Pointwise Distance Distributions of periodic point sets

Daniel Widdowson
Department of Computer Science
and Materials Innovation Factory
University of Liverpool, UK

Vitaliy Kurlin
Materials Innovation Factory
University of Liverpool, UK
vitaliy.kurlin@gmail.com

Abstract

The fundamental model of all solid crystalline materials (periodic crystals) is a periodic set of atomic centers considered up to rigid motion in Euclidean space. The major obstacle to materials discovery was highly ambiguous representations that didn't allow fast and reliable comparisons, and led to numerous (near-) duplicates in all experimental databases. This paper introduces the new invariants that are crystal descriptors without false negatives and are called Pointwise Distance Distributions (PDD). The PDD invariants are numerical matrices with a near-linear time complexity and an exactly computable metric. The strongest theoretical result is generic completeness (absence of false positives) for all finite and periodic sets of points in any dimension. The strength of PDD is demonstrated by 200B+ pairwise comparisons of all 660K+ periodic structures from the world's largest Cambridge Structural Database of 1.17M+ known crystals over two days on a modest desktop.

1 Motivations for resolving the data ambiguity challenge in Problem 1.1

This paper resolves the long-standing challenge of ambiguous data representation for periodic structures that model all solid crystalline materials (crystals). Any real crystal is best modeled as a periodic set $S \subset \mathbb{R}^n$ of points at all atomic centers, whose positions have a physical meaning and are determined via X-ray diffraction patterns. Edges between points are excluded because they only abstractly represent inter-atomic bonds that depend on thresholds for distances and angles [32].

The simplest example is a *lattice* $\Lambda \subset \mathbb{R}^n$ consisting of all integer linear combinations of a basis whose vectors span a *unit cell* U, whose translational copies are shown in Fig. 1 only for convenience.

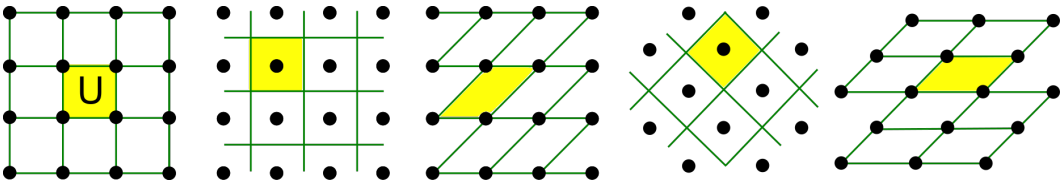

Figure 1: These isometric lattices are given by different cells and motifs. **1st**: $U = \langle (1,0),(0,1) \rangle$, $M = \{(0,0)\}$. **2nd**: $U = \langle (1,0),(0,1) \rangle$, $M = \{(\frac{1}{2},\frac{1}{2})\}$. **3rd**: $U = \langle (1,0),(1,1) \rangle$, $M = \{(0,0)\}$. **4th**: $U = \langle (\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}),(-\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}) \rangle$, $M = \{(\frac{1}{2},\frac{1}{2})\}$. **5th**: $U = \langle (\sqrt{2},0),(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}) \rangle$, $M = \{(0,0)\}$.

Materials discovery still relies on trial-and-error because periodic crystals are traditionally represented by non-invariants (descriptors with false negatives) or discontinuous invariants such as symmetry groups that break down under tiny perturbations. These conventional descriptions cannot identify fraudulent structures in experimental datasets that keep depositing numerous (near-)duplicates without reliable tools for justified comparisons [33]. The ambiguity challenge will be rigorously stated in Problem 1.1 as a classification of periodic sets up to isometry preserving the rigid form of crystals.

Preprint. Under review.

Fig. 1 illustrates the first obstacle: the same lattice can be generated by infinitely many different bases or unit cells. So distinguishing only lattices up to isometry is already non-trivial. Then any *periodic point set* S is a sum $\Lambda + M = \{\vec{u} + \vec{v} : u \in \Lambda, v \in M\}$, where a *motif* M is a finite set of points in the basis of U. Any lattice Λ is considered as a periodic set with a 1-point motif $M = \{p\}$. A single point p can be arbitrarily chosen in a unit cell U as in the first two pictures of Fig. 1. Basis vectors of U and atomic coordinates of motif points (atomic centers) in M form a conventional Crystallographic Information File (CIF). Fig. 2 (left) shows the ambiguity of the CIF pair (Λ, M) even if a basis of U is fixed. The recent work by Edelsbrunner et al [24] initiated a new research area in classifications of periodic point sets up to isometry. An *isometry* of Euclidean space $\mathbb{R}^n$ is any map that maintains inter-point distances. Any orientation-preserving isometry can be realized as a continuous rigid motion, for example any composition of translations and rotations in $\mathbb{R}^3$. This equivalence is most natural for periodic point sets that represent real rigid structures.

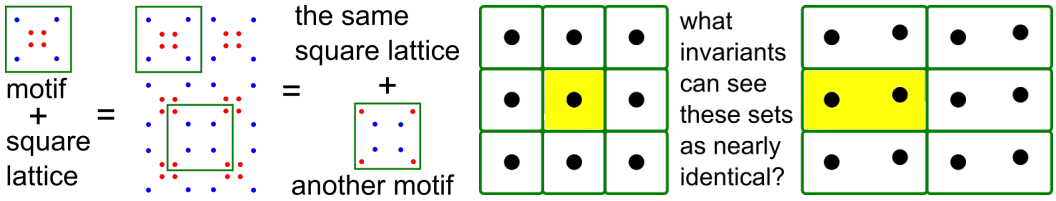

Figure 2: **Left** : even for a fixed cell of a lattice Λ, different motifs M can define isometric periodic sets $\Lambda + M$. **Right**: for almost any perturbation, the symmetry group and (the minimum volume of) any reduced cell discontinuously change, which justifies continuity (1.1d) in Problem 1.1.

Crystals can be reliably distinguished up to isometry only by an isometry invariant that takes the same value on all isometric sets, hence having no false negatives. If a descriptor allows false negatives, we can make *no reliable conclusions* because equivalent objects can have different representations as in Fig. 1 and 2 (left). Hence, non-invariants such as edge-lengths and angles of a unit cell, or coordinates of motif points in a cell basis cannot be used to justifiably compare crystals [6]. It suffices to classify up to isometry including mirror reflections. For any basis $v_1, \dots, v_n$ of $\mathbb{R}^n$, an isometry f preserves orientation if $f(v_1), \dots, f(v_n)$ in the same basis have a determinant $\det(f(v_1), \dots, f(v_n)) > 0$.

The traditional approach to identifying a periodic crystal is to use its conventional or reduced cell [46, section 9.3]. This reduced cell has been known to be discontinuous under perturbations [1] even for lattices when a motif M is a single point. More formally, [24, section 1] and [34, Theorem 15] proved that a continuous reduced cell cannot be defined for all lattices. For more general periodic sets, discontinuity of many past discrete invariants such as symmetry groups becomes clearer in Fig. 2 (right) showing that even real-valued invariants struggle to continuously quantify the similarity between nearly sets. The minimum volume of a cell U can easily double, while the density (the number or mass of points divided by the cell volume) remains constant under perturbations of points.

The continuous isometry classification of periodic sets has been an open problem since 1980 [1].

Problem 1.1. *Find a function I on all periodic sets of unlabeled points in $\mathbb{R}^n$ such that*

(1.1a) invariance *: if any periodic point sets $S \cong Q$ are isometric in $\mathbb{R}^n$, then $I(S) = I(Q)$, so the invariant I has* no false negatives*;*

(1.1b) completeness *: if $I(S) = I(Q)$ for any periodic point sets S, Q, then $S \cong Q$ are isometric, so the invariant I has* no false positives*;*

(1.1c) metric *: a distance d between values of I satisfies all axioms; 1) $d(I_1, I_2) = 0$ if and only if $I_1 = I_2$, 2) symmetry $d(I_1, I_2) = d(I_2, I_1)$, 3) triangle inequality $d(I_1, I_3) \le d(I_1, I_2) + d(I_2, I_3)$;*

(1.1d) continuity *: if Q is obtained from a periodic set $S \subset \mathbb{R}^n$ by shifting each point of S by at most ε, then $d(I(S), I(Q)) \le C\varepsilon$ for a fixed constant C and any such periodic point sets $S, Q \subset \mathbb{R}^n$;*

(1.1e) computability *: the invariant I, the metric d and verification of $I(S) = I(Q)$ should be computed in a near-linear time in the number of motif points of periodic sets for a fixed dimension n;*

(1.1f) inverse design *: any periodic point set $S \subset \mathbb{R}^n$ can be reconstructed from its invariant $I(S)$.* ■

Problem 1.1 is the ultimate *Data Science challenge* for all periodic crystals S whose non-invariant input (a cell basis and a motif) should be transformed into a complete invariant $I(S)$, which uniquely and unambiguously represents any S. Such a complete invariant can be considered as a materials genome [20] or a DNA-type code that also allows an explicit reconstruction for any periodic crystal.

For example, Computer Vision tries to identify humans and other objects such as road signs by using pixel-based images as input. Similar to other real objects, any periodic crystal can be given by (infinitely) many inputs. Problem 1.1 is an example of the data ambiguity challenge for almost any application that needs to uniquely identify a real object. Biology has partially solved this challenge by a DNA code, though a living creature cannot be easily created from a given code. Condition (1.1f) goes further and requires a full reconstruction of a periodic crystal S from its DNA-style code $I(S)$.

The proposed solution to Problem 1.1 is the isometry invariant I called the Pointwise Distance Distribution PDD. Theorems 3.2, 4.3, 5.1 , 4.4 prove that PDD satisfies all conditions of Problem 1.1, even (1.1b) at least for generic sets. More exactly, Theorem 4.4 shows that any periodic point set $S \subset \mathbb{R}^n$ in general position can be explicitly reconstructed from PDD and lattice invariants.

