Informatics meets energetics: combining crystal structure prediction and knowledge of the crystal landscape

E. Hawking[1]

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

ehawking@ccdc.cam.ac.uk

The significance of polymorphism in pharmaceuticals necessitates careful selection of a solid form to ensure thermodynamic or kinetic stability throughout drug manufacturing and production. Despite extensive polymorph screening methods, finding a more stable polymorph remains a risk, therefore experimental workflows are supplemented by computational methods such as Crystal Structure Prediction (CSP) and informatics-based risk assessments to derisk emerging of new forms during the development of formulated products. CSP has evolved to become integral in pharmaceutical materials science workflows to investigate the solid form landscape, where predicted crystal structures are ranked by their relative energies and compared against experimental structures [1]. However, the landscapes of predicted crystal packing arrangements are complex, with contributions from hydrogen bonding, molecular conformation and weak intermolecular interactions. Therefore, it can be challenging to for crystal engineers to understand the contribution of individual factors to the overall relative energy of different putative structures. Furthermore, these computational methods remain resource intensive and typically only applied to a selected number of molecules.

The Cambridge Structural Database (CSD), comprising over 1.3 million experimentally determined crystal structures, serves as a comprehensive resource for understanding the vast diversity of intramolecular geometries and intermolecular interactions [2]. This dataset provides valuable insights into the structural factors that influence solid-state properties. Solid Form Informatics leverages this wealth of crystallographic data to analyse and contextualize the various contributors to polymorphism. By integrating experimental observations, the use of informatics can help rationalize the solid form landscape and understand the risk of emerging new forms.

Analysis of CSP landscapes has been conducted to provide insight into the crystal forms using the tools from informatics based risk assessments. This has focused on generating and mining structural data from CSP landscapes, as well as the CSD, to de-risk solid forms. Insights from energetically ranked CSP landscapes have been compared with landscapes generated using CSD entries as templates to compare trends between different methods. Machine learning models were generated to classify low and high energy structures for individual CSP landscapes and analysis of the models was conducted to elucidate the importance of individual features. This has facilitated the understanding of relative importance of conformation, hydrogen bonding and weak intermolecular features, and demonstrates how CSP can be leveraged alongside structural data to understand contributing factors to low and high energy polymorphs. This has enabled rationalization of the factors contributing to the thermodynamic stability of small molecule crystal structures, providing context and insight into the crystal landscape and rationalise relationships of experimentally observed forms.

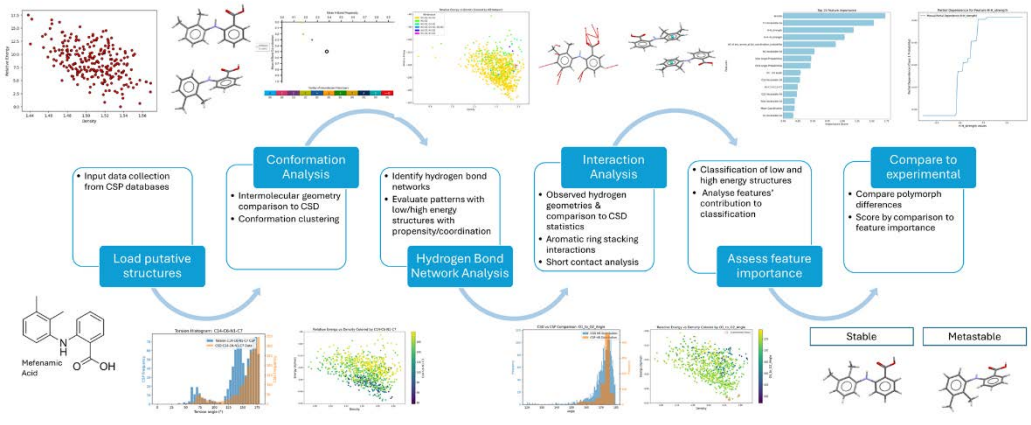

Fig. 1. Informatics meets energetics workflow leveraging CSD data to rationalize experimental stability.

[1] Toward computational polymorph prediction. Price SL, Price LS. 2018:133– 157. [2] Chupas, P. J., Ciraolo, M. F., Hanson, J. C. & Grey, C. P. (2001). *J. Am. Chem. Soc.*, **123**, 1694.

[2] A million crystal structures: the whole is greater than the sum of its parts. Taylor R, Wood PA. Chem Rev. 2019;119(16):9427–9477.