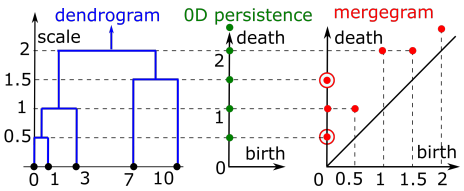
Persistence vs easier, faster, and stronger invariants

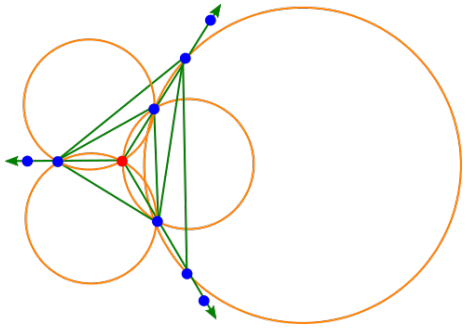
Philip Smith, Daniel Widdowson, Yury Elkin, Vitaliy Kurlin, http://kurlin.org,
Computer Science and Materials Innovation Factory, University of Liverpool, UK

Persistence is an *isometry invariant* of a cloud $S \subset \mathbb{R}^n$ of unlabeled points for standard filtrations of Vietoris-Rips, Cech, and Delaunay complexes. How strong is persistence as an isometry invariant of a cloud?

0D persistence for a point set extends to the stronger *mergegram*, which is continuous in the bottleneck distance and has the same asymptotic time [1].

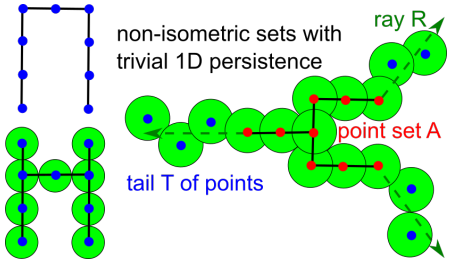


In the Delaunay-based 1D persistence, each (birth,death) comes from an acute Delaunay triangle with circumradius = death because all non-acute triangles enter the filtration at half-length of their longest edge. If all Delaunay triangles are *non-acute*, the resulting 1D persistence is trivial.



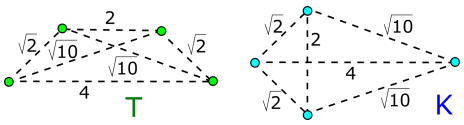
A huge generic family of point sets have identical or even trivial 1D persistence.

Any point set $S \subset \mathbb{R}^n$ can be extended [1] to a large family of non-isometric sets $S \cup T$ that have the same 1D persistence as S, by adding a 'tail' T of points 'angularly' close to a ray R attached to a 'corner' point $v \in S$.



New isometry invariants

For each point, $p \in S$, write the row of ordered distances to the k nearest neighbors of p in the full (discrete or periodic) set S. If k of m points in S have identical rows, collapse them into one row with weight k/m.

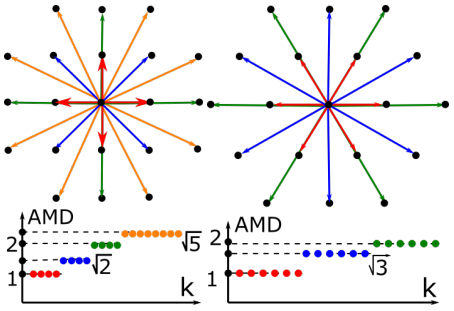


$$\mathrm{PDD}(T;3) = \left(\begin{array}{c|ccc} 1/2 & \sqrt{2} & 2 & \sqrt{10} \\ 1/2 & \sqrt{2} & \sqrt{10} & 4 \end{array}\right) \neq$$

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

The matrix of rows with weights in the extra column is the Pointwise Distance Distribution $\mathrm{PDD}(S;k)$.

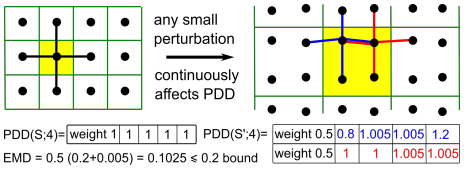
The 4-point sets T, K have the same 6 pairwise distances but are distinguished by PDD [4] quickly computed due to a fast k-nearest neighbor search [5]. By taking the weighted average of each column in $\mathrm{PDD}(S;k)$, we get the *Average Minimum Distance* [1] $\mathrm{AMD}_k = \frac{1}{m}\sum_{i=1}^{m} d_{ik}$. The square and hexagonal lattices have these AMDs:



These invariants are defined for periodic sets modeling all crystals whose structures are rigidly determined.

PDD: continuous invariants

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



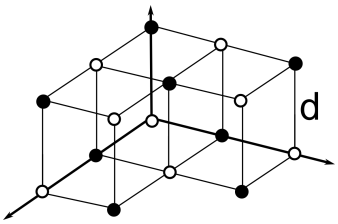
PDD : generically complete

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

Any such crystal is uniquely reconstructed from lattice invariants and $\mathrm{PDD}(S;k)$ for a large enough k [4].

