CSD Lecture & Tutorial – Knowledge-based methods for structural analysis

Peter Wood
CCDC; wood@ccdc.cam.ac.uk

Introduction

The physical properties of a material depend on the nature and mutual arrangement of its constituents. In crystalline materials these constituents, usually molecules or ions, are arranged in essentially infinite, repeating three-dimensional (3D) patterns determined by space-group symmetry. However, these same constituents can typically adopt multiple distinct 3D patterns to form different crystal structures – the phenomenon of polymorphism [1] – or include additional components to form co-crystals, salts or solvates. All of these different forms can exhibit different physicochemical properties including stability, solubility, photo-reactivity, melting point, crystal shape and many more.

An effective approach for understanding and examining structural stabilities, polymorph likelihoods and other structural relationships involves analysis of existing crystal structures of compounds that are similar in some way to the compounds under study. All the crystal structures of organic and metal-organic species that have been reported in the literature have been collected, curated and organised by the Cambridge Crystallographic Data Centre (CCDC) to form the >850,000 entries in the Cambridge Structural Database (CSD; Figure 1) [2]. Thus the CSD contains millions of discrete pieces of information about intramolecular geometry, as well as similarly extensive information on the intermolecular interactions of atoms and chemical functional groups.

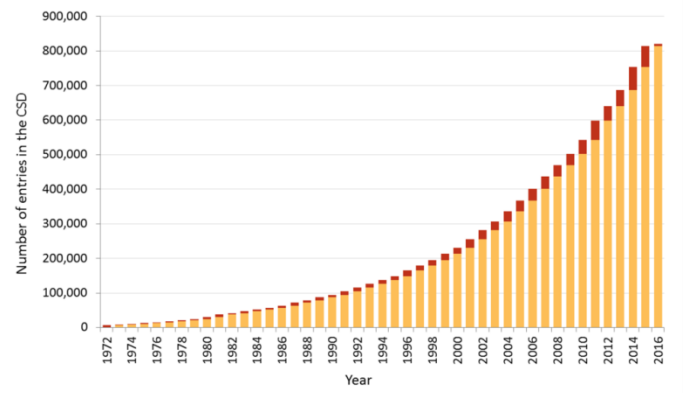

Figure 1. Growth of entries in the CSD 1970-2016

Software provided alongside the CSD (see Section 2) allows easy access to all data, particularly to distributions of geometrical parameters, both bonded and non-bonded, and to the frequencies of occurrence of a wide variety of functional group interactions. Research applications of the CSD have generated some 3,000 publications since the late 1970s, and many of these applications are reviewed elsewhere [3].

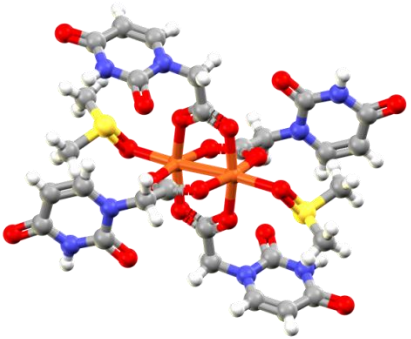

Figure 2. CSD refcode TUWMOP – the 800,000th entry in the CSD

By evaluating a structure in the context of existing knowledge in the CSD, it is relatively straightforward to identify both common and unusual structural features, *e.g.* an unusual conformation of a molecule, ring or functional group, geometrically unusual hydrogen bonded interaction, or an unusual donor-acceptor combination. Unusual features might suggest problems with structural refinement, or identify structural instabilities indicating that alternative crystal forms where molecules aggregate without these compromises might possibly exist [4].

Similarly, analysis of the CSD can quickly identify what is typical for a certain intramolecular or intermolecular geometry. This is an ideal approach, for example, to learn the typical coordination number and geometry for a given transition metal interacting with specific ligand types. Comparative CSD analysis can give answers easily and quickly and can influence the direction of experimental research.