The strength of PDD was experimentally checked for all 660K+ periodic crystals in the world's largest Cambridge Structural Database (CSD). Despite the CSD being curated to contain only real and distinct structures [33], the new invariants identified several pairs of duplicates. All the underlying publications are now being investigated for data integrity by five journals, see details in section 6.

Problem 1.1 is stated in the hardest scenario when points are unordered and unlabeled because many real crystals have identical compositions. For example, diamond and graphite (whose 2-dimensional layer is famous graphene) consist of pure carbon but have vastly different physical properties.

Conditions (1.1cd) for a continuous metric are stronger than a complete classification in (1.1ab): detecting an isometry gives a discontinuous metric $d(S,Q) = 1$ (or another positive number) for all non-isometric $S \not\cong Q$ even if S, Q are near duplicates as in Fig. 2 (right). Continuity under perturbations is practically important because atoms vibrate, and any real measurement of a crystal produces slightly different coordinates of its unit cell and motif. Any simulation of periodic structures introduces floating point errors because of inevitable approximations by iterative optimization.

Thousands of near-duplicates are routinely produced, though only a few structures are synthesized. Five real structures of 5679 predicted on a supercomputer over 12 weeks are a typical example [32]. This 'embarrassment of over-prediction' wastes resources and time to run simulations and then analyze results, often by visual inspection, because there were no invariants that resolve Problem 1.1.

Computability condition (1.1e) avoids the trivial function $I(S) = S$ in Problem 1.1. Inverse design in (1.1f) allows one to replace the traditional blind sampling (of ambiguous cells and motifs leading to (near-)duplicates via optimization) with a guided exploration of the crystal space parameterized by complete and reversible invariants. Section 2 shows that the state-of-the-art tools remain stuck with conditions (1.1ab) while the new invariants satisfy the stronger practical requirements (1.1cdef).

2 A review of the related state-of-the-art on comparing periodic point sets

Any point p in $\mathbb{R}^n$ can be identified with the vector $\vec{p}$ from the origin 0 to p. The Euclidean distance between $p, q \in \mathbb{R}^n$ is denoted by $|p-q|$, which is the length of the difference vector $\vec{p} - \vec{q}$.

Classical methods of Computational Geometry. Isometry questions related to Problem 1.1 were extensively studied for finite sets of points. The fastest algorithm to test an isometry between m-point sets in $\mathbb{R}^n$ runs in time $O(m^{\lceil n/3 \rceil} \log m)$ by [12], which can be improved to $O(m \log m)$ in $\mathbb{R}^4$ [48]. Significant results on matching bounded rigid shapes and registration of finite point sets were obtained in [64, 38, 35, 23]. These methods focus on true/false answers without continuous invariants.

The classical isometry invariant of a finite set in any metric space is the distribution of pairwise distances who completeness in general position was proved in 2004 [11]. The same paper described an infinite family of singular counter-examples to the full completeness, see the non-isometric 4-point sets in Fig. 3 (left). Persistent homology is a more recent isometry invariant of finite point sets, which is continuous under perturbation but turned out to be weaker than previously anticipated [56, 26, 27, 67]. Mémoli's seminal work on *distributions of distances* [53], also known as *shape*

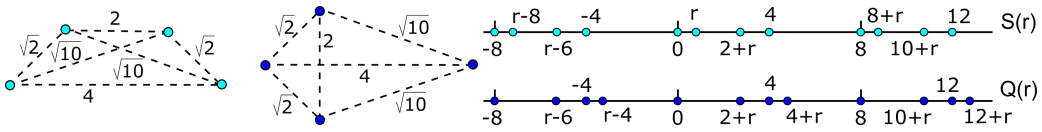

Figure 3: **Left**: point sets $K = \{(\pm 2, 0),\ (\pm 1, 1)\}$ and $T = \{(\pm 2, 0), (-1, \pm 1)\}$ can not be distinguished by their six pairwise distances $\sqrt{2}, \sqrt{2}, 2, \sqrt{10}, \sqrt{10}, 4$. **Right**: 1D periodic sets $S(r) = \{0, r, 2 + r, 4\} + 8\mathbb{Z}$ and $Q(r) = \{0, 2 + r, 4, 4 + r\} + 8\mathbb{Z}$ for $0 < r \leq 1$ have the same Patterson function [59, p. 197, Fig. 2]. All these pairs are distinguished by PDD in section 3.

distributions [55, 8, 44, 40, 36], for bounded metric spaces is closest to the proposed Pointwise Distance Distributions (PDD) for periodic point sets. However, Problem 1.1 is not reducible to the finite sets by taking a cube or a ball of a fixed (even very large) cut-off radius within a periodic point set. Indeed, one can easily find non-isometric subsets of the same lattice as in Fig. 4 (left).

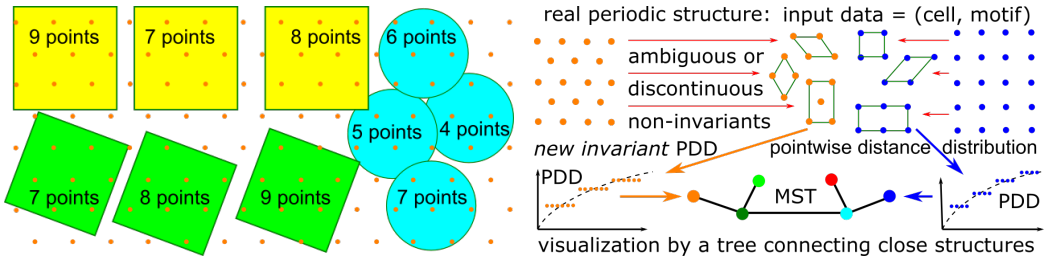

Figure 4: **Left**: isometric periodic sets have many non-isometric finite subsets that cannot be used for comparison, discontinuity under perturbations is similar to Fig. 2 (right). **Right**: an ambiguous input is transformed into the invariant PDD to visualize any dataset as a Minimum Spanning Tree (MST).

Traditional methods of Crystallography and Materials Science. Numerous crystal descriptors include quantities such as cell parameters and atomic coordinates that are not isometry invariants, hence cannot help in Problem 1.1. Many approaches produce finite subsets by using a fixed number of neighbors [16] or a fixed cut-off radius [7, 6, 43]. A reduction to one subset can never guarantee a complete and continuous classification because, under tiny perturbations, a *primitive* (minimal by volume) cell can become larger than any bounded subset of a fixed size, see Fig. 2 (right).

Another inconvenience of the tools such as SOAP (Smooth Overlapped Atomic Positions [6]) and ACSF (Atom-Centered Symmetry Functions [7]) is their dependence on parameters such as cut-off radii that require manual choices. If these parameters change, then so do the underlying invariants.

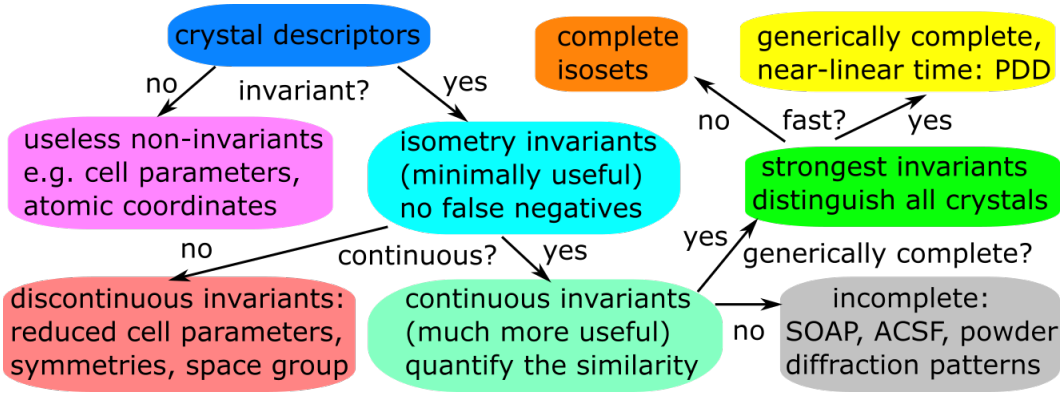

Figure 5: A hierarchy of crystal descriptors. A minimally useful invariant should have *no false negatives*. Most isometry invariants are discontinuous [71] or incomplete [57], hence fail conditions (1.1bcd) in Problem 1.1. The fastest and generically complete invariants are PDD in this paper.

One of the oldest invariants of crystals is the X-ray diffraction pattern whose single crystal form is best for determining a 3D structure of an experimental crystal [19]. Since not all materials can be grown as single crystals, the powder X-ray diffraction pattern (PXRD) is more common. All periodic structures with identical PXRDs are called *homometric*,[62], see the periodic versions $S(1) = \{0, 1, 3, 4\} + 8\mathbb{Z}$ and $Q(1) = \{0, 3, 4, 5\} + 8\mathbb{Z}$ of the 4-point set T, K in Fig. 3. The more general sets $S(r), Q(r)$ with identical Pair Distribution Functions (PDF) will be distinguished by PDD in section 3.

The Mercury software visually compares periodic structures [16] by minimizing the Root Mean Square Deviation (RMSD) of atomic positions from up to a given number m (15 by default) of closest molecules in two structures. This comparison depends on many parameters (maximum number of matched molecules, thresholds for matched distances and angles), fails the triangle inequality in condition (1.1c), and is too slow for pairwise comparisons, see section 6. In fact, the above RMSD tries to measure the maximum displacement of atoms only by using finite subsets of crystals.

