Artificial Intelligence in Crystallography: Navigating Big Data Challenges

D. W. M. Hofmann

Via dei Colombi 3, 09012 Capoterra, Italy

dwmhofmann@aol.com

The burgeoning volume of crystallographic data housed in structural databases like the CSD, ICSD, and PDB, coupled with the expanding powder diffraction database PDF®, urgently calls for advanced automated analytical approaches. This presentation will highlight the transformative potential of artificial intelligence (AI), specifically machine learning, in confronting the intricate challenges inherent in modern crystallography.

We will commence with a concise overview of foundational machine learning concepts, including essential elements such as similarity indices, input and output vectors, and the various problem categories addressed by these methods. These include classification through supervised learning, multidimensional regression, clustering, dimensionality reduction, and anomaly detection. Each technique will be vividly exemplified through its application to compelling crystallographic and chemical inquiries.

Supervised learning will be demonstrated for the derivation of atom-atom interactions. A brief discussion will follow on common algorithms such as support vector machines and neural networks, contrasting their respective advantages and disadvantages. For multidimensional regression, the prediction of material density will serve as a practical illustration. The utility of clustering will be showcased in critical areas like crystal structure prediction, polymorph recognition, and the analysis of powder diffraction patterns. Furthermore, dimensionality reduction will be explored through its application to predicting solubility. Finally, we will illustrate how anomaly detection, when applied to the CSD, effectively identifies both systematic and random errors within this vast repository. While drawing upon foundational work initiated in 2003 [4], the core of this talk will align with a recent article published in International Tables for Crystallography [1], offering a comprehensive perspective on AI's pivotal role in advancing crystallographic science.

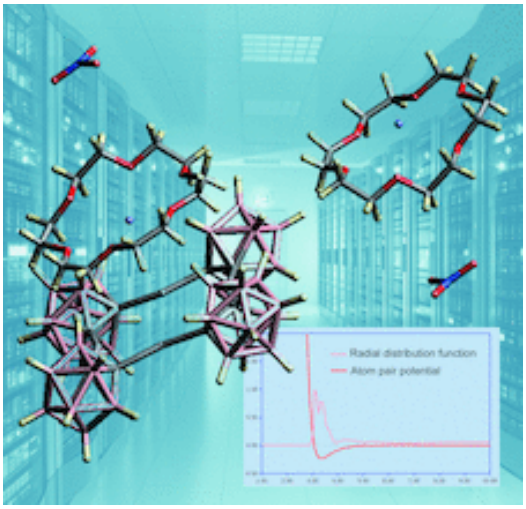

Figure 1. Supervised Learning allows to extract a lot of informatiom about intermolecular interactions from crystal structures [2]

[1] Hofmann, D. W. M & Kuleshova, L. N.. (2022). *Data mining. I. Machine learning in crystallographyexture Analysis in Materials Science*. Chester: International Union of Crystallography.

[2] Hofmann, D. W. M., & Kuleshova, L. N. (2023). *Acta Crystallogr. A*, **70(2)**, 132.

[3] Hofmann, D. W. M & Kuleshova, L. N.. (Eds.) (2009). *Data Mining in Crystallography*. Heidelberg: Springer Science & Business Media.

[4] Hofmann, D. W., & Apostolakis, J. (2003). *J. Mol. Struc.*, ***647*(1-3)**, 17.