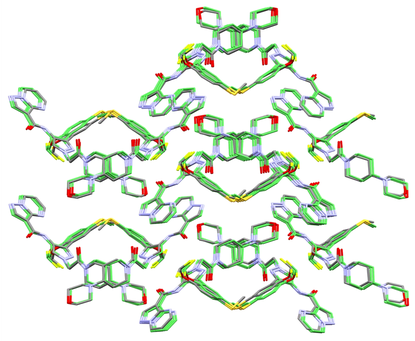
# Comparison of crystal structure similarity algorithms for large sets of theoretically predicted structures

**Nicholas Francia, Lily Hunnisett, Jonas Nyman, Isaac Sugden, Ghazala Sadiq, Jason Cole**

**The Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, UK**

## Introduction

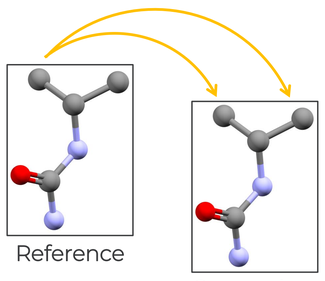
Computational Crystal Structure Prediction (CSP) methods are now able to predict polymorphs of molecules of considerable size and conformational complexity [1]. This is reflected in the targets for the recent 7th CSP Blind Test, organised by the Cambridge Crystallographic Data Centre (CCDC), featuring highly flexible molecules, multi-component systems and challenging molecular sizes [2].

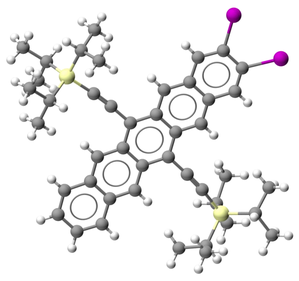


## Aim of the Study

Crystal structure similarity algorithms have been used in analysing CSP-generated sets to both remove duplicates and identify experimental crystal forms. We here compare traditionally used analysis tools, such as the Crystal Packing Similarity algorithm and PXRD Similarity program, together with the Pointwise Distance Distributions (PDD) [3, 4] and the VC-PWDF [5]. We identify the strengths and limits of each method by looking at a few case studies from the 7th Blind Test.

## Topological Symmetries



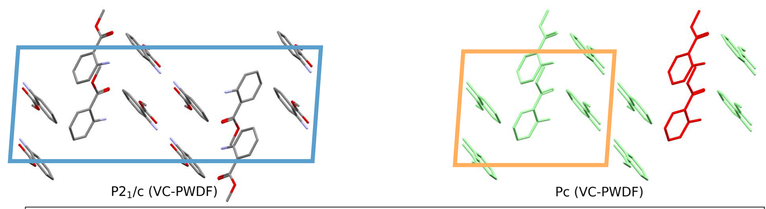
Molecule XXVII

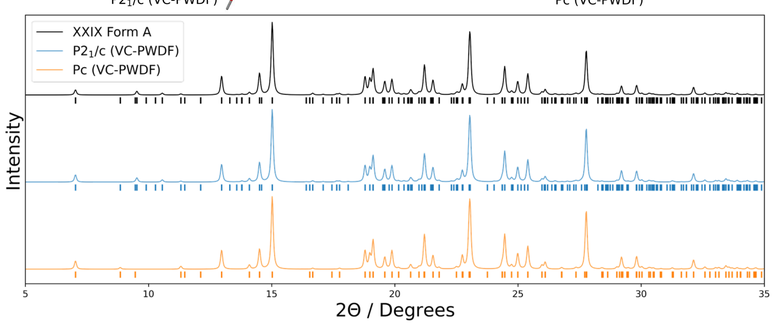
Topological symmetries increase the time required to compare structures with Crystal Packing Similarity. Comparisons were computationally demanding for molecule XXVII which has more than 4500 possible permutations.

This prompted the introduction of distance constraints within Crystal Packing Similarity which drastically reduced the number of possible permutations, speeding up the comparison of two structures. This new update is available in the 2024.1 release.