For full infinite crystals, the theoretically better alternative is the *bottleneck distance* $d_B(S, Q)$ equal to the maximum Euclidean distance needed to perturb every point $p \in S$ to its unique match in Q. Formally, $d_B(S, Q) = \inf\limits_{g:S\to Q} \sup\limits_{p\in S} |p - g(p)|$ is minimized over all bijections g and maximized over all points $p \in S$ for a fixed bijection. But d_B is not a metric on all periodic sets because $d_B = +\infty$ for nearly identical lattices. Also, d_B is discontinuous under perturbations of lattice bases.

Example 2.1 (infinite bottleneck)**.** *We show that $S = \mathbb{Z}$ and $Q = (1 + \delta)\mathbb{Z}$ for any $\delta > 0$ have $d_B(S, Q) = +\infty$. Assuming that $d_B(S, Q)$ is finite, consider an interval $[-N, N] \subset \mathbb{R}$ containing $2N + 1$ points of S. If there is a bijection $g : S \to Q$ such that $|p - g(p)| \le d_B$ for all points $p \in S$, the image of $2N + 1$ points $S \cap [-N, N]$ under this bijection g should be within the interval $[-N - d_B, N + d_B]$. The last interval contains only $1 + \frac{2(N+d_B)}{1+\delta}$ points, which is smaller than $1 + 2N$ when $\frac{N+d_B}{1+\delta} < N$ or $d_B < \delta N$. We get a contradiction by choosing a large $N > \frac{d_B}{\delta}$.* ■

If we consider only periodic point sets $S, Q \subset \mathbb{R}^n$ with the same density (or unit cells of the same size), the bottleneck distance $d_B(S, Q)$ takes only finite values and becomes a well-defined wobbling distance [37], but remains discontinuous as shown below.

Example 2.2 (discontinuous wobbling)**.** *Slightly perturb the basis $(1, 0), (0, 1)$ of the integer lattice $\mathbb{Z}^2$ to the basis vectors $(1, 0), (\varepsilon, 1)$ of the new lattice Λ. We prove that $d_B(\Lambda, \mathbb{Z}^2) \ge \frac{1}{3}$ for any $\varepsilon > 0$. Map $\mathbb{R}^2$ by $\mathbb{Z}^2$-translations to the unit square $[0, 1]^2$ with identified opposite sides (a torus). Then the whole square lattice $\mathbb{Z}^2$ is mapped to the single point represented by the corners of the square $[0, 1]^2$. The perturbed lattice Λ maps to the sequence of points $\{k\varepsilon \pmod 1\}_{k=0}^{+\infty} \times \{0, 1\}$ in the horizontal edges. If $d_B(\Lambda, \mathbb{Z}^2) = r < \frac{1}{3}$, then all above points should be covered by the closed disks of the radius r centered at the corners of $[0, 1]^2$. For $0 < \varepsilon < \frac{1}{3} - r$, we can find $k \in \mathbb{Z}$ so that $k\varepsilon$ is strictly between $r, 1 - r$, hence not covered by these disks, so $d_B(\Lambda, \mathbb{Z}_2) \ge \frac{1}{3}$.* ■

The discontinuity under perturbations is the major weakness of many past invariants including Voronoi diagrams, which should be matched via infinitely many rotations [54], space groups and other group-theoretic invariants [42]. The key example in Fig. 2 (right) shows that a continuous distance between nearly identical sets should be close to 0, not identically 0. These sets have different symmetries and can be related only by pseudo-symmetries depending on manual thresholds [71].

Recent methods of Periodic Geometry. For any $k \ge 1$, the k-th *density function* $\psi_k(t)$ of a periodic point set $S = \Lambda + M \subset \mathbb{R}^3$ is the total volume of the regions within the unit cell U of Λ covered by exactly k balls $B(p; t)$ with a radius $t \ge 0$ and centres at motif points $p \in M$, divided by the unit cell volume $\mathrm{Vol}[U]$. The density function $\psi_k(t)$ was proved in [24] to be invariant under isometry, continuous under perturbations, complete for periodic sets satisfying certain conditions of general position in $\mathbb{R}^3$, and computable in time $O(mk^3)$, where m is the motif size of S. Section 5 in [24] gave the counter-example to completeness: the 1-dimensional periodic sets $S_{15} = X + Y + 15\mathbb{Z}$ and $Q_{15} = X - Y + 15\mathbb{Z}$ for $X = \{0, 4, 9\}$ and $Y = \{0, 1, 3\}$ [45, section 4] have the same density functions for all $k \ge 1$ [5, Example 10] but were distinguished in [34, Example 5b].

The latest advance [4] reduces the isometry classification of all periodic point sets to an *isoset* of isometry classes of α-clusters around points in a motif at a certain radius α, which was motivated by the seminal work of Dolbilin with co-authors about Delone sets [22, 10, 21]. The continuous metrics on isosets [3, section 5] have only an approximate algorithm. Despite the progress in dimension 1 [51] and lattices in dimension 2 [50, 13] and 3 [49, 14], Problem 1.1 remains open in general.

3 The Pointwise Distance Distribution of a finite or periodic point set

Distances to neighbors were considered in [34, Definition 5], though only their average was proved to be invariant. Definition 3.1 introduces the matrix PDD whose ordered rows guarantee invariance in Theorem 3.2 and weights of rows that guarantee continuity under perturbations in Theorem 4.3.

Definition 3.1 (Pointwise Distance Distribution PDD)**.** *For a motif of points $M = \{p_1, \dots, p_m\}$ in a unit cell U of a lattice Λ, let $S \subset \mathbb{R}^n$ be a finite set coinciding with M or a periodic set $S = \Lambda + M$. For an integer $k \geq 1$, consider the $m \times k$ matrix $D(S;k)$, whose i-th row consists of the ordered distances $d_{i1} \leq \dots \leq d_{ik}$ measured from p_i to its first k nearest neighbors in the full set S. The rows of $D(S;k)$ are* lexicographically *ordered as follows. A row $(d_{i1}, \dots, d_{ik})$ is* smaller *than $(d_{j1}, \dots, d_{jk})$ if a few first distances coincide: $d_{i1} = d_{j1}, \dots, d_{il} = d_{jl}$ for $l \in \{1, \dots, k-1\}$ and the next $(l+1)$-st distances satisfy $d_{i,l+1} < d_{j,l+1}$. If w rows are identical to each other, any such group is collapsed to one row with the* weight w/m*. For each row, put this weight in the first column. The final $m \times (k+1)$-matrix is the* Pointwise Distance Distribution $\mathrm{PDD}(S;k)$. ■

The matrix $D(T;3)$ in Table 1 has two pairs of identical rows, so the matrix $\mathrm{PDD}(T;3)$ consists of two rows of weight $\frac{1}{2}$ below. The matrix $D(K;3)$ in Table 1 has only one pair of identical rows, so $\mathrm{PDD}(K;3)$ has three rows of weights $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{4}$. Then $\mathrm{PDD}(T;3) \neq \mathrm{PDD}(K;3)$

Table 1: Each point in $T, K \subset \mathbb{R}^2$ from Figure 3 has ordered distances to three other points.

T points	neighb.1	neighb.2	neighb.3
$(-2,0)$	$\sqrt{2}$	$\sqrt{10}$	4
$(+2,0)$	$\sqrt{2}$	$\sqrt{10}$	4
$(-1,1)$	$\sqrt{2}$	2	$\sqrt{10}$
$(+1,1)$	$\sqrt{2}$	2	$\sqrt{10}$

K points	neighb.1	neighb.2	neighb.3
$(-2,0)$	$\sqrt{2}$	$\sqrt{2}$	4
$(+2,0)$	$\sqrt{10}$	$\sqrt{10}$	4
$(-1,-1)$	$\sqrt{2}$	2	$\sqrt{10}$
$(-1,+1)$	$\sqrt{2}$	2	$\sqrt{10}$

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

Theorem 3.2 (isometry invariance of $\mathrm{PDD}(S;k)$)**.** *For any finite or periodic set $S \subset \mathbb{R}^n$, $\mathrm{PDD}(S;k)$ from Definition 3.1 is an isometry invariant of the set S for any $k \geq 1$.* ■

Table 2: Distances from each motif point of $S(r)$ and $Q(r)$ to their closest neighbors in Fig. 3.

