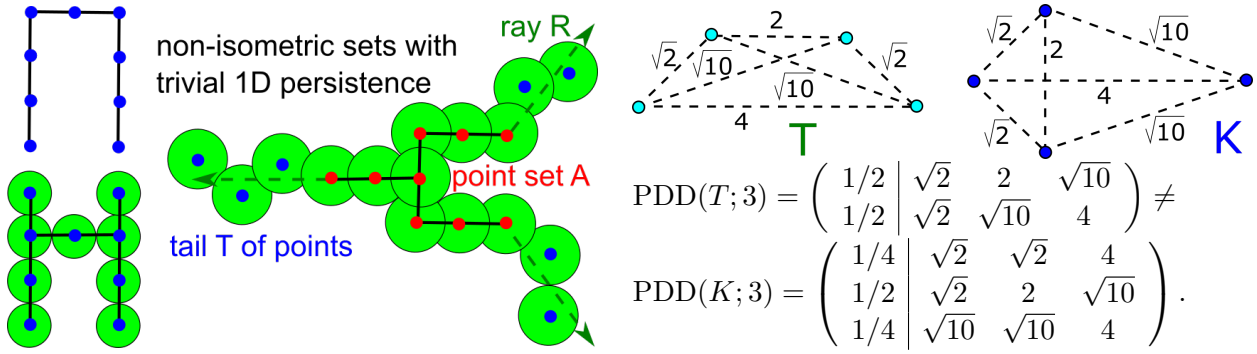
# Persistence vs newer isometry invariants of discrete sets of points

Daniel Widdowson, Philip Smith, Yury Elkin, Vitaliy Kurlin (vitaliy.kurlin@gmail.com)

Computer Science and Materials Innovation Factory, University of Liverpool, UK

If we consider standard filtrations (of Vietoris-Rips, Čech, Delaunay complexes) on a finite set of points $A \subset \mathbb{R}^N$, persistent homology is an isometry invariant of $A$, preserved by all transformations that maintain the inter-point distances [2]. Any set $A$ can be extended [3] to a large family of non-isometric sets $A \cup T$ that have the same 1D persistence as $A$, by adding a 'tail' $T$ of points 'angularly' close to a ray $R$ attached to a corner point $v \in A$ in Figure 1.




$$\mathrm{PDD}(T;3) = \left(\begin{array}{c|ccc} 1/2 & \sqrt{2} & 2 & \sqrt{10} \\ 1/2 & \sqrt{2} & \sqrt{10} & 4 \end{array}\right) \neq$$

$$\mathrm{PDD}(K;3) = \left(\begin{array}{c|ccc} 1/4 & \sqrt{2} & \sqrt{2} & 4 \\ 1/2 & \sqrt{2} & 2 & \sqrt{10} \\ 1/4 & \sqrt{10} & \sqrt{10} & 4 \end{array}\right).$$

Figure 1: **Left**: any cloud on the left can be extended by adding 'tails' $T$ at red corners to a large family of non-isometric clouds with trivial 1D persistence. **Right**: the 4-point sets $T, K$ have the same 6 pairwise distances but are distinguished by the new PDD invariants.

All finite and periodic sets of points $A \subset \mathbb{R}^N$ in general position are distinguished up to isometry by the *Pointwise Distance Distribution* PDD$(A;k)$ defined below [4]. For each point, $p \in A$, write the row of ordered distances to the $k$ nearest neighbors of $p$ in the full set $S$. If $k$ of $m$ points in $A$ have identical rows, collapse them into one row with weight $k/m$.

If each point of $A$ is perturbed by $\varepsilon$, the resulting matrix PDD$(A;k)$ changes up to $2\varepsilon$ in the Earth Mover's Distance (EMD). The neighbor number $k$ can be considered as a degree of approximation because increasing $k$ only adds more columns to PDD$(A;k)$. PDD has a parameterised near-linear time in the number of points by the $k$-nearest neighbor search [1].

The 200B+ comparisons of 660K+ crystals over a couple of days on a modest desktop identified five duplicates in the Cambridge Structural Database [5, section 7], which were missed by all past tools, now under investigation by five journals for data integrity. Hence the new isometry invariants are easier to interpret, faster, and stronger than persistence. The latest versions of authors' papers are at http://kurlin.org/projects/periodic-geometry-topology.php.

[1] Elkin, Y., Kurlin, V.: A new compressed cover tree guarantees a near linear parameterized complexity for all $k$-nearest neighbors search in metric spaces. arxiv:2111.15478 (2021)

[2] Frosini, P., Landi, C.: Size theory as a topological tool for computer vision. Pattern Recognition and Image Analysis **9**(4), 596–603 (1999)

[3] Smith, P., Kurlin, V.: Families of point sets with identical 1d persistence. arxiv:2202.00577

[4] Widdowson, D., Kurlin, V.: Pointwise Distance Distributions. arxiv:2108.04798

[5] Widdowson, D., et al.: Average Minimum Distances of periodic point sets - invariants for mapping periodic crystals. MATCH Comm. Math. Comp. Chemistry **87**, 529–559 (2022)

## **Geometric Data Science** extends TDA and Periodic Geometry

The area of Geometric Data Science aims to resolve the big data challenges by describing continuous metrics on spaces of discrete objects up to practical equivalences such as rigid motion or isometry of periodic crystals whose structures are determined in a rigid form.