Overview of CSD-Enterprise

For all academic users, a CSD subscription now includes access to the CCDC's entire range of structure-based analysis, visualization and communication applications. This includes all of the following packages:

- **CSD-System** – The CSD database itself and the core set of analysis applications for all scientists involved in any area of structural chemistry
- **CSD-Discovery** – Our specialist set of applications for discovery chemistry including protein visualisation, searching and docking as well as drug design
- **CSD-Materials** – Our specialist set of applications for users studying the structure and properties of crystalline materials

Molecular Geometry

The understanding of molecular geometry is a key application of the CSD across the breadth of structural chemistry. The most flexible approach for analysing or predicting molecular geometries is to define and search for substructures using *ConQuest* and then analyse the results manually using an application like the **Data Analysis** function in *Mercury*. A number of illustrations of this approach are shown elsewhere by Sykes and co-workers [5], including use of Principal Components Analysis (PCA) for cases when there are too many potentially correlated variables to plot on a single chart.

Probably the most common way to analyse molecular geometry using the CSD, however, is to utilise *Mogul*, the CSD's pre-derived knowledge-base of intramolecular geometry [6]. Here all the bond lengths, valence angles, torsion angles and rings in the CSD have been

analysed and grouped into chemically equivalent histograms for easy use. This approach makes it very easy to either analyse a known structure, from *Mercury*, or draw a new molecule in *Mogul* and determine the geometrical preferences.
Using the same knowledge-base of conformational information – *Mogul* – it is even possible to generate the mostly likely conformers for a given molecule based on known experimental data. The **Conformer Generator** application [7] is driven by *Mogul* data and can be used to generate an ensemble of likely conformers (Figure 3) from a 3D molecular input as well as score the conformers in terms of a likelihood score. This application can be accessed within *Mercury* under **CSD-Discovery** or **CSD-Materials**.

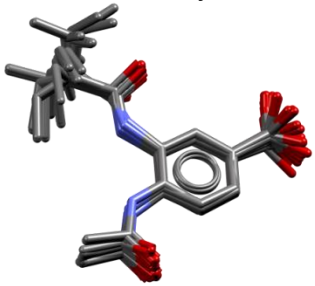

Figure 3. An ensemble of likely conformers generated using knowledge-based methods

Intermolecular Interactions
As with molecular geometries, it is possible to analyse and predict intermolecular interactions through a sketched substructure search in *ConQuest* and then analyse the results. Like with geometries again though, there is a pre-derived knowledge-base of information to help out with this – *IsoStar* – the CSD's knowledge-base containing pre-derived scatterplots of intermolecular interactions [8]. This knowledge-base of information allows the user to browse to the functional group of interest and immediately call up a 3D scatterplot showing how that functional group interacts with any one of a whole range of probe groups in the CSD.
With this extensive library of intermolecular interaction data it is possible to go one step further though and to analyse the interaction preferences of a whole molecule at once by combining the scatterplots for each individual functional group. The **Full Interaction Map** application [9], available within *Mercury* under **CSD-Discovery** or **CSD-Materials**, enables you to visualise knowledge-based interaction preferences for one or more molecules in the context of a crystal structure (Figure 4).

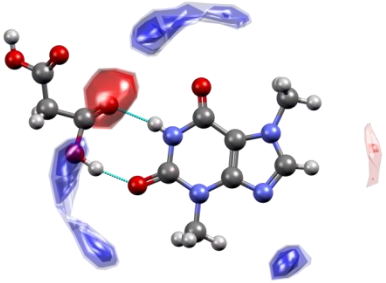

Figure 4. Full interaction maps displayed around the molecule theobromine to explore co-crystal design opportunities

Supramolecular Analysis

Crystal structures are, of course, the combination of molecular (or polymeric) geometries with intermolecular interactions through symmetry to form infinite supramolecular arrays. More complete comparison and understanding of crystal structures, particularly as applied to topics like polymorphism, co-crystal design and crystal structure prediction, requires analysis of these supramolecular packings and properties.

