New generalized crystallographic descriptors for structural machine learning

R. Zhang[1], S. Seth[2], J. Cumby[1]

[1]School of Chemistry, University of Edinburgh, Edinburgh, UK,
[2]School of Informatics, University of Edinburgh, Edinburgh, UK

james.cumby@ed.ac.uk

The ever-growing amount of crystallographic data offers the potential to uncover a range of scientific discoveries, from rapidly predicting physical properties to suggesting new materials with desirable functional behaviours. This is further enhanced by the current growth in machine learning (ML) algorithm development and implementation. There is, however, a significant obstacle to this goal; standard crystallographic information are not suitable inputs for ML algorithms. This arises due to the inherent flexibility of crystallography, such as non-unique unit cell definitions and symmetry. To overcome this problem, significant progress has been made in devising 'descriptors' for crystallographic ML, compressing and standardising crystallographic information into a smaller feature space. Much of the existing focus has been on molecular crystals, where the finite extent of individual molecules imposes a limit on the size of feature vector required. A large number of approaches have been proposed but do not easily extrapolate to extended (i.e. inorganic) materials. [1] The descriptors that are suitable for extended solids tend to be either hand-crafted for a specific problem, or have so many dimensions that extremely large datasets must be used to train reliable ML models. In addition, many do not scale well with variable numbers of atomic species.

Here, we present two new descriptors for crystallographic materials which are generally applicable and invariant to compositional complexity. The first is based on a real-space view of the structure, the second on a reciprocal (or diffraction) space view. Both descriptions are invariant to atomic permutations and unit cell choice, and can be considered as an 'extended' (i.e. more information-rich) version of the atomic radial distribution function (RDF) and powder diffraction pattern, respectively. The more complete features offered by these descriptors results in better physical property predictions. For example, our 'extended' RDF can predict bulk modulus from crystal structures obtained from the Materials Project [2] with a much lower error than the 'simple' RDF using linear ridge regression (Figure 1). It is notable that the error approaches current state-of-the-art results, [3] without any knowledge of the atom types involved.

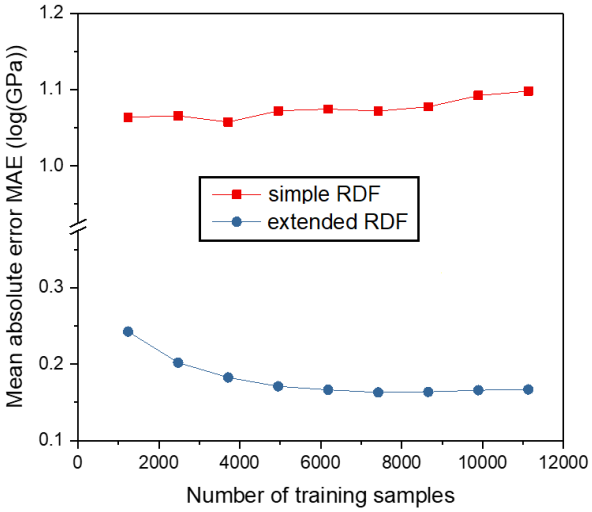

Figure 1. The variation in bulk modulus prediction error with increasing training examples showing better bulk modulus prediction using our 'extended' RDF than the existing (simple) RDF descriptor.

[1] Rossi, K. & Cumby, J. (2020). *Int. J. Quantum Chem.*, **120**, e26151.

[2] Jain, A., Ong, S. P., Hautier, G., *et al*. (2013). *APL Mater.*, 1(1), 011002.

[3] Chen, C., Ye, W., Zuo, Y. *et al.* (2019). *Chem. Mater.*, 31, 3564.

Keywords: machine learning; crystal descriptors; physical property prediction; RDF; radial distribution function

Acta Cryst. (2021), A**77**, C476