a crystal structure has been published available in

a single database is something that the discipline is rightly proud of.

Received: July 24, 2013

[1] a) W. L. Bragg, *Proc. Cambridge Philos. Soc.* **1912**, *17*, 43 – 57; b) W. L. Bragg, *Proc. R. Soc. London Ser. A* **1913**, *89*, 248 – 277; c) W. H. Bragg, *Proc. R. Soc. London Ser. A* **1913**, *89*, 430 – 438.

[2] a) P. P. Ewald, C. Hermann, *Strukturbericht 1913 – 1928*, Akademische Verlagsgesellschaft, Leipzig, **1929**; b) *Structure Reports 1940 – 1950*, *Vol. 8 – 13* (Eds.: A. J. C. Wilson, N. C. Baenziger, J. M. Bijvoet, J. M. Robertson), Reidel, Dordrecht, **1990**. See http://www.iucr.org/books/structure-reports (accessed 3 July **2013**).

[3] L. Pauling, *J. Am. Chem. Soc.* **1929**, *51*, 1010 – 1026.

[4] http://www.iucr.org/people/nobel-prize (accessed July 3, **2013**).

[5] a) F. H. Allen, S. Bellard, M. D. Brice, B. A. Cartwright, A. Doubleday, H. Higgs, T. Hummelink, B. G. Hummelink-Peters, O. Kennard, W. D. S. Motherwell, J. R. Rodgers, D. G. Watson, *Acta Crystallogr. Sect. B* **1979**, *35*, 2332 – 2339; b) F. H. Allen, J. E. Davies, J. J. Galloy, O. Johnson, O. Kennard, C. F. Macrae, E. M. Mitchell, G. F. Mitchell, J. M. Smith, D. G. Watson, *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 187 – 204; c) F. H. Allen, *Acta Crystallogr. Sect. B* **2002**, *58*, 380 – 388.

[6] a) ICSD: A. Belsky, M. Hellenbrandt, V. L. Karen, P. Luksch, *Acta Crystallogr. Sect. B* **2002**, *58*, 364 – 369; b) CRYSTMET: P. S. White, J. R. Rodgers, Y. Le Page, *Acta Crystallogr. Sect. B* **2002**, *58*, 343 – 348; c) H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235 – 242; d) http://www.wwpdb.org/ (accessed July 17, **2013**); e) NDB: H. M. Berman, J. Westbrook, Z. Feng, L. Iype, B. Schneider, C. Zardecki, *Acta Crystallogr. Sect. D* **2002**, *58*, 889 – 898.

[7] J. D. H. Donnay, G. Donnay, E. G. Cox, O. Kennard, M. V. King, *Crystal Data*: *Determinative Tables*, American Crystallographic Association, **1963**.

[8] L. E. Sutton, *Tables of Interatomic Distances in Molecules and Ions*, The Chemical Society, London, Special Publications Nos. 11 and 18, **1959** and **1965**.

[9] a) The Royal Society Scientific Information Conference: report and papers submitted, The Royal Society, London, **1948**; b) see also: B. Vickery, *J. Doc.* **1998**, *54*, 281 – 283.

[10] O. Kennard, personal communication of lecture notes entitled *In the Beginning was Bernal*, **1988**.

[11] a) F. H. Allen, O. Kennard, W. D. S. Motherwell, W. G. Town, D. G. Watson, T. J. Scott, A. C. Larson, *J. Appl. Crystallogr.* **1974**, *7*, 73 – 78; b) F. H. Allen, O. Kennard, D. Watson, K. M. Crennell, *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 129 – 139; c) F. H. Allen, *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 68 – 76.

[12] a) *Molecular Structures and Dimensions*, *Bibliographic Vol. 1 – 16 and Vol. A1* (variously edited by O.Kennard, D. G. Watson and others), Reidel, Dordrecht, **1970 – 1986**; b) G. A. Jeffrey, *Acta Crystallogr. Sect. B* **1978**, *34*, 3847.

[13] a) K. A. Brameld, B. Kuhn, D. C. Reuter, M. Stahl, *J. Chem. Inf. Model.* **2008**, *48*, 1 – 24; b) C. Bissantz, B. Kuhn, M. Stahl, *J. Med. Chem.* **2010**, *53*, 5061 – 5084; c) B. Kuhn, P. Mohr, M. Stahl, *J. Med. Chem.* **2010**, *53*, 2601 – 2611.

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

[15] http://journals.iucr.org/e/services/authorservices.html (accessed July 3, **2013**).

[16] F. H. Allen, O. Johnson, G. P. Shields, B. R. Smith, M. Towler, *J. Appl. Crystallogr.* **2004**, *37*, 335 – 338.

[17] I. J. Bruno, G. P. Shields, R. Taylor, *Acta Crystallogr. Sect. B* **2011**, *67*, 333 – 349.