$S(r)$ points	distance to neighbor 1	distance to neighbor 2	distance to neighbor 3
$p_1 = 0$	$\lvert 0 - r\rvert = r$	$\lvert 0 - (2+r)\rvert = 2 + r$	$\lvert 0 - 4\rvert = 4$
$p_2 = r$	$\lvert r - 0\rvert = r$	$\lvert r - (2+r)\rvert = 2$	$\lvert r - 4\rvert = 4 - r$
$p_3 = 2 + r$	$\lvert (2+r) - 4\rvert = 2 - r$	$\lvert (2+r) - r\rvert = 2$	$\lvert (2+r) - 0\rvert = 2 + r$
$p_4 = 4$	$\lvert 4 - (2+r)\rvert = 2 - r$	$\lvert 4 - r\rvert = 4 - r$	$\lvert 4 - 0\rvert = 4$
$Q(r)$ points	distance to neighbor 1	distance to neighbor 2	distance to neighbor 3
$p_1 = 0$	$\lvert 0 - (2+r)\rvert = 2 + r$	$\lvert 0 - (r+4-8)\rvert = 4 - r$	$\lvert 0 - 4\rvert = 4$
$p_2 = 2 + r$	$\lvert (2+r) - 4\rvert = 2 - r$	$\lvert (2+r) - (4+r)\rvert = 2$	$\lvert (2+r) - 0\rvert = 2 + r$
$p_3 = 4$	$\lvert 4 - (4+r)\rvert = r$	$\lvert 4 - (2+r)\rvert = 2 - r$	$\lvert 4 - 0\rvert = 4$
$p_4 = 4 + r$	$\lvert (4+r) - 4\rvert = r$	$\lvert (4+r) - (2+r)\rvert = 2$	$\lvert (4+r) - 8\rvert = 4 - r$

For the 1D periodic sets $S(r) = \{0, r, 2+r, 4\} + 8\mathbb{Z}$ and $Q(r) = \{0, 2+r, 4, 4+r\} + 8\mathbb{Z}$ in Fig. 3, Table 2 shows that $S(r), Q(r)$ are not isometric for any parameter $0 < r \leq 1$.

$$\mathrm{PDD}(S(r);8) = \left(\begin{array}{c|cccccccc} 1/4 & r & 2+r & 4 & 4 & 6-r & 8-r & 8 & 8 \\ 1/4 & r & 2 & 4-r & 4+r & 6 & 8-r & 8 & 8 \\ 1/4 & 2-r & 2 & 2+r & 6-r & 6 & 6+r & 8 & 8 \\ 1/4 & 2-r & 4-r & 4 & 4 & 4+r & 6+r & 8 & 8 \end{array}\right) \neq$$

$$\mathrm{PDD}(Q(r);8) = \left(\begin{array}{c|cccccccc} 1/4 & r & 2-r & 4 & 4 & 6+r & 8-r & 8 & 8 \\ 1/4 & r & 2 & 4-r & 4+r & 6 & 8-r & 8 & 8 \\ 1/4 & 2-r & 2 & 2+r & 6-r & 6 & 6+r & 8 & 8 \\ 1/4 & 2+r & 4-r & 4 & 4 & 4+r & 6-r & 8 & 8 \end{array}\right).$$

The averages of columns in $\mathrm{PDD}(S;k)$ form the vector $\mathrm{AMD}(S;k)$ of *Average Minimum Distances* [34]. Any lattice $\Lambda \subset \mathbb{R}^n$ has a 1-point motif $M = \{p\}$, hence $\mathrm{PDD}(S;k)$ is a single row of increasing distances from p to all other points $\Lambda - \{p\}$. Fig. 6 (right) shows a honeycomb set S whose motif consists of two points that have the same distances to all their neighbors, hence two rows of $D(S;k)$ collapse to a single vector. So $\mathrm{PDD}(S;k) = \mathrm{AMD}(S;k)$ in the cases above.

Example 3.3 (PDD is stronger than AMD)**.** *Since both sets $S(r), Q(r)$ in Fig. 3 (right) have period 8, the matrices $\mathrm{PDD}(S(r);k)$ and $\mathrm{PDD}(Q(r);k)$ have distance 8 in each row for columns 7 and 8 as shown above. All further distances are obtained from the first eight by adding a multiple of period 8. The vector $\mathrm{AMD}(S(r);k)$ of column averages for any $k \geq 8$ is determined by $\mathrm{AMD}(S(r);8) = (1, 2.5, 3.5, 4.5, 5.5, 7, 8, 8)$. Since the components of $\mathrm{AMD}(S(r);k)$ do not depend on the parameter $0 < r < 1$, the sets $S(r)$ are counter-examples to the completeness of* AMD*, now distinguished by $\mathrm{PDD}(S(r);k)$ already for $k = 1$. Hence $\mathrm{PDD}(S;k)$ is strictly stronger than $\mathrm{AMD}(S;k)$.* ■

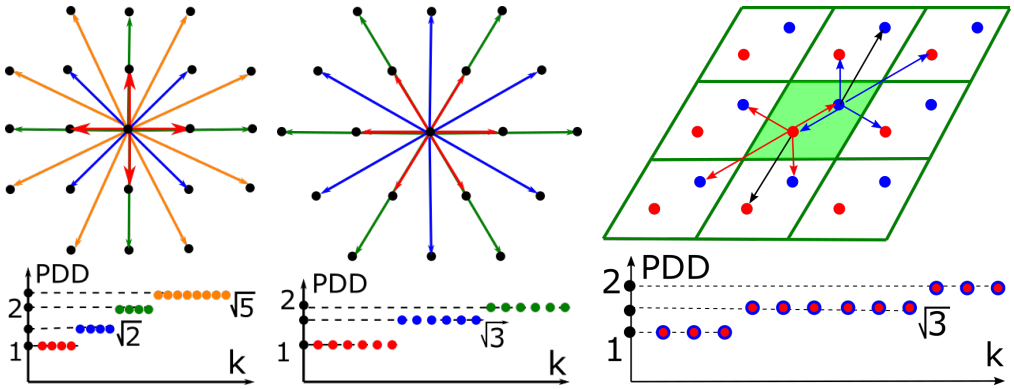

Figure 6: The square lattice (left), hexagonal lattice (middle), and honeycomb periodic set (right) with a minimum inter-point distance of 1 have $\mathrm{PDD}(S;k)$ with a single row of increasing distances.

For a periodic set S, the number k in $\mathrm{PDD}(S;k)$ can be considered as a degree of approximation (or a count of decimal places), not as a parameter that affects invariant values. If we increase k, we extract more distant geometric data from S by adding more columns to $\mathrm{PDD}(S;k)$ and keeping all previous distances. If some rows are identical in $D(S;k-1)$ and become different in $D(S;k)$, we recompute weights but not distances. The past tools [16], [6] strongly depend on extra parameters.

Now we compare PDD with the closest past invariant called the Pair Distribution Function (PDF). For a periodic point set $S \subset \mathbb{R}^n$ with a motif M, the *exact* PDF consists of ordered distances from all points $p \in M$ to all other points $q \in S - \{p\}$. So the infinite sequence $\mathrm{ePDF}(S)$ is obtained from PDD by combining all rows into one sequence and losing weights. Additionally, we keep only one distance from each pair $|p-q| = |q-p|$. For any fixed $0 < r \leq 1$, the sets $S(r), Q(r)$ have the same sequences starting with $\mathrm{ePDF}(S(r)) = \mathrm{ePDF}(Q(r)) = \{r, 2-r, 2, 2+r, 4-r, 4, 4, 4+r, \dots\}$. This example shows that PDD for $k = 1$ is strictly stronger than ePDF as an isometry invariant.

For any lattice $\Lambda \subset \mathbb{R}^n$, the vector $\mathrm{ePDF}(\Lambda;k)$ up to k distances concides with $\mathrm{PDD}(S;k)$. For the honeycomb set S in Fig. 6 (right), $\mathrm{ePDF}(S;2k)$ is obtained from $\mathrm{PDD}(S;k)$ by repeating every distance twice. If a periodic set is perturbed and a unit cell doubles as in Fig. 2 (right), then every distance in ePDF is replaced by a couple of (near-) duplicate distances, so $\mathrm{ePDF}(S;k)$ discontinuously changes by including twice as many short distances and losing longer distances.

This typical discontinuity was roughly repaired by replacing every single distance d with its Gaussian distribution $\exp(-(x-d)^2/2\sigma)$ with a parameter $\sigma > 0$. Then a normalized sum of such 'blurred' distances [69] becomes the smooth Pair Distribution Function $\mathrm{PDF}(S;\sigma)$. Since algorithms can compare only finite vectors, this $\mathrm{PDF}(S;\sigma)$ is then uniformly sampled, which creates dependence on σ. So PDD provides a straightforward alternative to this counter-intuitive PDF pipeline {discrete sequence} $\to$ {smooth function} $\to$ {discrete sequence}, whose continuity wasn't formally proved.

4 Continuity and generic completeness of Pointwise Distance Distributions

Continuity of $\mathrm{PDD}(S;k)$ under perturbations of S in the bottleneck distance d_B will be measured by the Earth Mover's Distance [63], which can be applied to any weighted distributions of different sizes. Definition 4.1 is for any vector $I(S) = ([w_1(S), R_1(S)], \dots, [w_{m(S)}, R_{m(S)}(S)])$ of pointwise invariants of a set S with weights $w_i(S) \in (0,1]$ satisfying $\sum\limits_{i=1}^{m(S)} w_i(S) = 1$.

Later we consider only the case when $[w_i, R_i]$ is the i-th row of $\mathrm{PDD}(S;k)$. Then $m(S)$ is the number of rows in $\mathrm{PDD}(S;k)$. Each row $R_i(S)$ should have a size independent of S, for example a number k of neighbors in $\mathrm{PDD}(S;k)$. For any vectors $R_i = (r_{i1}, \dots, r_{ik})$ and $R_j = (r_{j1}, \dots, r_{jk})$ of a length k, we use the L_∞-distance $|R_i - R_j|_\infty = \max\limits_{l=1,\dots,k} |r_{il} - r_{jl}|_\infty$.

