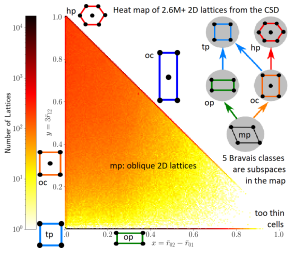
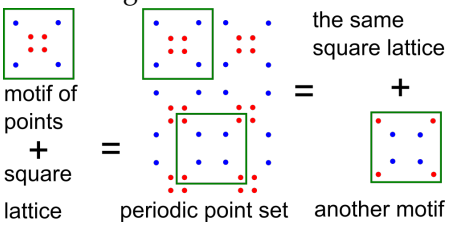
CRISP : the Crystal Isometry Principle

Andy Cooper and Vitaliy Kurlin's Data Science group
Materials Innovation Factory, University of Liverpool



What is a *periodic crystal S*?

A crystal S is traditionally defined by a motif of atoms periodically translated along basis vectors of a unit cell.



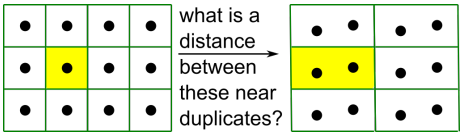
Since crystals are determined in a rigid form, their strongest *equivalence* is *rigid motion*, which is a compositions of translations and rotations.

A **new** definition of a *crystal*

A **periodic crystal** is a class of all (infinitely many) periodic point sets that are equivalent to each other under rigid motion, or a slightly weaker *isometry* = rigid motion + reflection.

A Crystallographic Information File is only one of the infinitely many representations (photos) of a crystal but comparisons need a continuous DNA-style code (materials genome).

Crystal descriptors should be *invariant* (preserved under any rigid motion). The symmetry group and reduced cell are invariant but discontinuous under almost any perturbation.



Mapping crystals problem

Find a map I : {periodic crystals} $\to$ a simpler space with the conditions:

invariance : $S \simeq Q$ are isometric $\Rightarrow$ $I(S) = I(Q)$, so *no false negatives*;

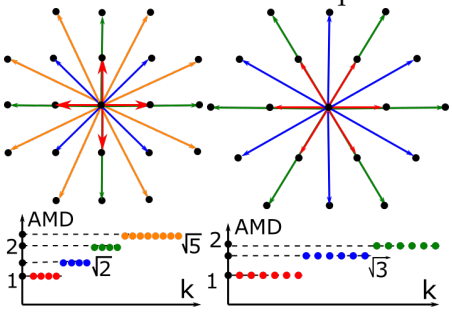
completeness : $I(S) = I(Q) \Rightarrow S \simeq Q$ are isometric, hence *no false positives*;

continuity : $I(S)$ continuously changes under perturbations of S in a distance metric d satisfying the axioms $d(I(S), I(Q)) = 0 \Leftrightarrow S \simeq Q$ are isometric, $d(I, I') = d(I', I)$, inequality $d(I, I') \le d(I, I'') + d(I'', I')$;

any S can be *reconstructed* from $I(S)$.

AMD invariants of crystals

For any p_i (one of m motif points) in a cell of a crystal S, let d_{ik} be the distance to its k-th closest neighbour in the infinite set S. The Average Minimum Distance [1] is $\mathrm{AMD}_k = \frac{1}{m}\sum_{i=1}^{m} d_{ik}$. The square and hexagonal lattices have these AMD sequences:



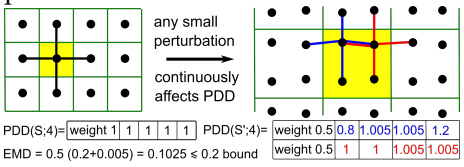
Stronger invariant PDD [1]

Pointwise Distance Distribution

For any motif point p_i, put its distances $d_{i1} \le \cdots \le d_{ik}$ into a row of the $m \times k$ matrix. If j of m rows are identical, collapse them into one row of *weight* j/m. The matrix $\mathrm{PDD}(S;k)$ is an unordered distribution of rows with weights, strictly stronger than the Pair Distribution Function. Increasing k adds more columns to PDD without changing the first columns.

