Poster

Ultra-fast detection of (near-)duplicate structures across major crystal databases

Daniel Widdowson[1], Vitaliy Kurlin[1]

[1]*Computer Science department, Materials Innovation Factory, University of Liverpool, United Kingdom*

D.E.Widdowson@liverpool.ac.uk

The Cambridge Structural Database (CSD) and the Crystallography Open Database (COD) contain thousands of structures. Large portions of these databases overlap, often because their entries originate from the same publication. Constructing a list of CSD-COD cross entries is difficult because data can be reformatted in the curation process, many data points do not reliably distinguish or identify crystals, and most ways of finding matches are slow, requiring on the order of $_{10}{}^{10}$ comparisons.

Using geometry-based structural invariants (PDD [1]: Pointwise Distance Distribution and its simplified version AMD [2]: Average Minimum Distance), we compared all-vs-all entries in the CSD and COD, discovering the extent of their overlap for the first time.

PDD is invariant under any choices of unit cells; if two crystals are rigidly equivalent, their PDDs are identical. Unlike the density of a crystal, which can coincide for some crystals by chance, PDD is stronger than the Pair Distribution Function and distinguishes all different periodic crystals in the CSD in under one hour. If all atoms are sightly displaced, PDD continuously changes and hence can find near duplicates in noisy data.

Using these invariants, we compared 1,214,848 entries from the CSD against 508,392 from COD, taking only 17 minutes on a typical desktop computer. Over 400,000 crystals were matched, an overlap of 33% of the CSD and 80% of COD. We also found a significant overlap of COD with the ICSD (over 50,000 entries), as well as at least minor overlaps with the Materials Project database. We additionally searched for duplicate entries, finding several thousand in both the CSD and COD, many of which are not listed anywhere as known duplicates. The recent 'GNoME' dataset [3] of AI-predicted crystals has over 1,000 duplicates.

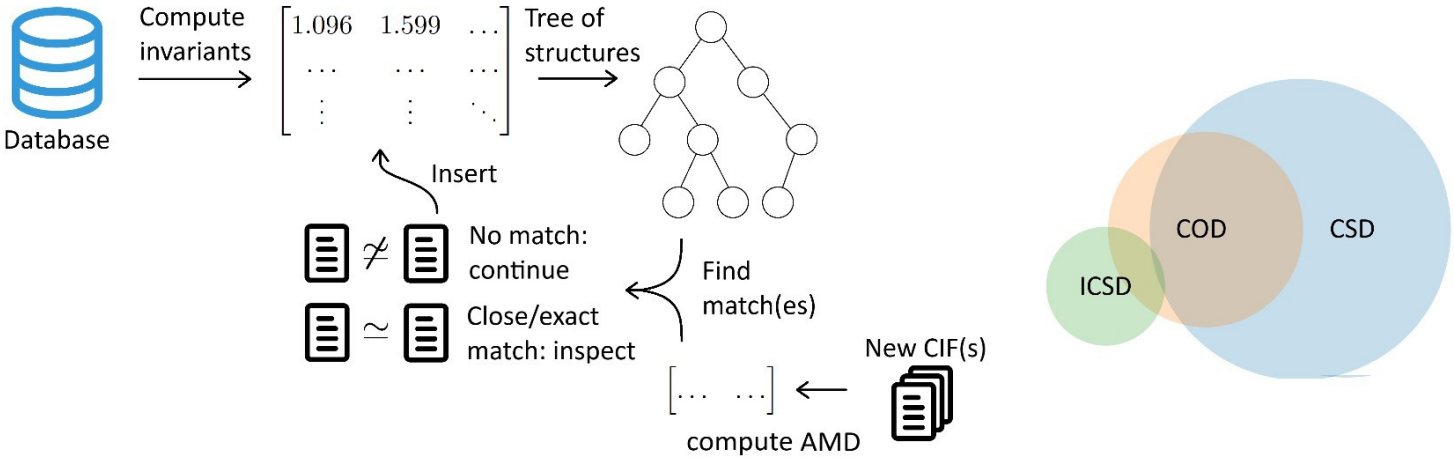

Figure 1. Left: Proposed workflow for incorporating structural invariants into curation. Right: Venn diagram of the discovered extent of overlap between the CSD, ICSD and COD.

[1] Widdowson, D., Kurlin, V. Resolving the data ambiguity for periodic crystals. Adv. Neural Inform. Proc. Systems, v.35, p.24625-24638, (2022).

[2] MATCH Communications Math. Computer Chemistry 87, 529-559 (2022).

[3] Nature 624, 80-85 (2023).