Definition 4.1 (EMD). *Let finite or periodic sets $S, Q \subset \mathbb{R}^n$ have weighted vectors $I(S), I(Q)$ as discussed above. A* flow *from $I(S)$ to $I(Q)$ is an $m(S) \times m(Q)$ matrix whose element $f_{ij} \in [0,1]$ represents a partial* flow *from $R_i(S)$ to $R_j(Q)$. The* Earth Mover's Distance *is the minimum* cost $\mathrm{EMD}(I(S), I(Q)) = \sum\limits_{i=1}^{m(S)} \sum\limits_{j=1}^{m(Q)} f_{ij} |R_i(S) - R_j(Q)|$ *for $f_{ij} \in [0,1]$ subject to* $\sum\limits_{j=1}^{m(Q)} f_{ij} \le w_i(S)$ *for $i = 1, \dots, m(S)$,* $\sum\limits_{i=1}^{m(S)} f_{ij} \le w_j(Q)$ *for $j = 1, \dots, m(Q)$,* $\sum\limits_{i=1}^{m(S)} \sum\limits_{j=1}^{m(Q)} f_{ij} = 1$. ■

The first condition $\sum\limits_{j=1}^{m(Q)} f_{ij} \le w_i(S)$ means that not more than the weight $w_i(S)$ of the component $R_i(S)$ 'flows' into all components $R_j(Q)$ via 'flows' f_{ij}, $j = 1, \dots, m(Q)$. Similarly, the second condition $\sum\limits_{i=1}^{m(S)} f_{ij} = w_j(Q)$ means that all 'flows' f_{ij} from $R_i(S)$ for $i = 1, \dots, m(S)$ 'flow' into $R_j(Q)$ up to the maximum weight $w_j(Q)$. The last condition $\sum\limits_{i=1}^{m(S)} \sum\limits_{j=1}^{m(Q)} f_{ij} = 1$ forces to 'flow' all rows $R_i(S)$ to all rows $R_j(Q)$. The EMD satisfies all metric axioms [63, appendix], needs $O(m^3 \log m)$ time for distributions of a maximum size m and is approximated in $O(m)$ time [66, 39].

Theorem 4.2 (lower bound of EMD). *For finite or periodic point sets $S, Q \subset \mathbb{R}^n$, and $k \ge 1$, the distances satisfy* $\mathrm{EMD}(\mathrm{PDD}(S;k), \mathrm{PDD}(Q;k)) \ge ||\mathrm{AMD}(S;k) - \mathrm{AMD}(Q;k))||_\infty$. ■

Theorem 4.3 uses the bottleneck distance $d_B(S,Q) = \inf\limits_{g: S \to Q} \sup\limits_{p \in S} |p - g(p)|$ and the *packing radius* $r(S)$, which is the minimum half-distance between any points of S. Equivalently, $r(S)$ is the maximum radius r to have disjoint open balls of radius r centered at all points of S.

Theorem 4.3 (continuity of PDD). *For any $k \ge 1$, if finite or periodic sets $S, Q \subset \mathbb{R}^n$ satisfy $d_B(S,Q) < r(S)$, then* $\mathrm{EMD}(\mathrm{PDD}(S;k), \mathrm{PDD}(Q;k)) \le 2 d_B(S,Q)$. ■

Continuity Theorem 4.3 means that any small perturbation of atomic positions in the bottleneck distance d_B leads to a small change of the Pointwise Distance Distribution in the Earth Mover's Distance. Theorem 4.3 extends the following fact for 2-point sets ($k = 1$). If we perturb two points by at most ε, the distance between them changes by at most 2ε.

For any set $S \subset \mathbb{R}^n$ of m points with distinct inter-point distances, completeness of $\mathrm{PDD}(S; m-1)$ follows from [34, Theorem 16]. Following the earlier work [24, section 5.1], appendix C will define a distance-generic set that can approximate any periodic point set $S = \Lambda + M \subset \mathbb{R}^n$. The number m of points in a unit cell U is an isometry invariant because any isometry maps U to another cell with the same number m of points. In dimensions $n = 2, 3$, a lattice Λ can be reconstructed from its isometry invariants in [18, 50]. Theorem 4.4 assumes that a lattice Λ is given and reconstructs a periodic point set $S = \Lambda + M$ in any dimension $n \ge 2$.

Theorem 4.4 (generic completeness of PDD). *Let $S = \Lambda + M \subset \mathbb{R}^n$ be a distance-generic periodic set with m points in a motif M. Let $R(\Lambda)$ be the smallest radius such that all closed balls with centers $p \in \Lambda$ cover $\mathbb{R}^n$. Let $2R(\Lambda)$ be smaller than all distances in the last column of* $\mathrm{PDD}(S;k)$ *for a big enough k. The set S is uniquely reconstructed up to isometry from Λ, m,* $\mathrm{PDD}(S;k)$. ■

5 Polynomial time algorithms and experimental comparisons of PDD

The algorithm for PDD in Theorem 5.1, found several pairs of unexpected duplicates, which were missed by all past tools, through 200B+ pairwise comparisons of 660K+ real periodic crystals over a couple of days on AMD Ryzen 5 5600X (6-core) @4.60Ghz, 32GB DDR4 RAM @3600 Mhz.

The key parameters of $\mathrm{PDD}(S;k)$ is the number m of points in a unit cell U and the number k of neighbors. So the complexity in Theorem 5.1 is near-linear in both k, m for a fixed dimension n. Inputs of the algorithm are a periodic point set $S \subset \mathbb{R}^n$ and an integer $k > 0$. The output $\mathrm{PDD}(S;k)$ is a matrix with at most m rows and exactly $k+1$ columns, where m is the number of motif points. The first column contains the weights of rows, which sum to 1 and are proportional to the number of appearances of the row before collapsing, see the detailed code in appendix B.

Theorem 5.1 (PDD complexity)**.** *Let a periodic set $S \subset \mathbb{R}^n$ have m points in a unit cell U. For a fixed dimension n, $\mathrm{PDD}(S;k)$ is computed in a near-linear time $O(km(5\nu)^n V_n \log(m)\log^2(k))$, where V_n is the unit ball volume in $\mathbb{R}^n$, d and $\nu = \frac{d}{\sqrt[n]{\mathrm{Vol}[U]}}$ are the diameter and* skewness *of U.* ■

Section 2 reviewed that all past tools are based on ambiguous non-invariant data or discontinuous invariants that miss (near-)duplicates, or the resulting algorithms are too slow for pairwise comparisons of millions of crystal structures. The recently discovered continuous invariants with theoretical (not exactly computable) metrics [24, section 6] and [2, section 8] require cubic algorithms, which turned out to be unrealistic for large data. We tried our best and ran several algorithms below.

The Cambridge Crystallographic Data Centre (CCDC) is a multi-million company curating the Cambridge Structural Database (CSD) since 1960s. Now the CSD is the largest in the world with 1.17M+ known structures. A new crystalline material is deposited in the CSD only after a peer-reviewed publication. The CCDC checks that a new structure is genuine and not a duplicate of an earlier one because their data is trusted by all pharmaceutical giants developing new drugs in a crystalline form. The CSD is a huge list of Crystallographic Information Files representing crystals by unit cells and motifs of points in coordinates of a cell basis with limited search and slow comparison.

The Nature paper [32] reported four experimental T2 crystals (based on the same molecule T2) that were successfully synthesized after predicting 5679 crystals through 12-week simulations on a supercomputer. All initial 2M+ randomly sampled crystals were iteratively optimized to the 'most stable' approximations of local energy minima. This is a typical 'embarrassment of over-prediction' when many (near-)duplicates are found around the same local minimum but remain undetected.

One striking example is the pair of crystals 14 and 15 in Fig. 7, see more details of the available data in appendix A. When this pair was compared by another free software Platon [68], a bug was discovered, which is still not fixed for a couple of months. Such bugs will keep emerging because the discontinuity of past invariants and metrics was not addressed as in Problem 1.1.

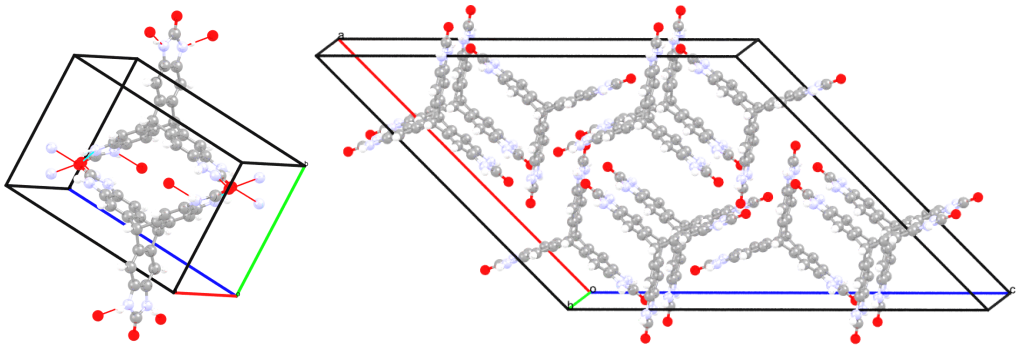

Figure 7: Crystals 14, 15 based on the T2 molecule have very different Crystallographic Information Files (with different motifs in unit cells of distinct shapes) but are nearly identical up to isometry.

For example, a rough sampling of the density functions $\psi_k(t)$ from [24] of 5679 crystals for up to $k = 8$ took more than four days on a comparable machine. This experiment detected the T2-δ crystal that was accidentally not deposited in the CSD because of a visual confusion with another structure.

The most popular packing similarity [16] algorithm COMPACK is available in the free software Mercury. The 4950 comparisons of the 100 lowest energy crystals close to T2-δ in density by packing similarity took 3 hours 53 min, 2.825 seconds per comparison. Extrapolating this time for comparing any new structure with the whole CSD gives 38 days. In contrast, a typical comparison by PDD takes around 10 milliseconds, so comparing 100 crystals pairwise takes less than one minute.

Table 3: Most comparisons of 100 lowest energy crystals close to the T2-δ by packing similarity [16] matched small numbers of molecules for the default maximum 15, which means a failure.

molecules	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
comparisons	2784	1150	773	69	85	21	31	6	2	7	12	3	1	0	6

Most importantly, more than half of all comparisons in Table 3 matched only one molecule from two crystals. Since all crystals consist of the same rigid molecule T2, this output means a complete failure: one common molecule, no other conclusions. Since the CCDC deposits hundreds of new structures daily, the short-cut approach is to compare the chemical composition of atoms. But even the water H_2O has at least 15 forms of ice crystals, while other compositions have many more polymorphic forms in the CSD. This comparison by composition can miss duplicates where one atom is incorrect.

