Poster

Comparison of crystal structure similarity algorithms and analysis of large sets of theoretically predicted structures

N. Francia[1], L. M. Hunnisett[1], J. Nyman[1], C. J. Kingsbury[1], I. Sugden[1], G. Sadiq[1], J. Cole[1]

[1]The Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, UK
nfrancia@ccdc.cam.ac.uk

Computational Crystal Structure Prediction (CSP) methods have evolved over the past few decades from studying small rigid molecules to being able to predict polymorphs of molecules of increasing molecular size and conformational complexity, attracting the interests of pharmaceutical and functional materials companies [1]. This is reflected in the targets for the recent 7 th CSP Blind Test, organised by the Cambridge Crystallographic Data Centre (CCDC), featuring highly flexible molecules, multi-component systems and challenging molecular sizes [2]. CSP methods typically consist of a first-generation step that produces 10^3-10^6 structures, followed by a series of increasingly more accurate energy evaluations, with the highest-energy structures rejected at each stage [3].

Large datasets of computationally generated crystal structures, such as those produced in the recent 7th CSP Blind Test, necessitate efficient and automated data analysis tools. This work presents the improvements of traditional crystal structure similarity algorithms (e.g., Crystal Packing Similarity, PXRD Similarity [4]) and the exploration of novel, computationally faster techniques (e.g., Pointwise Distance Distributions [5]) on this extensive dataset (>100,000 structures over 6 systems). This allowed us not only to identify theoretical-experimental structure matches but also to explore the agreement between different structure generation and ranking methods. Moreover, different analysis tools are being tested against such large datasets to determine recurring structural motifs, such as the presence of dimers as the building block of multiple structures or common chains and layers of hydrogen-bonded molecules.

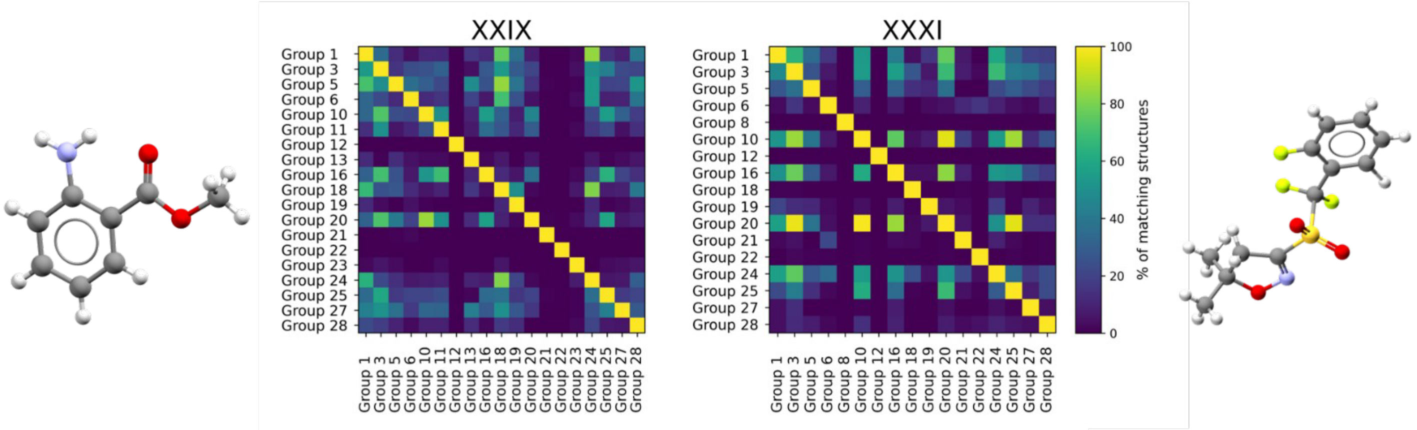

Figure 1. Crystal structure set similarity heat maps for molecules XXIX and XXXI showing the percentage of structures from the group on the horizontal axis that match a structure from the group on the vertical axis.

[1] Nyman, J., & Reutzel-Edens, S. M. (2018). *Faraday Discussions*, **211**, 459-476.

[2] L. M. Hunnisett et al., (2024) *Acta Cryst. B, submitted*.

[3] Day, G. M. (2011). *Crystallography Reviews*, **17(1)**, 3-52.

[4] C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, Acta Crystallographica Section B, **B72**, 171-179, 2016

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

We thank Prof. Vitaliy Kurlin and Dr Daniel Widdowson for sharing the code for Pointwise Distance Distributions.