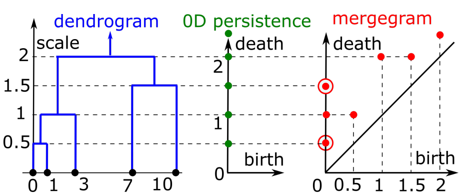
Persistence vs easier, faster, and stronger invariants

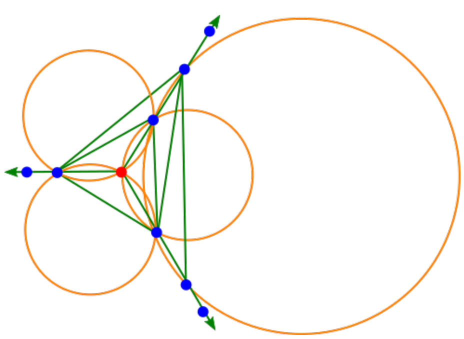
Philip Smith, Daniel Widdowson, Yury Elkin, Vitaliy Kurlin, http://kurlin.org,
Computer Science and Materials Innovation Factory, University of Liverpool, UK

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].

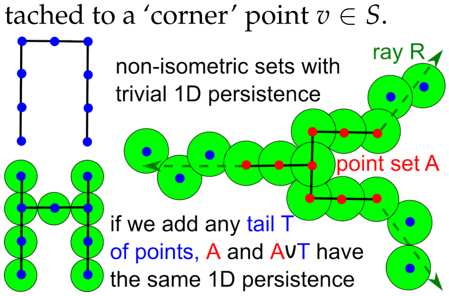


In the Delaunay-based 1D persistence, each (birth,death) comes from an acute Delaunay triangle with circumradius = death because all non-acute triangles enter the filtration at half-length of their longest edge. If all Delaunay triangles are *non-acute*, the resulting 1D persistence is trivial.



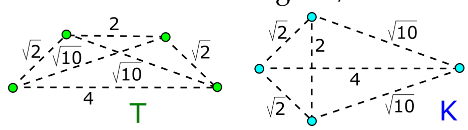
A huge generic family of point sets have identical or even trivial 1D persistence.

Any point set $S \subset \mathbb{R}^n$ can be extended [2] 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.

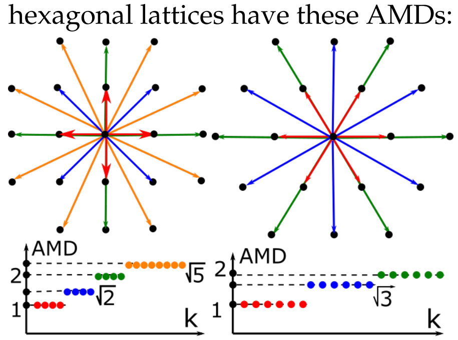


$$\mathrm{PDD}(T;3) = \left(\begin{array}{c|ccc} 1/2 & \sqrt{2} & 2 & \sqrt{10} \\ 1/2 & \sqrt{2} & \sqrt{10} & 4 \end{array}\right) \neq$$

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

The Pointwise Distance Distribution $\mathrm{PDD}(S;k)$ is the matrix of rows with weights in the extra column, extended to complete invariants in $\mathbb{R}^n$ [3].

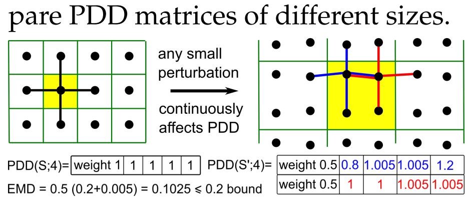
The 4-point sets T, K have the same 6 pairwise distances but are distinguished by PDD [4] quickly computed due to a fast k-nearest neighbor search [5]. By taking the weighted average of each column in $\mathrm{PDD}(S;k)$, we get the *Average Minimum Distance* [4] $\mathrm{AMD}_k = \frac{1}{m}\sum_{i=1}^{m}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



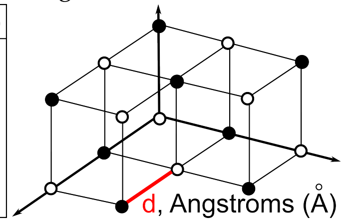
PDD : generically complete

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].

Feynman's first lecture geometrically distinguished 7 crystals below, now extended to all periodic crystals in the CSD: Cambridge Structural Database.

●	○	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+ comparisons of all 660K+ periodic crystals in the CSD over two days on a modest desktop detected 5 pairs of entries with identical geometry and one atom replacement [4], e.g. HIFCAB vs 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.

Map: periodic crystals → periodic point sets is *injective* modulo isometry, so any periodic crystal is determined by its atomic geometry without chemical types. Hence all known and undiscovered crystals live in one *Crystal Isometry Space* parametrized by complete isometry invariants.