[18] a) I. J. Bruno, J. C. Cole, P. R. Edgington, M. Kessler, C. F. Macrae, P. McCabe, J. Pearson, R. Taylor, *Acta Crystallogr. Sect. B* **2002**, *58*, 389 – 397; b) I. R. Thomas, I. J. Bruno, J. C. Cole, C. F. Macrae, E. Pidcock, P. A. Wood, *J. Appl. Crystallogr.* **2010**, *43*, 362 – 366.

[19] D. J. Sutor, *Acta Crystallogr.* **1958**, *11*, 453 – 458.

[20] a) C. F. Macrae, P. R. Edgington, P. McCabe, E. Pidcock, G. P. Shields, R. Taylor, M. Towler, J. van de Streek, *J. Appl. Crystallogr.* **2006**, *39*, 453 – 457; b) C. F. Macrae, I. J. Bruno, J. A. Chisholm, P. R. Edgington, P. McCabe, E. Pidcock, L. Rodriguez-Monge, R. Taylor, J. van de Streek, P. A. Wood, *J. Appl. Crystallogr.* **2008**, *41*, 466 – 470.

[21] B. R. Penfold, J. C. B. White, *Acta Crystallogr.* **1959**, *12*, 130 – 135.

[22] J. B. Wetherington, J. W. Moncrief, *Acta Crystallogr.* **1973**, *29*, 1520 – 1525.

[23] a) G. R. Desiraju, *Angew. Chem.* **1995**, *107*, 2541 – 2558; *Angew. Chem. Int. Ed. Engl.* **1995**, *34*, 2311 – 2327; b) G. R. Desiraju, *J. Am. Chem. Soc.* **2013**, *135*, 9952 – 9967.

[24] J. Bernstein, R. E. Davis, L. Shimoni, N.-L. Chang, *Angew. Chem.* **1995**, *107*, 1689 – 1708; *Angew. Chem. Int. Ed. Engl.* **1995**, *34*, 1555 – 1573.

[25] R. A. Sykes, P. McCabe, F. H. Allen, G. M. Battle, I. J. Bruno, P. A. Wood, *J. Appl. Crystallogr.* **2011**, *44*, 882 – 886.

[26] a) 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, A. G. Orpen, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2133 – 2144; b) S. J. Cottrell, T. S. G. Olsson, R. Taylor, J. C. Cole, J. W. Liebeschuetz, *J. Chem. Inf. Model.* **2012**, *52*, 956 – 962.

[27] D. J. Watkin, R. J. Cooper, http://www.xtl.ox.ac.uk/crystals.html (accessed July 3, **2013**).

[28] J. W. Liebeschuetz, J. Hennemann, T. S. G. Olsson, C. R. Groom, *J. Comput.-Aided Mol. Des.* **2012**, *26*, 169 – 183.

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

[30] I. C. Hayes, A. J. Stone, *J. Mol. Phys.* **1984**, *53*, 83 – 105.

[31] a) G. Jones, P. Willett, R. C. Glen, A. R. Leach, R. Taylor, *J. Mol. Biol.* **1997**, *267*, 727 – 748; b) M. L. Verdonk, J. C. Cole, M. J. Hartshorn, C. W. Murray, R. D. Taylor, *Proteins* **2003**, *52*, 609 – 623; c) J. W. Liebeschuetz, J. C. Cole, O. Korb, *J. Comput.-Aided Mol. Des.* **2012**, *26*, 737 – 748.

[32] a) M. L. Verdonk, J. C. Cole, R. Taylor, *J. Mol. Biol.* **1999**, *289*, 1093 – 1108; b) M. L. Verdonk, J. C. Cole, P. Watson, V. Gillet, P. Willett, *J. Mol. Biol.* **2001**, *307*, 841 – 859; c) J. W. M. Nissink, C. W. Murray, M. J. Hartshorn, M. L. Verdonk, J. C. Cole, R. Taylor, *Proteins* **2002**, *49*, 457 – 471.

[33] a) M. Hendlich, A. Bergner, J. Guenther, G. Klebe, *J. Mol. Biol.* **2003**, *326*, 607 – 620; b) J. Guenther, A. Bergner, M. Hendlich, G. Klebe, *J. Mol. Biol.* **2003**, *326*, 621 – 636.

[34] W. I. F. David, K. Shankland, J. van de Streek, E. Pidcock, W. D. S. Motherwell, J. Cole, *J. Appl. Crystallogr.* **2006**, *39*, 910 – 915.

[35] E. Pidcock, J. van de Streek, M. U. Schmidt, *Z. Kristallogr.* **2007**, *222*, 713 – 717.

[36] a) http://www.ccdc.cam.ac.uk/ResearchAndConsultancy/CCDCResearch/Pages/WebCite.aspx (accessed July 3, **2013**); b) R. Wong, F. H. Allen, P. Willett, *J. Appl. Crystallogr.* **2010**, *43*, 811 – 824.

