MS41-2-2 Algorithms for automated detection of (near-)duplicate periodic crystals
#MS41-2-2

V. Kurlin [1]
[1]**Materials Innovation Factory, University of Liverpool - Liverpool (United Kingdom)**

Abstract
A periodic crystal is usually given in a Crystallographic Information File as a pair of a (primitive, conventional, or reduced) unit cell and a motif of atoms with fractional coordinates with respect to a cell basis. This traditional representation is highly ambiguous because the same crystal can be given by infinitely many pairs of a cell basis and a motif of atoms.
Since crystal structures are determined in a rigid form, their most practical equivalence is rigid motion (a composition of translations and rotations) or isometry (including reflections), which maintain all inter-atomic distances. The most natural definition of a periodic crystal is not a single set of points given but an isometry class of infinitely many periodic point sets that are all equivalent to each other up to isometry.
This new definition clarifies the key question “same or different” that remained unresolved [2] by classical tools such as the powder diffraction pattern (PXRD) and packing similarity (RMSD implemented by Mercury). Both similarities RMSD and 1-PXRD fail the triangle inequality [3], which should be satisfied by any well-defined metric. Figure 1 (left) shows that almost any tiny displacement (thermal vibration) of atoms discontinuously affects classical invariants based on symmetry groups or reduced (conventional) cells that can double in size.
To reliably detect (near-)duplicates appearing in all experimental databases and paper mills [1], periodic crystals should be compared by their complete isometry invariants, which are crystal descriptors with no false negatives and no false positives.
The first complete invariants of all periodic crystals are the isosets [4] whose key ingredients are local clusters of a radius alpha around motif points. Figure 1 (middle and right) shows how these clusters grow when the radius alpha is increasing. Isometry classes of all alpha-clusters are arranged into a hierarchical isotree showing their evolution for all alpha, see Figure 2. The isotree always stabilizes for any periodic point set and algorithmically detects an isometry in a time cubically depending on the size of a motif.
A continuous metric on isosets was approximated with a factor of only 4 in 3-space [5]. These results were confirmed by 200B+ pairwise comparisons of all 660K+ periodic crystals in the Cambridge Structural Database, which took a couple of days on a modest desktop [6] and detected five pairs of isometric duplicates where one atom is replaced by a different one, for example, Cd by Mn in HIFCAB vs JEPLIA.

References
[1] Bimler, David. Better Living through Coordination Chemistry: A descriptive study of a prolific papermill that combines crystallography and medicine, IUCr Newsletter April 2022
https://www.chemistryworld.com/news/800-crystallography-related-papers-appear-to-stem-from-one-paper-mill/4015589.article.
[2] Sacchi, P., Matteo Lusi, M., Cruz-Cabeza, A.J., Nauhac, E.Joel Bernstein, J. Same or different – that is the question: identification of crystal forms from crystal structure data. CrystEngComm 22 (43), 7170-7185, 2020.
[3] Widdowson, D., Kurlin, V., Pointwise Distance Distributions of finite and periodic point sets. Arxiv.org:2108.04798. The latest version is at http://kurlin.org/projects/periodic-geometry-topology/PDD.pdf
[4] Anosova, O., Kurlin, V. An isometry classification of periodic point sets. Peer-reviewed Proceedings of Discrete Geometry and Mathematical Morphology 2021, p. 229-241. Available at http://kurlin.org/projects/projects/periodic-geometry-topology/crystal-isosets-complete.pdf
[5] Anosova, O., Kurlin, V. Algorithms for continuous metrics on periodic crystals. Arxiv: 2205.15298. The latest version is available at http://kurlin.org/projects/projects/periodic-geometry-topology/isoset-metric.pdf
[6] Widdowson, D., Mosca, M.M., Pulido, A., Kurlin, V., Cooper, A.I. Average Minimum Distances of periodic point sets - fundamental invariants for mapping all periodic crystals. MATCH Communications in Mathematical and in Computer Chemistry, v.87(3), p.529-559, 2022. Available at http://kurlin.org/projects/periodic-geometry-topology/AMD.pdf