[1] Y.Elkin, V.Kurlin. The mergegram of a dendrogram and its stability. Mathematics, 9 (17), 2021.

[2] P.Smith, V.Kurlin. APCT 2024, doi:10.1007/s41468-024-00177-6.

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

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

[5] Y.Elkin. V.Kurlin. A new compressed cover tree for k-nearest neighbor search. Proceedings ICML 2023.

Geometric Data Science extends TDA and Geometric Deep Learning

Geometric Data Science aims to continuously parametrize moduli spaces of discrete objects (finite and periodic sets and graphs) up to important relations (rigid motion and isometry), e.g. the space of all periodic crystals whose structures are determined in a rigid form.

The key obstacle for crystals was the ambiguity of conventional representations based on reduced cells, which are discontinuous under perturbations. Without continuously quantifying the crystal similarity, the brute-force Crystal Structure Prediction produces millions of nearly identical approximations to numerous local energy minima, see the red peaks in Fig. 1 (left).

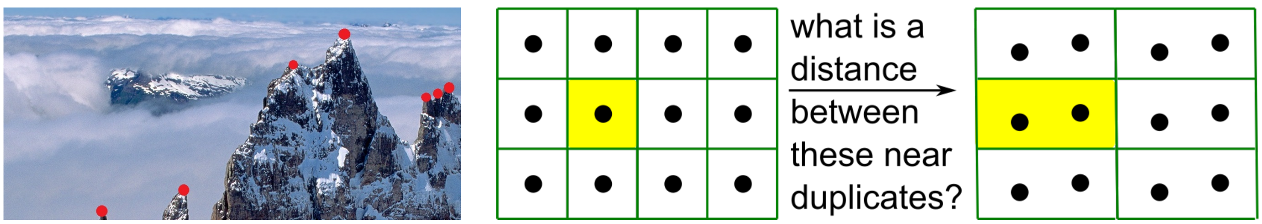


Figure 1: **Left**: energy landscapes show crystals as isolated peaks of height=-energy. To see beyond the ‘fog’, we need a map with invariant coordinates and continuous distances satisfying metric axioms. **Right**: all cell-based invariants including symmetries are discontinuous.

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

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

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

(c) *continuity* : if any point of a set $S \subset \mathbb{R}^n$ is perturbed by at most ε, the invariant I changes up to $\lambda\varepsilon$ for a fixed constant λ and a suitable metric d on invariant values satisfying the metric axioms: (1) coincidence $d(I(S), I(T)) = 0$ if and only if $S \simeq T$ are isometric, (2) symmetry $d(I, I') = d(I', I)$, (3) triangle inequality $d(I, I') + d(I', I'') \geq d(I, I'')$ for any values I, I', I'';

(d) *inverse design* : parametrize all $I(S)$ that can be efficiently inverted to a set $S \subset \mathbb{R}^n$;

(e) *computability* : I, d, the verification of $I(S) = I(T)$, and reconstruction of S from $I(S)$ should be obtained in polynomial time in the number of given points for a fixed dimension n.

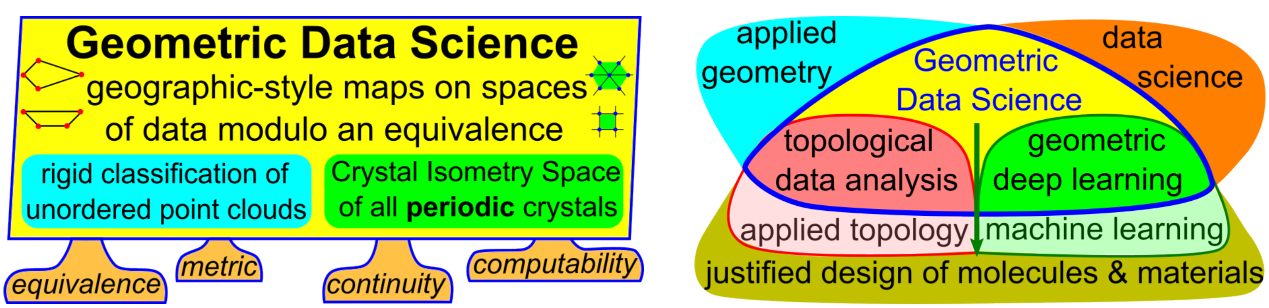


Figure 2: **Left**: Geometric Data Science (GDS) is based on equivalence, metric continuity, and computability. The first breakthrough is the *Crystal Isometry Principle* (CRISP): all periodic crystals live in a common Crystal Isometry Space continuously parametrized by complete invariants. **Right**: GDS extends Topological Data Analysis, which studies the persistence of cycles in data, and Geometric Deep Learning, which experimentally searches for invariants.