[37] a) H.-B. Bürgi, J. D. Dunitz, *Structure Correlation*, VCH, Weinheim, **1997**; b) H.-B. Bürgi, J. D. Dunitz, *Acc. Chem. Res.* **1983**, *16*, 153 – 161.

[38] a) R. Taylor, O. Kennard, *Acc. Chem. Res.* **1984**, *17*, 320 – 326. Fascinating historical accounts of the controversies that existed prior to the appearance of this paper have recently been given by: b) C. Schwalbe, *Cryst. Rev.* **2012**, *18*, 191 – 206, and c) J. Bernstein, *Cryst. Growth Des.* **2013**, *13*, 961—964.

[39] See, for example: a) G. A. Jeffrey, W. Saenger, *Hydrogen Bonding in Biological Structures*, Springer, Berlin, **1991**; b) G. R. Desiraju, T. Steiner, *The Weak Hydrogen Bond in Structural Chemistry and Biology*, Oxford University Press, Oxford, **1999**; c) M. Nishio, *Phys. Chem. Chem. Phys.* **2011**, *13*, 13873 – 13900.

[40] a) F. H. Allen, O. Kennard, D. G. Watson, L. Brammer, A. G. Orpen, 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, R. Taylor, *J. Chem. Soc. Dalton Trans.* **1989**, S1 – S83.

[41] a) F. H. Allen, W. D. S. Motherwell, *Acta Crystallogr. Sect. B* **2002**, *58*, 407 – 422; b) R. Taylor, *Acta Crystallogr. Sect. D* **2002**, *58*, 879 – 888; c) F. H. Allen, J. A. Chisholm, P. A. Wood, P. T. A. Galek, L. Fábián, O. Korb, A. J. Cruz-Cabeza, J. W. Liebeschuetz, C. R. Groom, E. Pidcock, *Supramolecular Chemistry: From Molecules to Nanomaterials*, Wiley, Chichester, **2012**, pp. 2927 – 2946.

[42] E. K. Wilson, *Chem. Eng. News* **2012**, *90*, 39 – 40.

[43] P. T. A. Galek, E. Pidcock, P. A. Wood, I. J. Bruno, C. R. Groom, *CrystEngComm* **2012**, *14*, 2391 – 2403.

[44] a) P. T. A. Galek, L. Fábián, W. D. S. Motherwell, F. H. Allen, N. Feeder, *Acta Crystallogr. Sect. B* **2007**, *63*, 768 – 782; b) P. T. A. Galek, L. Fábián, F. H. Allen, *CrystEngComm* **2010**, *12*, 2091 – 2099; c) P. T. A. Galek, L. Fábián, F. H. Allen, *Acta Crystallogr. Sect. B* **2010**, *66*, 237 – 252.

[45] P. T. A. Galek, F. H. Allen, L. Fábián, N. Feeder, *CrystEngComm* **2009**, *11*, 2634 – 2639.

[46] a) A. Delori, P. T. A. Galek, E. Pidcock, W. Jones, *Chem. Eur. J.* **2012**, *18*, 6835 – 6846; b) A. Delori, P. T. A. Galek, E. Pidcock, M. Patni, W. Jones, *CrystEngComm* **2013**, *15*, 2916 – 2928; c) M. Majumder, G. Buckton, C. Rawlinson-Malone, A. C. Williams, M. J. Spillman, E. Pidcock, K. Shankland, *CrystEngComm* **2013**, *15*, 4041 – 4044.

[47] O. Korb, P. A. Wood, *Chem. Commun.* **2010**, *46*, 3318 – 3320.

[48] P. A. Wood, T. S. G. Olsson, J. C. Cole, S. J. Cottrell, N. Feeder, P. T. A. Galek, C. R. Groom, E. Pidcock, *CrystEngComm* **2013**, *15*, 65 – 72.

[49] T. Gelbrich, D. S. Hughes, M. B. Hursthouse, T. L. Threlfall, *CrystEngComm* **2008**, *10*, 1328 – 1334.

[50] http://www.ccdc.cam.ac.uk/Community/Requestastructure/pages/Requestastructure.aspx (accessed July 9, **2013**).

[51] a) http://www.crystallography.net/(accessed July 9, **2013**); b) S. Grazulis, D. Chateigner, R. T. Downs, A. F. T. Yokochi, M. Quirós, L. Lutterotti, E. Manakova, J. Butkus, P. Moeck, A. Le Bail, *J. Appl. Crystallogr.* **2009**, *42*, 726 – 729; c) http://wwmm.ch.cam.ac.uk/crystaleye/index.html (accessed July 9, **2013**); d) N. Day, J. Downing, S. Adams, N. W. England, P. Murray-Rust, *J. Appl. Crystallogr.* **2012**, *45*, 316 – 323.

[52] F. H. Allen, R. Taylor, *Chem. Commun.* **2005**, 5135 – 5140.