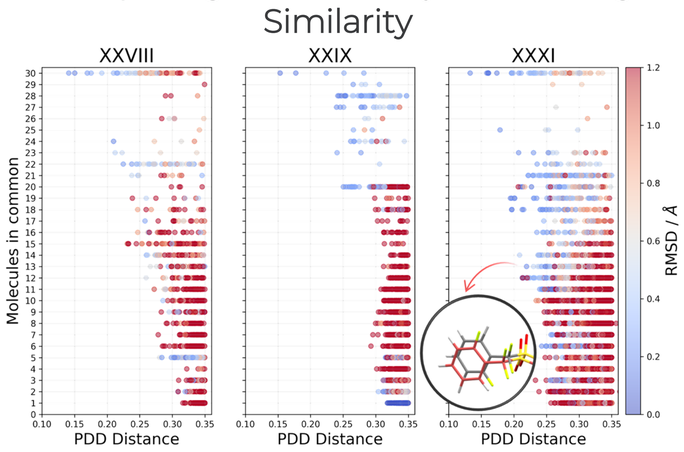
## Identifying Putative Polytypes

Form A of molecule XXIX is a $Z' = 3$ crystal in the $P2_1/c$ space group. In two of the submitted sets, a possible *Pc* polytype of the experimental form was found with 5 layers out of 6 which perfectly overlap with form A. This structure needs a computationally expensive 70-molecule molecular shell to be distinguished from the experimental one with Crystal Packing Similarity. However, despite being similar, its powder pattern fails to index the experimental one.



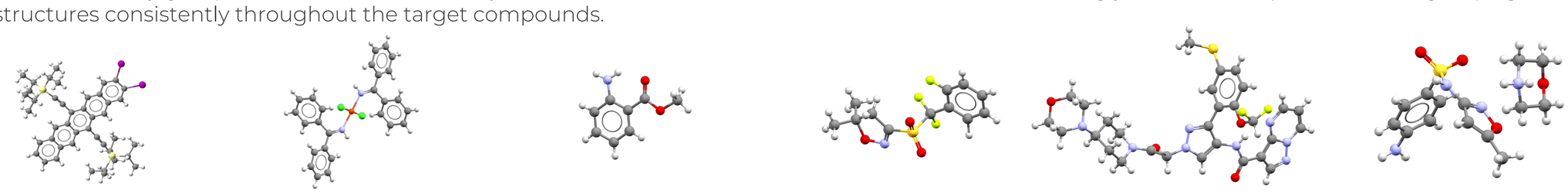

## Comparing PDD and Crystal Packing Similarity

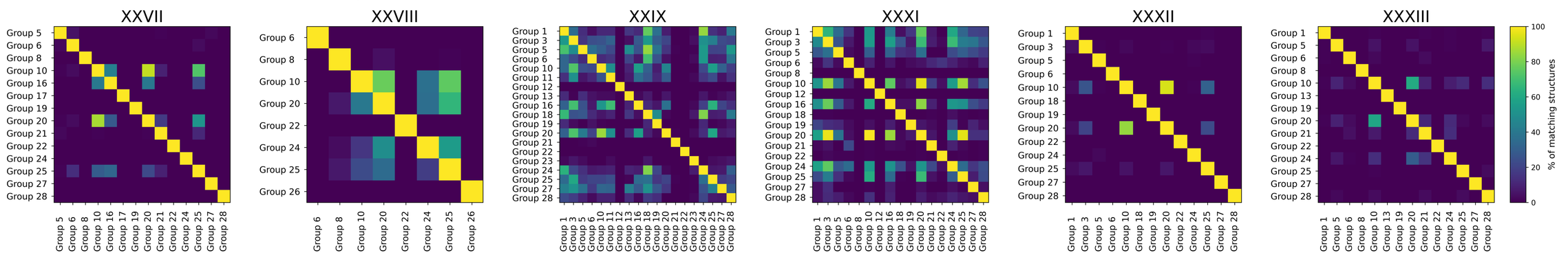



In most cases, a good agreement between the two methods can be seen when comparing CSP structures with experimental crystals. For molecule XXXI, PDD tends to overestimate the similarity due to a lack of chemical information. Despite this, the main advantage of PDD is its computational efficiency, being 1000x faster than Crystal Packing Similarity, making it possible to perform large-scale comparisons or be used to pre-select relevant comparisons.

