Acta Cryst B
STRUCTURAL SCIENCE
CRYSTAL ENGINEERING
MATERIALS

ISSN 2052-5206

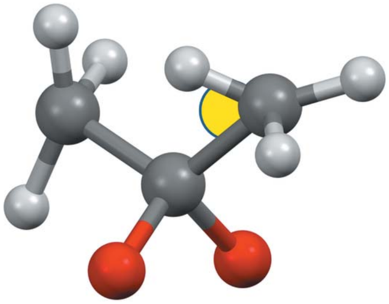
The development and use of a crystallographic database

Carl Henrik Görbitz*

Department of Chemistry, University of Oslo, PO Box 1033 Blindern, N-0315 Oslo, Norway. *Correspondence e-mail: c.h.gorbitz@kjemi.uio.no

Received 29 March 2016
Accepted 30 March 2016

Keywords: Cambridge Structural Database; CIF archive; open data; crystal structure database.



Crystallographers constitute a privileged group of scientists, not only because they provide an understanding of nature to an extent and at a resolution that is truly unique, but also because their results are archived in a manner that is unrivaled among physical methods used for characterization of molecular properties. This archive, the Cambridge Structural Database (CSD), developed and maintained by the Cambridge Crystallographic Data Center (CCDC), presently holds more than 800 000 entries for organic and organometallic compounds. Other databases include the Protein Data Bank (PDB; Berman *et al.*, 2000) and the Inorganic Crystal Structure Database (ICSD; Belsky *et al.*, 2002). The value of this resource is documented by the fact that the currently used reference to the CSD (Allen, 2002) has received more than 10 000 citations related to all fields of crystallography (Wong *et al.*, 2010). The paper by Groom *et al.* (2016) supersedes Allen's article and will become the new standard reference to the CSD.

The authors describe the exponential growth in the number of entries in the CSD since the humble beginnings in 1965, and provide details in terms of the historical and gradual development of molecular complexity and size, journal used for publication and more. Two distinct ways in which the CSD provides value are pointed out: 'aggregation and standardization of structures, which facilitates access to individual entries', and 'the study of the collection of entries' (Kennard, 1997) related to the geometry of molecules and the interactions they make. Some key papers discuss structure correlation (Bürgi & Dunitz, 1983), C—H groups as donors (Taylor & Kennard, 1982) and geometric tables (Allen *et al.*, 1987; Orpen *et al.*, 1989). A recent example includes quantification of the symmetry preferences of intermolecular interactions in organic crystal structures (Taylor *et al.*, 2015).

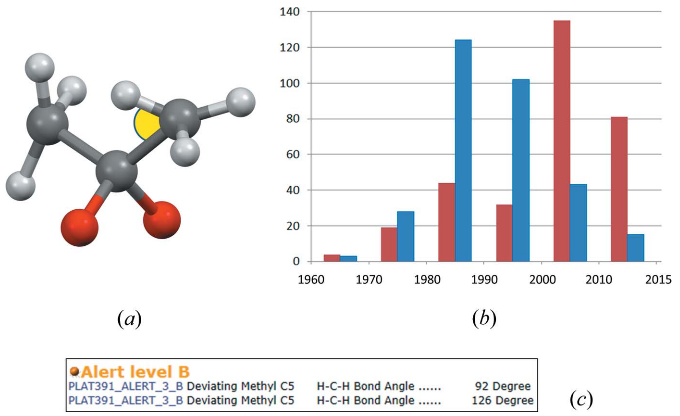
The process from deposition of a data set to a published paper is discussed in detail. In fact, most journals these days require that any crystallographic material has been deposited at the CCDC before the manuscript is submitted for review; individual structures subsequently being identified by their CCDC deposition numbers in the printed paper. Publication triggers immediate public release of the corresponding structure(s). Retrieval of data is then open to anyone through requests posted at the CCDC web site. The authors also point out that many structures are now published only and directly as CSD Communications (previously known as Private Communications) and foresee that this will soon become the most popular way in which to publish crystal structures. After discussing how the CSD is used, they outline future developments, including systems that handle structures derived from non-crystallographic sources, such as electron diffraction, atomic force microscopy, free electron lasers and NMR crystallography, but also from crystal structure prediction algorithms.

The tremendous success of the CSD has several, more or less independent components:

(i) Since each atomic position is defined by a set of three coordinates, crystallography is by its very nature amenable to digitalization.

(ii) Computer programs have been developed to help the crystallographer during the refinement and in particular deal with the positions of H atoms.

(iii) The introduction of the CIF file format (now the Crystallographic Information Framework syntax; Brown & McMahon, 2002; Bernstein *et al.*, 2016; Hall & McMahon, 2016) has not only facilitated online checking of data, in particular *checkCIF* (IUCr, 2016), before publication, but has also, by contributing to an all-digital flow of data, eliminated several sources of error associated with manual punching of coordinates *etc.*



Alert level B
PLAT391_ALERT_3_B Deviating Methyl C5 H-C-H Bond Angle ...... 92 Degree
PLAT391_ALERT_3_B Deviating Methyl C5 H-C-H Bond Angle ...... 126 Degree

Figure 1
(*a*) Example of inferior methyl group geometries in a CSD structure (which is to remain anonymous; only part of the molecule is shown). The highlighted C—C—H angle is 67.5°. The same methyl group also has a 153.7° C—C—H angle. (*b*) Distribution of C—C—H angles $< 75°$ (blue bars) and water H—O—H angles $> 135°$ (red bars) in CSD structures as a function of decennium. (*c*) Typical *checkCIF* warning for a methyl group with suspicious geometry. In this case one methyl H atom, originally positioned by a *SHELXL* (Sheldrick, 2015) AFIX 33 command, was manually shifted slightly away from its theoretical position.