Discontinuity of most invariants under vibration

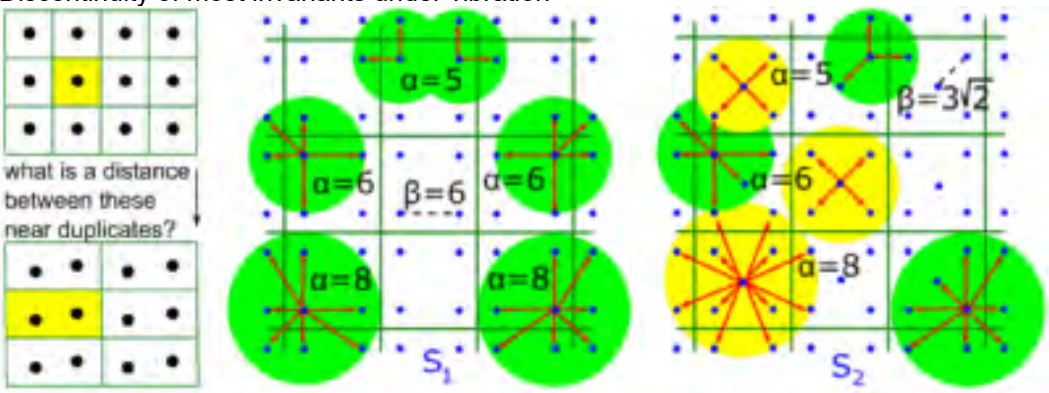

Figure 1: **Left**: most invariants based on symmetry groups and reduced or conventional cells are discontinuous under tiny perturbations. **Middle**: the periodic set $S_1 \subset \mathbb{R}^2$ has the square unit cell $[0,10)^2$ containing the four points $(2,2),(2,8),(8,2),(8,8)$. All local α-clusters of S_1 are isometric, shown by red arrows for $\alpha = 5,6,8$. **Right**: in comparison with S_1, the set S_2 has the extra point $(5,5)$ in the center of the cell $[0,10)^2$, so S_2 has green and yellow isometry types of α-clusters.

Isotree visualizing the complete invariant isoset

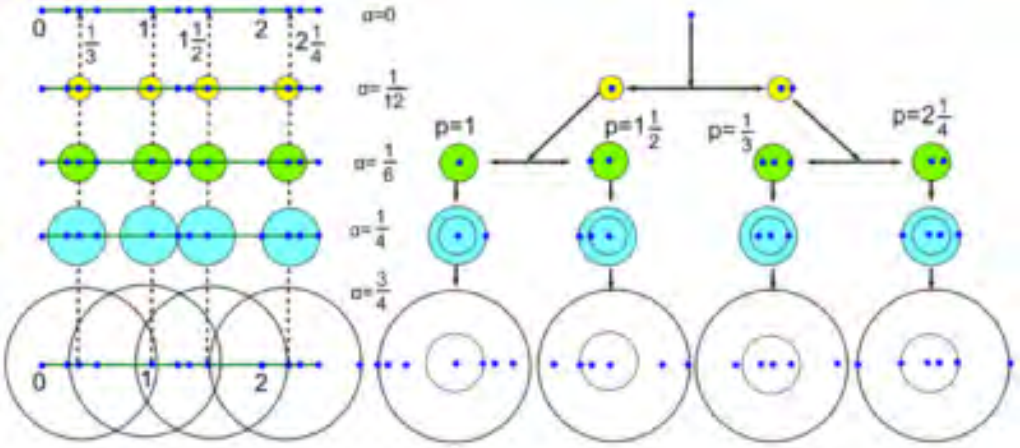

Figure 2: **Left**: the 1-dimensional periodic set $S_4 = \{0, \frac{3}{4}, \frac{1}{2}, \frac{1}{2}\} + \mathbb{Z}$ has four points in the unit cell $[0,1)$. **Right**: the isotree $\mathrm{IT}(S_4)$ consists of isometry classes of the local α-clusters of radius α around the points of $S_4 \subset \mathbb{R}$ with radii $\alpha = 0, \frac{1}{12}, \frac{1}{6}, \frac{1}{4}, \frac{3}{4}$. The α-clusters are intervals shown by colored disks.