Exactly this reliance on past tools allowed PDD to detect five pairs of unexpected duplicate 'needles in the haystack' of 660K+ periodic crystals. First, the simpler invariants $\mathrm{AMD}(S;100)$ were computed for all 660K+ periodic structures in the CSD, without disorder and with full geometric data.

The 200B+ pairwise comparisons of $\mathrm{AMD}(S;100)$ vectors revealed 6371 pairs S,Q with $|\mathrm{AMD}(S;100)-\mathrm{AMD}(Q;100)|_\infty \le 0.01$. As an AMD is simpler and faster to compare, up to the order of 10^{-7} seconds per comparison, this step took around 8 hours. This fact and Theorem 4.2 makes AMD a good filter for comparison before using the stronger invariants PDD.

Second, computing the L_∞-based EMD between the pairs above detected 182 pairs with $\mathrm{EMD} < 0.01$. Most of these pairs were expected and were the same crystal, or different aliases for the same database entry. The five pairs reported in [34, section 7] were unexpected because the underlying periodic sets of points at atomic centres were truly isometric (to the last decimal place) but one atom had different chemical elements in two crystals. The crystals with the CSD codes HIFCAB and JEPLIA are literally isometric but one Cadmium is replaced by Manganese at the same position. All past tools taking into account atomic types see these crystals as different. The CCDC agreed that such a coincidence is physically impossible because another atom should have slightly different distances to neighbors detected by PDD. Hence at least one of the structures in the pairs above cannot be correct. The five journals have started investigating the data integrity of the underlying publications.

This paper reports many more pairs in supplementary materials that were less obvious due to larger EMD values up to 0.1. The new pairs were found by comparing periodic sets of points at molecular centers instead of atomic centers. The pairs of the resulting sets of centers are exactly identical with $\mathrm{EMD}=0$ but differ by some atomic types as above. The CCDC is now investigating this new batch.

6 Conclusions and a discussion of remaining limitations and societal impact

The most important contribution is the introduction of the PDD invariants whose speed, continuity, and generic completeness in Problem 1.1 have no analogs among any state-of-the-art comparisons.

Many descriptors and similarities of crystals use only their finite subsets (only molecules or balls of a manually cut-off radius), which can have infinitely many non-isometric classes as shown in Fig. 4 (left). Any minimally useful descriptor should be invariant as in Problem 1.1a because any rigid motion cannot really change a solid crystalline material. A descriptor with false negatives cannot be used for comparison because equivalent objects can have different values of a non-invariant.

False negatives should be avoided first to detect different crystals $S \ncong Q$ due to $I(S) \ne I(S)$. Only for an invariant I, it makes sense to avoid false positives by proving the completeness of I in condition (1.1b) to justify the converse implication: $I(S)=I(Q) \Rightarrow S \cong Q$. All 'state-of-the-art' comparisons fail to distinguish many periodic crystals [41, 60, 57], hence cannot define a metric

satisfying the coincidence axiom: $d(S,Q)=0$ only for isometric crystals $S \cong Q$. Appendix A gives specific examples when past similarities RMSD [16] and 1-PXRD [65] fail the triangle axiom.

The traditional crystallography tools such as diffraction patterns and pair distribution function cannot distinguish any homometric crystals [58], even in dimension 1 as shown in Fig. 3 (right).

The simpler continuity in conditions (1.1c,d) is needed to quantify the similarity of (near-)duplicates but is failed by most invariants based on (reduced) cells or symmetry groups, see Fig. 2 (right).

Only the couple of recently discovered invariants satisfy the practically important conditions (1.1a,b,c,d,f): isosets [4] and PDD in this paper, see Fig. 5. The advantage of PDD is their near-linear time in Theorem 5.1. Pairwise comparisons of millions of real crystals are urgently needed to avoid the paper mills based on duplicates [9], whose new cases are reported in appendix A.

So all 'state-of-the-art' tools were too slow for large experiments such as 200B+ pairwise comparisons of all 660K+ periodic crystals in the Cambridge Structural Database (CSD). Hence it is possible to compare PDD with the past tools only on smaller datasets. The appendices below include

Appendix A: details of extra experimental comparisons with the state-of-the-art tools;

Appendix B: a pseudo-code with examples and instructions for the attached code;

Appendix C: rigorous proofs of all theoretical results in the main paper.

In conclusion, Theorems 3.2, 4.3, 5.1 , 4.4 fulfilled almost all conditions of Problem 1.1, while all past tools remained discontinuous or too slow for billions of real comparisons. The only limitation is a hypothetical existence of singular sets $S \not\cong Q$ with $\mathrm{PDD}(S;k)=\mathrm{PDD}(Q;k)$ for all $k \geq 1$. The PDD distinguished all known 660K+ periodic crystals in the CSD through 200B+ comparisons each running in nanoseconds on a modest desktop outperforming all tools by many orders of magnitude.

As a result, several pairs of potentially fraudulent structures are emerging, which might have some negative impact on past publications that could be retracted. More importantly, the experiments confirmed the Crystal Isometry Principle [34, section 7]: the map $\{\text{periodic crystals}\} \to \{\text{periodic point sets}\}$ is injective (doesn't lose information) modulo isometry. Hence all existing and undiscovered crystals live in the common space parameterized by complete isometry invariants. Its first continuous maps for 2D lattices appeared in [13]. We thank all reviewers for their valuable time and suggestions.

References

[1] LC Andrews, HJ Bernstein, and GA Pelletier. A perturbation stable cell comparison technique. *Acta Crystallographica Section A*, 36(2):248–252, 1980.

[2] O Anosova and V Kurlin. Introduction to periodic geometry and topology. *arXiv:2103.02749 (draft book, to be updated in July 2022)*.

[3] O Anosova and V Kurlin. Algorithms for continuous metrics on periodic crystals. *arxiv:2205.15298 (early draft)*, 2021.

[4] O Anosova and V Kurlin. An isometry classification of periodic point sets. In *Proceedings of Discrete Geometry and Mathematical Morphology*, 2021.

[5] Olga Anosova and Vitaliy Kurlin. Density functions of periodic sequences. *arxiv:2205.02226*, 2022.

[6] Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B*, 87(18):184115, 2013.

[7] Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of chemical physics*, 134(7):074106, 2011.

[8] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *Transactions PAMI*, 24(4):509–522, 2002.

[9] David Bimler. Better living through coordination chemistry: A descriptive study of a prolific papermill that combines crystallography and medicine. 2022.

[10] M. Bouniaev and N. Dolbilin. Regular and multi-regular t-bonded systems. *J. Information Processing*, 25:735–740, 2017.

[11] M. Boutin and G. Kemper. On reconstructing n-point configurations from the distribution of distances or areas. *Adv. Appl. Math.*, 32(4):709–735, 2004.

[12] Peter Brass and Christian Knauer. Testing the congruence of d-dimensional point sets. In *Proceedings of SoCG*, pages 310–314, 2000.

[13] Matthew Bright, Andrew I Cooper, and Vitaliy Kurlin. Geographic-style maps for 2-dimensional lattices. *arxiv:2109.10885*, 2021.

[14] Matthew Bright, Andrew I Cooper, and Vitaliy Kurlin. Welcome to a continuous world of 3-dimensional lattices. *arxiv:2109.11538 (early draft)*, 2021.

[15] Mathew J. Bryant, Simon N. Black, Helen Blade, Robert Docherty, Andrew G.P. Maloney, and Stefan C. Taylor. The csd drug subset: The changing chemistry and crystallography of small molecule pharmaceuticals. *Journal of Pharmaceutical Sciences*, 108(5):1655–1662, 2019.

[16] J. Chisholm and S. Motherwell. Compack: a program for identifying crystal structure similarity using distances. *J. Applied Crystal.*, 38:228–231, 2005.

[17] Scott Cohen and Leonidas Guibas. The earth mover's distance: Lower bounds and invariance under translation. Technical report, Stanford University, 1997.

[18] J Conway and N Sloane. Low-dimensional lattices. vi. voronoi reduction of three-dimensional lattices. *Proceedings Royal Society A*, 436(1896):55–68, 1992.

[19] William IF David, Kenneth Shankland, Ch Baerlocher, LB McCusker, et al. *Structure determination from powder diffraction data*, volume 13. 2002.

[20] de Pablo et L. New frontiers for the materials genome initiative. *npj Computational Materials*, 5(1):1–23, 2019.

[21] N Dolbilin and M Bouniaev. Regular t-bonded systems in R^3. *European Journal of Combinatorics*, 80:89–101, 2019.

[22] N. Dolbilin, J Lagarias, and M. Senechal. Multiregular point systems. *Discrete & Computational Geometry*, 20(4):477–498, 1998.

[23] N Dym and S Kovalsky. Linearly converging quasi branch and bound algorithms for global rigid registration. In *Proceedings ICCV*, pages 1628–1636, 2019.

[24] H Edelsbrunner, T Heiss, V Kurlin, P Smith, and M Wintraecken. The density fingerprint of a periodic point set. In *Proceedings of SoCG*, pages 32:1–32:16, 2021.

[25] Herbert Edelsbrunner and Raimund Seidel. Voronoi diagrams and arrangements. *Discrete & Computational Geometry*, 1(1):25–44, 1986.

[26] Y. Elkin and V. Kurlin. The mergegram of a dendrogram and its stability. In *Proceedings of MFCS (Mathematical Foundations of Computer Science)*, 2020.

[27] Y. Elkin and V. Kurlin. Isometry invariant shape recognition of projectively perturbed point clouds by the mergegram extending 0d persistence. *Mathematics*, 9(17), 2021.

[28] Yury Elkin. New compressed cover tree for k-nearest neighbor search (phd thesis). *arxiv:2205.10194*, 2022.