(iv) Last, but not least, the continuous efforts made by the CCDC staff to develop the CSD and the associated software to become an indispensable tool for small molecule crystallographers.

As an example of this development, I searched the database for organic structures in which a methyl group has (at least) one C—C—H angle $< 75°$. A total of 349 such physically unrealistic narrow angles were found in 315 structures, one example being illustrated in Fig. 1(*a*). Sorted by year, the distribution in Fig. 1(*b*) shows that the number of such freak geometries has declined dramatically since the 1980s. There are two obvious reasons: the first is that most people now fix methyl groups in theoretical, staggered positions, the second that updated *checkCIF* routines issue a set of error messages when such geometries are encountered, Fig. 1(*c*). The fact that the number has still not dropped to zero does, however, raise some concern. Evidently, some structures are published by researchers who fail to use the refinement programs properly and care little about checking their results. Furthermore, some reviewers take their role too lightly and do not discover and address Alert level B errors like those in Fig. 1(*c*). From my own experience as a reviewer I find that water molecules are particularly prone to error, and the frequency of wide H—O—H bond angles $> 135°$ appears to still be on the rise, Fig. 1(*b*). More rigorous *checkCIF* algorithms will undoubtedly be available in the future, but in the end it is the responsibility of both authors and reviewers to correct such obvious errors before the structure enters the CSD.

I was myself introduced to the CSD when I worked with my masters thesis in the mid 1980s, but I used it in full for the first time a few years later in preparation of a paper on the hydrogen-bond distances and angles in the structures of amino acids and peptides (Görbitz, 1989). At the time elucidating information on intermolecular interactions was quite an undertaking, above all due to the obvious fact that we did not have graphical computers at the time (the first graphical CSD interfaces appeared in 1991, ‘modern’ interfaces arrived in 2002 with *ConQuest*; Bruno *et al.*, 2002; and *Mercury*; Macrae *et al.*, 2006). Lacking any visual input or output, making sure that you had found what you intended to find required excessively time-consuming manual checking of intricate tables of molecular connectivities. Also, although the investigation was carried out only on a small subset of 749 amino acid and peptide structures extracted from the 67 000 structures in the CSD at the time, the generation of neighbors for calculating intermolecular interaction geometries exhausted our computer resources to the extent that a simple search would run overnight. The change compared to the way the CSD is used today is simply incredible. And evidently, according to Groom *et al.* (2016), it is going to get even better. Lucky crystallographers!

References

Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.
Allen, F. H., Kennard, O., Watson, D. G., Brammer, L., Orpen, A. G. & Taylor, R. J. (1987). *J. Chem. Soc. Perkin Trans. 2*, pp. S1–S19.
Belsky, A., Hellenbrandt, M., Karen, V. L. & Luksch, P. (2002). *Acta Cryst.* B**58**, 364–369.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Bernstein, H. J., Bollinger, J. C., Brown, I. D., Gražulis, S., Hester, J. R., McMahon, B., Spadaccini, N., Westbrook, J. D. & Westrip, S. P. (2016). *J. Appl. Cryst.* **49**, 277–284.
Brown, I. D. & McMahon, B. (2002). *Acta Cryst.* B**58**, 317–324.
Bruno, I. J., Cole, J. C., Edgington, P. R., Kessler, M., Macrae, C. F., McCabe, P., Pearson, J. & Taylor, R. (2002). *Acta Cryst.* B**58**, 389–397.
Bürgi, H. B. & Dunitz, J. D. (1983). *Acc. Chem. Res.* **16**, 153–161.
Görbitz, C. H. (1989). *Acta Cryst.* B**45**, 390–395.
Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. (2016). *Acta Cryst.* B**72**, 171–179.
Hall, S. R. & McMahon, B. (2016). *Data Sci. J.* **15**, 1–15.
IUCr (2016). *checkCIF*, http://checkcif.iucr.org/.
Kennard, O. (1997). *The Impact of Electronic Publishing on the Academic Community*, pp. 159–166. London: Portland Press Ltd.
Macrae, C. F., Edgington, P. R., McCabe, P., Pidcock, E., Shields, G. P., Taylor, R., Towler, M. & van de Streek, J. (2006). *J. Appl. Cryst.* **39**, 453–457.
Orpen, A. G., Brammer, L., Allen, F. H., Kennard, O., Watson, D. G. & Taylor, R. (1989). *J. Chem. Soc. Dalton Trans.* pp. S1–S83.
Sheldrick, G. M. (2015). *Acta Cryst.* C**71**, 3–8.
Taylor, R., Allen, F. H. & Cole, J. C. (2015). *CrystEngComm*, **17**, 2651–2666.
Taylor, R. & Kennard, O. (1982). *J. Am. Chem. Soc.* **104**, 5063–5070.
Wong, R., Allen, F. H. & Willett, P. (2010). *J. Appl. Cryst.* **43**, 811–824.