

Challenges and opportunities in curating one million crystal structures

Amy Sarjeant^a, Suzanna Ward^b and Ian Bruno^b

^a The Cambridge Crystallographic Data Centre, 252 Nassau St. Princeton NJ 08542, USA, sarjeant@ccdc.cam.ac.uk

^b The Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge UK CB2 1EZ, ward@ccdc.cam.ac.uk, bruno@ccdc.cam.ac.uk

As the Cambridge Structural Database (CSD) is poised to add its 1 millionth structure this year, we stop to reflect on the challenges involved in curating the world's repository of small molecule crystal structure data and the opportunities that such big data can afford. Ensuring the completeness and quality of the data entrusted to us, accommodating data from new sources and experimental techniques, and continuing to provide relevant data in an ever-growing database are among a few of the challenges we face. The power of the CSD to inform various fields of chemical and biochemical research, stems from validation and enrichment of the data, as well as placing those data into a broader context. Only by continuing to evolve are we able to meet the needs of such a dynamic research community.

This talk will explore the part we can play in helping the community set and adhere to new best practices for data management that will enable researchers to get the most from crystal structure data. It will focus on the steps the CCDC has made in the past to address challenges, as well as some of the projects we are investigating as we seek to add the next million structures to the CSD.