PDD is a continuous invariant

If atoms are perturbed up to ε, then $\mathrm{PDD}(S;k)$ changes up to 2ε in Earth Mover's Distance EMD, which compares PDD matrices of different sizes.

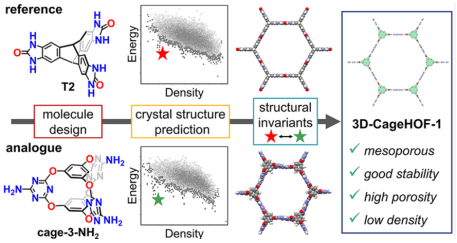


PDD generically complete & fast

Under a tiny perturbation, any crystal becomes *generic*, e.g. has no repeated distances except due to periodicity.

Any generic periodic crystal can be reconstructed, uniquely up to isometry in 3D, from lattice invariants and $\mathrm{PDD}(S;k)$ for a large enough k, and computed in near-linear time in m, k.

New crystal by PDD analogy



PDD can include atom attributes but compares any periodic sets of atoms or molecular centers [2] without fixing a symmetry group or chemistry.

'Needles in a haystack'

More than 200 billion comparisons of AMD and PDD of all 660K+ periodic crystals (no disorder, full 3D structure) in the Cambridge Structural Database for $k = 100$ (now in one hour) on a modest desktop detected 5 pairs of geometric duplicates with one atom replacement [1], which seems physically impossible, e.g.

HIFCAB vs JEPLIA (Cd $\leftrightarrow$ Mn).

Five journals are investigating the integrity of the underlying articles.

Crystal Isometry Principle

Map: periodic crystals $\to$ periodic point sets is *injective* modulo isometry, so any periodic crystal is determined by the geometry of its atomic centers without chemical types. Replacing one atom with a different one should perturb distances to atom neighbors.

Hence all known and undiscovered crystals live in one *Crystal Isometry Space* (CRIS) parametrized by complete isometry invariants. The case of finite atomic clouds is solved in [3].

[1] D.Widdowson, V.Kurlin. Resolving the data ambiguity for periodic crystals. Proceedings NeurIPS 2022.

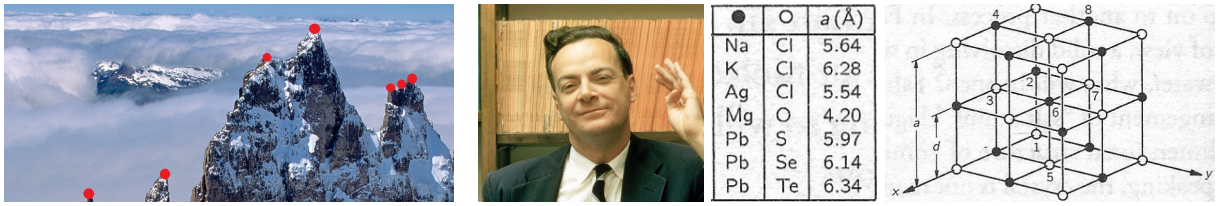
[2] Q.Zhu et al. Analogy powered by prediction and structural invariants. JACS, 2022, 144, 22, 9893–9901.

[3] Widdowson, Kurlin. Recognizing rigid patterns of clouds of unordered points. Proceedings of CVPR 2023.

Geometric Data Science (GDS) develops continuous maps of data objects

The vision is to map (continuously parametrize) the space of any data objects considered up to practical equivalences. While Geometric Deep Learning experimentally outputs equivariant descriptors of clouds or graphs, GDS developed analytic, complete and continuous invariants for any finite and generic periodic sets of unordered points in $\mathbb{R}^n$, see the papers in NeurIPS 2022 and CVPR 2023 at http://kurlin.org/research-papers.php#Geometric-Data-Science.