The necessity of a continuous metric is clearer for periodic crystals whose conventional representations by reduced cells are discontinuous under the ever-present atomic vibrations. Without continuously quantifying the crystal similarity, the brute-force Crystal Structure Prediction produces millions of nearly identical approximations to numerous local energy minima.

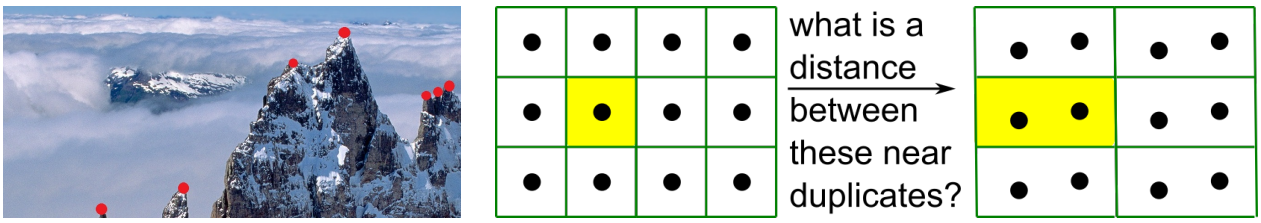



Figure 2: **Left**: energy landscapes show crystals as isolated peaks of height=-energy. To see beyond the 'fog', we need a map with invariant coordinates and continuous distances satisfying metric axioms. **Right**: most crystal invariants are discontinuous, a minimal cell can double.

**Problem**: isometry classification of discrete sets with continuous metrics and fast algorithms. Find a function $I$ on all finite or periodic sets of unlabeled points in $\mathbb{R}^n$ such that

(a) *invariance* : if sets $S \cong Q$ are isometric, then $I(S) = I(Q)$, so $I$ has *no false negatives*;

(b) *completeness* : if $I(S) = I(Q)$, then $S \cong Q$ are isometric, so $I$ has *no false positives*;

(c) *metric* : a distance $d$ between values of $I$ satisfies all axioms; $d(I_1, I_2) = 0$ if and only if $I_1 = I_2$, symmetry $d(I_1, I_2) = d(I_2, I_1)$, triangle inequality $d(I_1, I_3) \le d(I_1, I_2) + d(I_2, I_3)$;

(d) *continuity* : if $Q$ is obtained from a point set $S \subset \mathbb{R}^n$ by shifting each point of $S$ by at most $\varepsilon$, then $d(I(S), I(Q)) \le C\varepsilon$ for a fixed constant $C$ and any such point sets $S, Q \subset \mathbb{R}^n$;

(e) *computability* : the invariant $I$, the metric $d$ and verification of $I(S) = I(Q)$ should be computable in a near-linear time in the number of given points for a fixed dimension $n$;

(f) *inverse design* : any point set $S \subset \mathbb{R}^n$ can be reconstructed from its invariant $I(S)$.

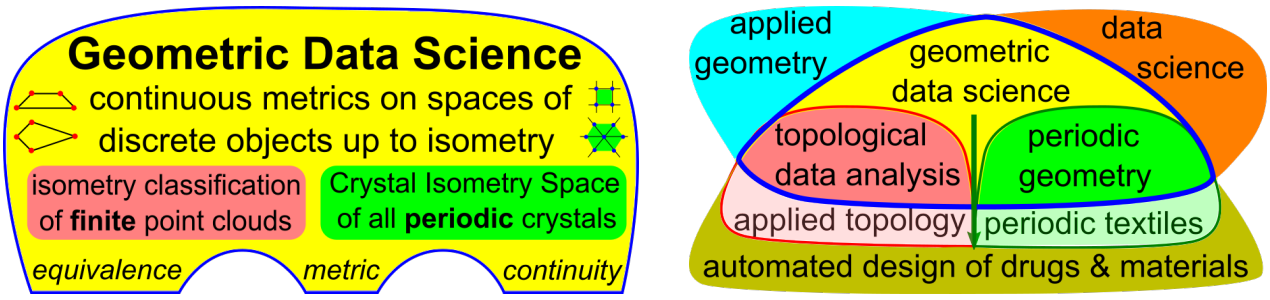



Figure 3: **Left**: GDS develops continuous metrics on isometry classes of data objects. **Right**: Periodic Geometry justified the Crystal Isometry Principle (CRISP): all real crystals live in the common Crystal Isometry Space continuously extending Mendeleev's table of elements.