A wide range of applications to probe these areas are available within **CSD-Materials** [10] include **Crystal Packing Similarity** (Figure 5) for comparing structures, **Hydrogen Bond Propensities** [11] for predicting the likelihood of polymorphism and the **Hydrate Analyser** for studying hydrated structures. There are too many to explore here, but we will cover some of these in the hands-on tutorial.

Finally, in the area of supramolecular analysis there are applications for the study of protein-ligand structures. You can perform protein-ligand docking experiments and virtual screening using *GOLD* [12], a part of **CSD-Discovery**. Also available in **CSD-Discovery** is *Relibase* [13], which allows you to search, analyse and visualize protein-ligand structures.

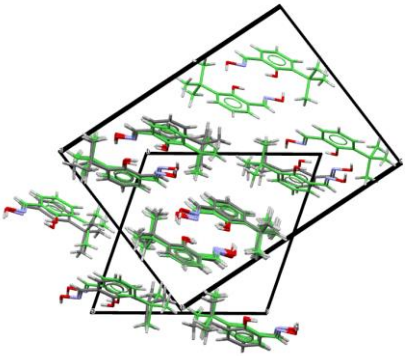

Figure 5. Crystal packing similarity overlay of the two polymorphs of 3-*tert*-butylsalicylaldoxime, before and after a pressure-induced phase transition

Conclusions

This overview has hopefully given you a rounded picture of what can be achieved using knowledge-based methods and of what software is available. The hands-on tutorial that will be presented at this school will focus primarily on advanced functionality for small molecule crystal structure comparison and analysis in the interests of time. There will be opportunities throughout the remainder of the school to investigate the wider options available from the CSD software and to ask further questions.

References

[1] J. Bernstein, Polymorphism in Molecular Crystals, Oxford University Press, Oxford, UK, 2002.

[2] C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, Acta Cryst. B, 2016, 72, 171-179.

[3] (a) F. H. Allen, W. D. S. Motherwell, Acta Cryst. B, 2002, 58, 407–422; (b) F. H. Allen, R. Taylor, Chem. Soc. Rev., 2004, 33, 463–475; (c) R. Wong, F. H. Allen, P. J. Willett, Appl. Crystallogr., 2010, 43, 811–824.

[4] (a) J. A. Chisholm, E. Pidcock, J. van de Streek, L. Infantes, W. D. S. Motherwell, F. H. Allen, CrystEngComm, 2006, 8, 11–28; (b) E. Pidcock, J. A. Chisholm, P. A. Wood, P. T. A. Galek, L. Fábián, O. Korb, A. J. Cruz-Cabeza, J. W. Liebeschuetz, C. R. Groom, F. H. Allen, The Cambridge Structural Database System and its applications in supramolecular chemistry and materials design. In Supramolecular Chemistry, from Molecules to Nanomaterials, Wiley: New York, 2012, 2927–2946; (c) P. T. A. Galek, E. Pidcock, P. A. Wood, I. J. Bruno, C. R. Groom, CrystEngComm, 2012, 14, 2391–2403.
[5] R. A. Sykes, P. McCabe, F. H. Allen, G. M. Battle, I. J. Bruno, P. A. Wood, J. Appl. Crystallogr., 2011, 44, 882-886.
[6] I. J. Bruno, J. C. Cole, M. Kessler, Jie 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.
[7] 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.
[8] J. C. Cole, C. R. Groom, O. Korb, P. McCabe, G. Shields, J. Chem. Inf. Model., 2016, 56, 652-661.
[9] 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.
[10] 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 and P. A. Wood, J. Appl. Crystallogr., 2008, 41, 466-470.
[11] P. T. A. Galek, L. Fabian, W. D. S. Motherwell, F. H. Allen, N. Feeder., Acta Cryst. B, 2007, 63, 768-782.
[12] G. Jones, P. Willett, R. C. Glen, A. R. Leach and R. Taylor, J. Mol. Biol., 1997, 267, 727-748.
[13] M. Hendlich, A. Bergner, J. Günther, G. Klebe, J. Mol. Biol., 2003, 326, 607-620.