The key obstacle for periodic crystals was the ambiguity of conventional data based on minimal or reduced cells that are discontinuous under atomic displacements. Without continuously quantifying the crystal similarity, the brute-force Crystal Structure Prediction produces millions of nearly identical approximations to numerous local energy minima, see red peaks in Fig. 1.



●	○	a (Å)
Na	Cl	5.64
K	Cl	6.28
Ag	Cl	5.54
Mg	O	4.20
Pb	S	5.97
Pb	Se	6.14
Pb	Te	6.34

Figure 1: **Left**: energy landscapes show crystals as isolated peaks of height$=-$energy. To see beyond the 'fog', we need a map parametrized by invariant coordinates with a continuous metric. **Right**: R. Feynman's first lecture showed that 7 cubic crystals differ by side lengths, while our invariants distinguished all 850K+ periodic crystals in the CSD. These crystals have unique positions in a common *Crystal Isometry Space* whose one 2D projection is in Fig. 2.

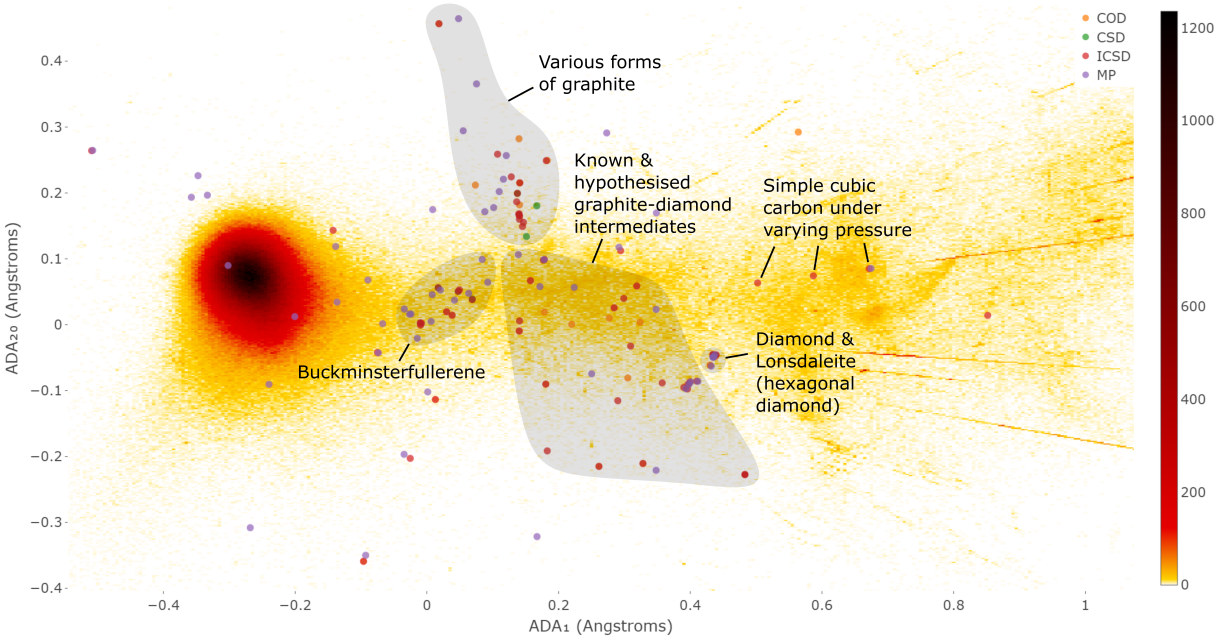


Figure 2: Carbon allotropes on a continuous map of periodic crystals in the Cambridge Structural Database (CSD), Crystallography Open Database (COD), Inorganic Crystal Structure Database (ICSD), and Materials Project (MP). The colour indicates the number of crystals whose AMD_k (average distance to the k-th atomic neighbour) are discretized to each pixel.

The crystal space can be visualized in other explicit coordinates from AMD vectors and PDD matrices. The density of S has been extracted from the asymptotic of $\mathrm{AMD}_k(S)$ as $k \to +\infty$.