[29] Yury Elkin and Vitaliy Kurlin. Counterexamples expose gaps in the proof of time complexity for cover trees introduced in 2006. In *Topological Data Analysis and Visualization (TopoInVis)*, pages 9–17, 2022.

[30] Yury Elkin and Vitaliy Kurlin. Paired compressed cover trees guarantee a near linear parametrized complexity for all k-nearest neighbors search in an arbitrary metric space. *arxiv:2201.06553*, 2022.

[31] Yury Elkin and Vitaliy Kurlin. A new near-linear time algorithm for k-nearest neighbor search using a compressed cover tree. In *International Conference on Machine Learning (ICML)*, volume 202, pages 9267–9311, 2023.

[32] A Pulido et al. Functional materials discovery using energy–structure maps. *Nature*, 543:657–664, 2017.

[33] C Groom et al. The Cambridge Structural Database. *Acta Cryst B*, 72(2):171–179, 2016.

[34] D Widdowson et al. Average minimum distances of periodic point sets. *MATCH Communications in Math. Comp. Chemistry*, 87:529–559, 2022.

[35] H Maron et al. Point registration via efficient convex relaxation. *ACM Trans. on Graphics*, 35(4):1–12, 2016.

[36] H Pottmann et al. Integral invariants for robust geometry processing. *Comp. Aided Geom. Design*, 26(1):37–60, 2009.

[37] Hans-Georg Carstens et al. Geometrical bijections in discrete lattices. *Combinatorics, Probability and Computing*, 8:109–129, 1999.

[38] J Yang et al. Go-icp: A globally optimal solution to 3d icp point-set registration. *Transactions PAMI*, 38:2241–2254, 2015.

[39] R Sato et al. Fast and robust comparison of probability measures in heterogeneous spaces. *arXiv:2002.01615*, 2020.

[40] S Manay et al. Integral invariants for shape matching. *Transactions PAMI*, 28(10):1602–1618, 2006.

[41] S Pozdnyakov et al. Incompleteness of atomic structure representations. *Phys. Rev. Lett.*, 125:166001, 2020.

[42] Giuseppe Fadda and Giovanni Zanzotto. On the arithmetic classification of crystal structures. *Acta Cryst. A: Foundations*, 57(5):492–506, 2001.

[43] G Flor, D. Orobengoa, E. Tasci, J. Perez-Mato, and M. Aroyo. Comparison of structures applying the tools available at the bilbao crystallographic server. *J. Applied Crystallography*, 49(2):653–664, 2016.

[44] Cosmin Grigorescu and Nicolai Petkov. Distance sets for shape filters and shape recognition. *IEEE transactions on image processing*, 12(10):1274–1286, 2003.

[45] Grünbaum and Moore. The use of higher-order invariants in the determination of generalized Patterson cyclotomic sets. *Acta Cryst A*, 51:310–323, 1995.

[46] T Hahn, U Shmueli, and J Arthur. *Internat. tables for crystallography*, volume 1. 1983.

[47] Maria Angeles Junquera Pulido, Linjiang Chen, Tomasz Kaczorowski, Daniel Holden, Marc A Little, Samantha Y Chong, Benjamin J Slater, David Mcmahon, Baltasar Bonillo, Chloe J Stackhouse, et al. Additional computational data (related to" functional materials discovery using energy–structure–function maps" manuscript). 2017.

[48] Heuna Kim and Günter Rote. Congruence testing of point sets in 4 dimensions. *arXiv:1603.07269*, 2016.

[49] V Kurlin. A complete isometry classification of 3-dimensional lattices. *arxiv:2201.10543*, 2022.

[50] V Kurlin. Mathematics of 2-dimensional lattices. *arxiv:2201.05150*, 2022.

[51] Vitaliy Kurlin. A computable and continuous metric on isometry classes of high-dimensional periodic sequences. *arxiv:2205.04388*, 2022.

[52] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426*, 2018.

[53] Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of Comp. Mathematics*, 11(4):417–487, 2011.

[54] M Mosca and V Kurlin. Voronoi-based similarity distances between arbitrary crystal lattices. *Crystal Research and Technology*, 55(5):1900197, 2020.

[55] Robert Osada, Thomas Funkhouser, Bernard Chazelle, and David Dobkin. Shape distributions. *ACM Transactions on Graphics (TOG)*, 21(4):807–832, 2002.

[56] Steve Oudot and Elchanan Solomon. Inverse problems in topological persistence. In *Topological Data Analysis*, pages 405–433. Springer, 2020.

[57] Behnam Parsaeifard and Stefan Goedecker. Manifolds of quasi-constant soap and acsf fingerprints and the resulting failure to machine learn four-body interactions. *The Journal of Chemical Physics*, 156(3):034302, 2022.

[58] A Patterson. Homometric structures. *Nature*, 143:939–940, 1939.

[59] A Patterson. Ambiguities in the x-ray analysis of structures. *Phys. Rev.*, 65:195–201, 1944.

[60] Sergey N Pozdnyakov and Michele Ceriotti. Incompleteness of graph convolutional neural networks for points clouds in three dimensions. *arXiv:2201.07136*, 2022.

[61] Daniel Probst and Jean-Louis Reymond. Visualization of very large high-dimensional data sets as minimum spanning trees. *Cheminformatics*, 12:1–13, 2020.

[62] Joseph Rosenblatt and Paul D Seymour. The structure of homometric sets. *SIAM Journal on Algebraic Discrete Methods*, 3(3):343–350, 1982.

[63] Y. Rubner, C. Tomasi, and L. Guibas. The earth mover's distance as a metric for image retrieval. *Intern. Journal of Computer Vision*, 40(2):99–121, 2000.

[64] Mauro R Ruggeri and Dietmar Saupe. Isometry-invariant matching of point set surfaces. In *3DOR*, pages 17–24. Citeseer, 2008.

[65] Pietro Sacchi, Matteo Lusi, Aurora J Cruz-Cabeza, Elisa Nauha, and Joel Bernstein. Same or different–that is the question: identification of crystal forms from crystal structure data. *CrystEngComm*, 22(43):7170–7185, 2020.

[66] S Shirdhonkar and D Jacobs. Approximate earth mover's distance in linear time. In *Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[67] P Smith and V Kurlin. Families of point sets with identical 1d persistence,. *arxiv:2202.00577 (early draft)*, 2022.

[68] ALJ Spek. Single-crystal structure validation with the program platon. *Journal of applied crystallography*, 36(1):7–13, 2003.

[69] Maxwell W Terban and Simon JL Billinge. Structural analysis of molecular materials using the pair distribution function. *Chemical Reviews*, 122:1208–1272, 2022.

[70] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[71] Peter Zwart, Ralf Grosse-Kunstleve, Andrey Lebedev, Garib Murshudov, and Paul Adams. Surprises and pitfalls arising from (pseudo) symmetry. *Acta Cryst. D*, 64:99–107, 2008.

A Appendix A: details of experimental comparisons with the state-of-the-art

This appendix provides extra details for large-scale experimental comparisons of the new invariant PDD with the most widely used past tools: the Root Mean Square Deviation (RMSD) for comparing finite subsets [16] and 1-PXRD similarity based on powder X-ray diffraction patterns. The popularity of these tools does not help to resolve the key question for crystals : 'the same or different?' [65].

The key advantage of Pointwise Distance Distributions (PDD) for searching (near-)duplicates and nearest neighbors in huge datasets is their hierarchical nature. The weighted averages of columns in the PDD matrix form the simpler AMD vectors (Average Minimum Distances), which can be quickly subtracted and compared. Since the Earth Mover's Distance (EMD) between PDD matrices can be only larger than the L_∞ distance between their AMD vectors by Theorem 4.2, it suffices to compute the slower EMD only on pairs of crystals whose AMD vectors are very close to each other.

Any crystal dataset can be further organized into hierarchical families whose levels are parameterized by the number i of atomic neighbors. The i-th level of this hierarchy can consist of disjoint or overlapping clusters of crystals that have close values of the i-th coordinate AMD_i in the vector $\mathrm{AMD}(S;k) = (\mathrm{AMD}_1, \dots, \mathrm{AMD}_k)$. Indeed, $L_\infty(\mathrm{AMD}(S;k), \mathrm{AMD}(Q;k)) \geq |\mathrm{AMD}_i(S) - \mathrm{AMD}_i(Q)|$. Hence, any crystals with distant values of AMD_i cannot have close AMD vectors.

We use the public dataset of 5679 T2 crystals base on the same T2 molecule and predicted by 12-week supercomputer simulations [47]. Only five were synthesized: T2-α, T2-β, T2-γ, T2-δ, T2-ε. By request we can share CIFs of all mentioned crystals, which can be opened by any text editor and visualised by the free software Mercury. Fig. 8 shows 10 PDD curves of T2-ε corresponding to the 10 isometrically unique atomic positions within the T2 molecule. The rotation through 120° around an axis keeps the T2 molecule and the whole infinite crystal T2-ε invariant, while the three oxygen atoms shown in red simply rotate their positions. Hence all oxygen atoms have the same ordered distances to their neighbors and their PDD curves coincide. The same conclusion holds for six nitrogens whose positions are isometrically equivalent in T2. Carbon atoms have five non-isometric positions, which generate five unique PDD curves in Fig. 8. All 10 PDD curves provide a finer classification of atomic types not only by chemical elements, but by their geometric positions within a molecule. For every atom, the first distance for $k = 1$ is the bond distance to its closest neighbor, see the zoomed beginnings of PDD curves. Though chemical elements can be included into PDD as an extra column, the 10 distinct PDD curves in Fig. 8 distinguish all chemical elements in the T2 molecule.