R. Feynman geometrically distinguished a few crystals below, which we have done for all real materials.

●	○	d (Å)
Na	Cl	2.82
K	Cl	3.14
Ag	Cl	2.77
Mg	O	2.10
Pb	S	2.98
Pb	Se	3.07
Pb	Te	3.17



200B+ pairwise comparisons of AMD, PDD for all 660K+ periodic crystals (no disorder, full 3D structure) in the CSD over two days on a modest desktop detected five pairs of suspicious entries with identical geometry and one atom replacement [3, 4], e.g. HIFCAB and JEPLIA (Cd ↔ Mn),...

Five journals are investigating the integrity of the underlying articles.

Crystal Isometry Principle:

periodic crystals → periodic point sets is *injective* on isometry classes.

Any periodic crystal is determined by geometry of its atomic centres (without chemical labels) because replacing one atom with a different one should perturb distances to atom neighbors.

All known and undiscovered crystals live in the common *Crystal Isometry Space* parameterized by invariants.

[1] Y.Elkin, V.Kurlin. The mergegram of a dendrogram and its stability. Proceedings of MFCS 2020.

[2] P.Smith, Kurlin. Sets with identical 1D persistence. arXiv:2202.00577.

[3] D.Widdowson et al. Average Minimum Distances of periodic sets. MATCH, v.87(3), p.529-559, 2022.

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

[5] Y.Elkin. New compressed cover tree for k-nearest neighbor search. PhD thesis, arXiv:2205.10194.

Geometric Data Science extends TDA and Geometric Deep Learning

The area of Geometric Data Science aims to solve the data challenges by defining continuous metrics on moduli spaces of discrete objects up to practical equivalences such as rigid motion or isometry, e.g. for all 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.

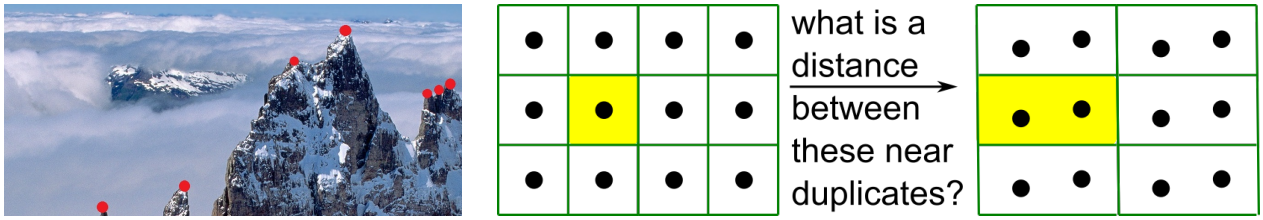


Figure 1: **Left**: energy landscapes show crystals as isolated peaks of height=-energy. To see beyond the 'fog', we need a map with invariant coordinates and continuous distances satisfying metric axioms. **Right**: most crystal invariants are discontinuous, a minimal cell can double.

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

(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 $d(I_1, I_2) + d(I_2, I_3) \geq d(I_1, I_3)$;

(d) *continuity* : if Q is obtained from a point set $S \subset \mathbb{R}^n$ by shifting each point of S by at most ε, then $d(I(S), I(Q)) \leq C\varepsilon$ 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 $I(S)$.

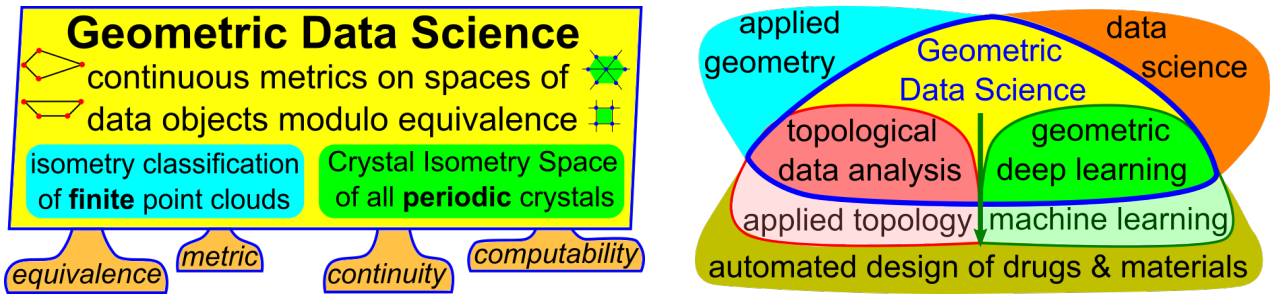


Figure 2: **Left**: Geometric Data Science (GDS) is based on equivalence, metric continuity, and computability. The major breakthrough is the Crystal Isometry Principle (CRISP): all real crystals live in the common Crystal Isometry Space continuously extending Mendeleev's periodic table of elements. **Right**: GDS extends Topological Data Analysis studying persistence of cycles hidden in data and Geometric Deep Learning for data with non-Euclidean metrics.