Introduction to invariant-based machine learning for periodic crystals

J. Ropers, M. M. Mosca, O. Anosova, V. Kurlin

University of Liverpool, Liverpool L69 3BX, United Kingdom, vitaliy.kurlin@liverpool.ac.uk

Machine learning can be justified only if input descriptors are crystal invariants independent of accidental choices. To use a household analogy, the average color of human clothes can be the easiest descriptor extracted from images but cannot be seriously used to predict height or any reliable data about people. Similarly, no properties of crystals can be reliably predicted from ambiguous parameters of a unit cell and a motif. Since crystal structures are determined in a rigid form, they should be considered *equivalent* modulo rigid motion or *isometry*, which preserves all interpoint distances. Then crystals can be justifiably distinguished only by *isometry invariants* that are independent of a unit cell and are preserved under any translations and rotations. Though Niggli's reduced cell is unique, it is discontinuous under atomic perturbations, which are always present in real crystals. This continuity of invariants is important to quantify similarities between near identical crystals obtained by Crystal Structure Prediction [1] as approximations to energy minima.

The pictures on the right of Figure 1 show that almost any small perturbation of points breaks most past invariants: symmetry groups and descriptors depending on primitive or reduced cells, because the volume of the primitive cell doubles. The density was used to represent a crystal structure in energy landscapes but is constant under perturbations and not enough to distinguish dense (really, almost all) crystals. But density functions [2] depending on a variable radius give more information about relative positions of atoms.

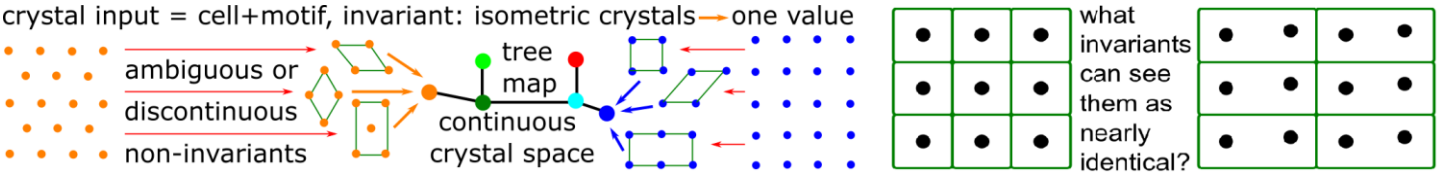

Figure 1. A traditional representation of a periodic crystal by ambiguous (cell, motif) on the left should be converted to isometry invariants such as complete isosets [3,4], which uniquely identify a crystal so that a similarity on the right is continuously quantified.

All machine learning approaches implicitly assume that a target property continuously depends on a given input, for example similar crystals should have close values of their lattice energy. We experimentally tested that the lattice energy is discontinuous with respect to the density, powder X-ray diffraction and packing similarity (root mean square deviation as computed by Mercury). For example, many crystals detected as similar by the above tools have very different energies. The new invariants [5-6] are not only theoretically continuous under perturbations but also satisfy continuity for energy learning: we experimentally identified a distance threshold d and a constant c such that any distance between AMD invariants smaller than d guarantees an energy difference smaller than c times d [7].

Standard machine learning tools were trained on AMD invariants without chemical data for 10 min and predicted the lattice energy with a mean average error of less than 5KJ/mole on a CSP dataset of 5679 crystals [1] containing about 250 atoms per unit cell.

Distances between AMD invariants are computed so fast that the pairs of all 229K organic molecular crystals from the Cambridge Structural Database (CSD) were processed overnight on a modest desktop and detected numerous near duplicates in the CSD [5-6].

[1] Pulido, A. et al, Functional materials discovery using energy–structure–function maps. *Nature*, *543*(7647), pp.657-664.

[2] Edelsbrunner, H., Heiss, T., Kurlin, V., Smith, P, Wintraecken, M. (2021). The density fingerprint of a periodic point set. Peer-reviewed proceedings of *Symposium on Computational Geometry*. Available at http://kurlin.org/research-papers.php#SoCG2021.

[3] Anosova, O., Kurlin, V. (2021). An isometry classification of periodic point sets. Peer-reviewed proceedings of *Discrete Geometry and Mathematical Morphology*, available at http://kurlin.org/research-papers.php#DGMM2021.

[4] Anosova, O., Kurlin, V. (2021). Introduction to Periodic Geometry and Topology. Available at https://arxiv.org/abs/2103.02749.

[5] Widdowson, D., Mosca, M.M., Pulido, A., Kurlin, V., Cooper, A.I. Average Minimum Distances of periodic point sets. To appear in MATCH Communications in Mathematical and in Computer Chemistry (2022). Available at https://arxiv.org/abs/2009.02488.

[6] Widdowson, D., Kurlin, V. Pointwise Distance Distributions of periodic sets. Available at http://arxiv.org/abs/2108.04798.

[7] Ropers, J., Mosca, M.M., Anosova, O., Kurlin, V., Cooper, A.I. Fast predictions of lattice energies by continuous isometry invariants of periodic crystals. Proceedings of DACOMSIN (Data and Computation for Materials Science and Innovation) 2021.

Keywords: machine learning; lattice energy prediction; crystal invariant and similarities; continuous classification of crystals