CrystEngComm

ROYAL SOCIETY OF CHEMISTRY

EDITORIAL

View Article Online
View Journal | View Issue

Check for updates

Cite this: *CrystEngComm*, 2020, **22**, 7143

DOI: 10.1039/d0ce90154g

rsc.li/crystengcomm

Introduction to the Cambridge Structural Database – a wealth of knowledge gained from a million structures

Suzanna C. Ward* and Ghazala Sadiq*

The Cambridge Structural Database (CSD) grew from a vision by J. D. Bernal and Olga Kennard who believed that the collective use of data would lead to the discovery of new knowledge, transcending the results of individual experiments.[1] In 1965 Olga and her team began to collect together published bibliographic, chemical and crystal structure data for all small molecules studied by X-ray or neutron diffraction. With the rapid developments in computing taking place at the time, the collection went on to be encoded in electronic form and the CSD was born and shared with the world, becoming one of the first numerical scientific databases to begin operations.[2]

Following on from reaching CSD One Million the database now contains more than 1 080 000 structures associated with over 500 000 articles from over 400 000 different authors.[3] It is made up of organic (43%) and metal–organic (57%) crystal structures, including around 100 000 metal–organic frameworks (MOFs), 12 000 approved drugs and 11 000 different polymorphic families. The CSD is in use in almost every chemistry department in the world. Pharmaceutical companies depend on the database to drive drug discovery projects and materials scientists use it to interrogate, probe and design ever more complex 3D network structures.[4]

This themed issue of *CrystEngComm* combines 33 articles which highlight some of the many applications of the CSD in celebration of the one millionth crystal structure, a significant community achievement in 2019. Throughout this issue, the research carried out by the authors demonstrates the breadth of information and the variety of applications arising from the data in the CSD. Over the last half a century the complexity and size of structures have expanded, and the techniques and instrumentation used to determine new structures have evolved considerably. The articles herein show how far the field has evolved.

Several articles gain new insights from the CSD by investigating different aspects of data collection, from bibliographic details to interactions and chirality to co-crystallisation. Cole *et al.* (DOI: 10.1039/D0CE00045K) explore the changing nature of the CSD through a bibliographic analysis, reflecting on how crystallographic studies are now more commonplace within wider research rather than being the main focus of an article. Delving deeper into the fundamental chemistry, Gavezzotti *et al.* (DOI: 10.1039/D0CE00334D) use a subset of structures in the CSD in combination with experimental sublimation enthalpies to calibrate a new intermolecular potential field for organic compounds. The wealth of data in the CSD can also provide insights into the nature of bonds and Vologzhanina *et al.* (DOI: 10.1039/D0CE00288G) explore the peculiarities of Br⋯Br bonding, demonstrating good potential to reveal extended architectures based on Br⋯Br interactions that are typical for electroconductive polybromide-based materials. Swift *et al.* (DOI: 10.1039/D0CE00273A) note that as the world's largest repository for structural information there are many new types of questions that can be asked of the CSD. Data informatics are employed to identify and analyze hydrate–anhydrate structural pairs, outlining a general approach that could be used to identify subtle trends across numerous structures.

The prevalence of metal–organic structures in the CSD is evidenced by a number of the articles. With the rapidly increasing number of MOFs in the database, Fairen-Jimenez *et al.* (DOI: 10.1039/D0CE00299B) provide a tutorial on the most useful tools for efficient exploration of the CSD for MOF applications. They go on to describe what developments could further enhance the discovery of MOFs in the future. The work by Blatov *et al.* (DOI: 10.1039/D0CE00265H) also focusses on MOFs and looks at the dimensionality and underlying topology of coordination networks. Throughout their study they employ machine learning methods and these techniques are clearly becoming more popular. Cooper *et al.* (DOI: 10.1039/D0CE00111B) explore increasing the performance, trustworthiness and practical value of machine learning

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

models through a case study predicting hydrogen bond network dimensionalities from molecular diagrams.

As well as metal–organic structures the CSD contains a wealth of data that is relevant to the pharmaceutical industry. Over the last few decades, multi-component crystals have attracted significant attention as a possibility to alter the physicochemical characteristics of a drug compound without changing its chemical structure. The article by Infantes and co-workers (DOI: 10.1039/D0CE00948B) outlines the use of CSD statistical tools for co-former selection for the antiretroviral drug nevirapine. Assessing the supramolecular chemistry of the co-formers coupled with an understanding of the target API leads to a useful strategy for co-former selection. Seaton *et al.* (DOI: 10.1039/D0CE00301H) also highlight how utilising a combination of database mining, computational prediction and experimental screening leads to the formation of solid solutions of chiral and achiral components. We have seen a number of new techniques emerge through the CSD, and in their article Woollam *et al.* (DOI: 10.1039/D0CE01216E) reveal a promising protocol enabling the exploitation of electron diffraction data with limited resolution obtained from beam sensitive organic material. Bringing everything together, Wilson *et al.* (DOI: 10.1039/D0CE00898B) look at end-to-end pharmaceutical manufacturing processes highlighting how workflows can benefit from combining the wealth of data in the CSD with structural informatics and analysis tools.

Fifty five years after it was founded, the widespread use of structural data worldwide and the reliance on the CSD and associated software in drug discovery and development, materials science and life sciences are testament to the fact that the collation and curation of individual experiments has enabled new insights and knowledge from the data. These articles demonstrate the value of a curated collection of structural data and the reality of the vision of Kennard and Bernal.

So where will the CSD go next? Some might argue that we have "enough" crystal structures, but when we dig into the detail it's easy to see this is untrue. The wealth of data now allows users to train far more elaborate models in various spaces but there are still gaps, and new chemistries appear each year that extend us into new areas. More data means more relevant data to specific problem domains, leading to more accurate data driven modelling. Furthermore, a single crystal structure of a compound is a single point on a landscape of many possible structures associated with that compound. It may be that in the future, a study of a compound will routinely feature multiple different structures discovered as part of the research, leading to an enormous expansion in the volume of data in the CSD. More data on observable polymorphs will allow scientists to gain far greater insights into the nature of crystallisation. Another direction is in the breadth of data: we can see a world where crystal structures have more information associated with them, allowing a broader range of use. This should lead to a rich array of research, and probably many more publications here in *CrystEngComm*!

We are grateful to all of the authors contributing to this special issue, our colleagues at the CCDC, in particular Dr Jason Cole for his many insights, and to the *CrystEngComm* editorial office for their support in making it possible. We are excited to see what more can be learnt from the CSD in the years to come.

References

1 O. Kennard, *The Impact of Electronic Publishing on the Academic Community, From private data to public knowledge*, Portland Press, 1997, https://portlandpress.com/DocumentLibrary/Umbrella/Wenner%20Gren/Vol%2073/Sesh%206/Chapter%202.pdf.

2 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 and D. G. Watson, *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.*, 1979, **35**, 2331–2339, DOI: 10.1107/S0567740879009249.

3 C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.*, 2016, **72**, 171–179, DOI: 10.1107/S2052520616003954.

4 C. Groom and J. Cole, *The newsletter of The Cambridge Crystallographic Data Centre*, June 2015.