## Crystal Structure Landscape Similarity

PDD approach was used to perform a purely geometrical crystal structure similarity comparison between the submitted structure sets and therefore assess search completeness. Matches were identified using a cutoff of 0.225 Å to reduce the impact of false positives, exclude poorly overlapping structures and balance the missed perfect matches with the inclusion of a few partial matches. Target systems XXIX and XXXI, both small molecules with a few conformations available, show a substantial overlap between many groups. As the size and flexibility of the molecule increase, the CSP sets become increasingly different. Despite this, a few groups generated similar structures consistently throughout the target compounds.

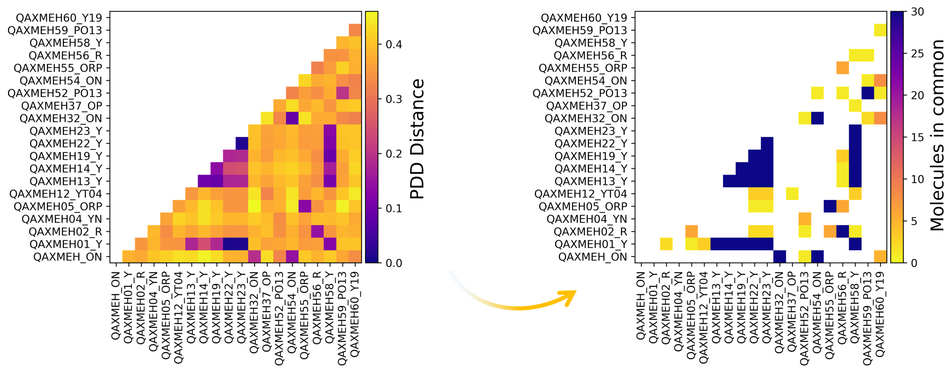



## Conclusions and Future Developments

The comparison of crystal structures and the identification of matches can be sensitive to the method applied, suggesting the use of alternative comparison approaches to exploit the advantages of each of them. Performance improvements have been made to Crystal Packing Similarity when topological symmetries are present. The analysis of powder patterns was found useful in distinguishing putative polytypes. The computationally efficient PDD approach made it possible to compare the sets of the 7th Blind Test and it is a valuable tool to be used in the early stages of CSP when the clustering of large sets of structures is essential to remove possible duplicates [6].

A two-step approach, available in the 2024.1 CSD release, in which Crystal Packing Similarity is used only on the best matches by PDD has been found to drastically reduce the computation time while maintaining the accuracy of Crystal Packing Similarity.




## References

[1] J. Nyman, S. Reutzel-Edens, Faraday Discuss., 211, 459-476 (2018)
[2] L. Hunnisett, J. Cole, G. Sadiq, Acta. Cryst. Sect. A., 78, a136-a136 (2022)
[3] D. Widdowson, M. Mosca, A. Pulido, A. Cooper, V. Kurlin, match., 87, 529-559 (2021)
[4] D. Widdowson and V. Kurlin, Adva. Neural Inf. Proc. Syst, 35 (2022)
[5] R. A. Mayo, K.M. Marczenko and E.R. Johnson, Chemical Science, 14(18), 4777-4785 (2023)
[6] G. Day, Crystallography Reviews, 17, 3-52 (2011)

## Acknowledgements

We would like to express our gratitude to the CCDC Development Team for implementing new features in both Mercury and the CSD Python API, as well as for various performance improvements to existing crystal structure similarity tools. We also extend our thanks to the participants of the 7th Blind Test for their valuable insights on crystal comparisons. Lastly, we would like to acknowledge Dr. Vitaliy Kurlin and Dr. Daniel Widdowson from the University of Liverpool for sharing the PDD code, and Dr. Alex Mayo, Prof. Alberto Otero-de-la-Roza, and Prof. Erin R. Johnson for the VC-PWDF code.