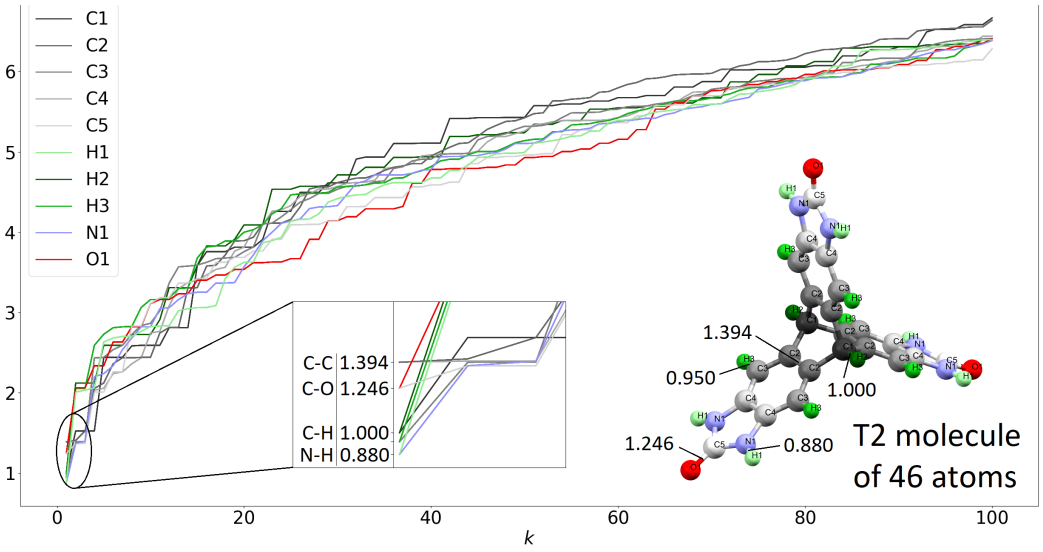

Figure 8: Each curve shows distances d_k from one atom of a T2 molecule to k-th neighbor in the T2-ε crystal. The 46 atoms form 10 groups, one for each atom in the asymmetric unit: three oxygens form one type with the PDD curve in red, 23 carbons have 5 non-isometric positions in T2 with PDD curves in 5 shades of gray. The zoomed part shows the initial bond distances in Å $= 10^{-10}$ m.

A key challenge in visualization is to justifiably represent invariant data in a simple form. Past algorithms such as t-SNE [70] and UMAP [52] are stochastic and can produce different outputs from repeated runs on the same input. Hence we used the more recent deterministic TreeMap [61] to draw a Minimum Spanning Tree (MST). Informally, any MST can be considered as an optimal road network between cities. Such a network can be drawn in many different ways; the most valuable new data is the distances between PDDs of crystals but the MST is drawn only for visualization. Similarly, positions of cities and distances between them are more important than a sketch of a road network.

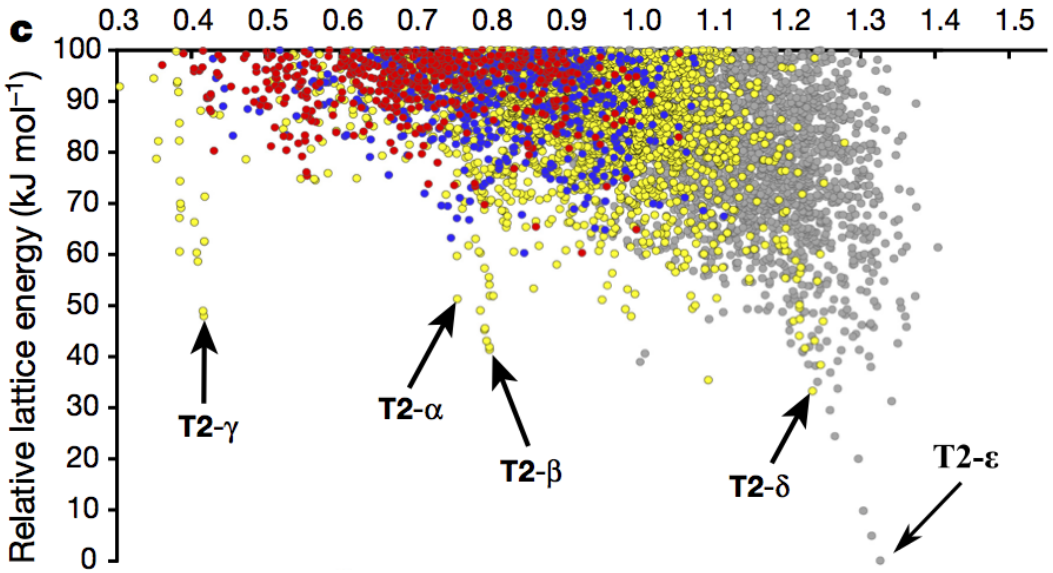

Figure 9: State-of-the-art visualization of a CSP dataset, modified from Fig. 2d in [32]. Every predicted crystal has two coordinates (density, energy). The arrows show the predicted crystals that were manually matched to the five experimental crystals T2-α, T2-β, T2-γ, T2-δ, T2-ε.

Fig. 10 illustrates key advantages of the invariant-based visualization in the right picture compared with the energy landscape in the left picture extracted from the state-of-the-art [32] in Fig. 9. Briefly, the MST based on stronger isometry invariants $\mathrm{PDD}(S;100)$ provides more information about crystal similarity than the cloud of points (density, energy). In the past, simulated versions were manually searched to match experimental crystals only by density. Measuring the energy of an experimental crystal needs an explosion disintegrating this crystal. For the experimental crystal T2-δ, dozens of simulated crystals were searched in the vertical strip in the left picture of Fig. 10. Simulated crystal 14 was visually chosen in [32] as the best match for T2-δ.

The new PDD invariant automates the search for closest crystals, because PDD can be computed for all types of simulated and experimental crystals. The MST in the right picture of Fig. 10 includes T2-δ as a red dot close to the near duplicate crystals 14 and 15, which were not filtered out by past tools because of very different unit cells and motifs in Fig. 7. There were many other pairs of crystals with almost identical values of (density, energy), for example crystals 5920 and 0049, whose structures were found away from each other in the MST. When comparing this pair by the COMPACK algorithm, only one molecule is matched in Table 4. The resulting zero value of RMSD only means that crystals 5920 and 0049 are different (no similarity found), while EMD provides a proper distance. Fig. 10 (right) shows crystals 5920, 0049 in different branches of the MST.

The 4950 comparisons of 100 crystals by COMPACK took 3 hours 53 min (2.825 sec per comparison), which is 3 orders of magnitude slower than EMD on PDD invariants. Table 3 says that over half the time (56%) only one molecule is matched, which means that no similarity was found between crystals based on identical molecules. In fact, in 95% of cases three or fewer molecules are matched, which is also considered a non-match for the default number of 15 maximum matched molecules. The number of matched molecules rarely exceed 10 (0.5%), and only 6 comparisons matched 15/15.

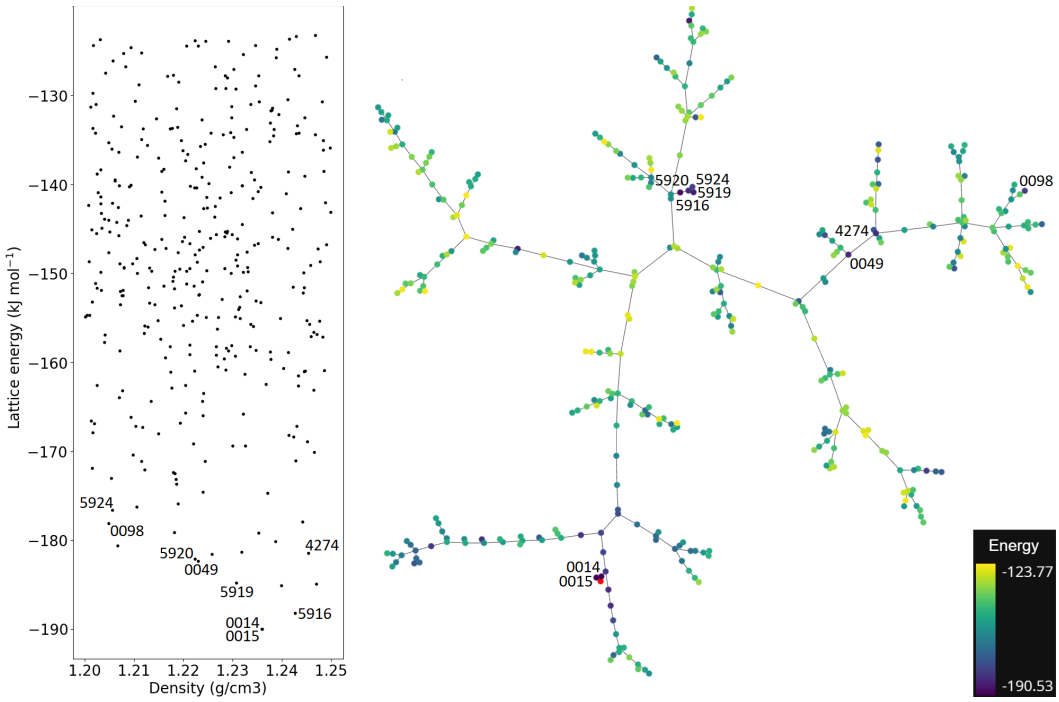

Figure 10: **Left**: a vertical 'strip' of the energy landscape from Fig. 9 in the density range $[1.2, 1.25]$ around the density of the experimental crystal T2-δ. Some of the lowest energy crystals are highlighted by their IDs in the T2 dataset reported in [32]. **Right**: MST of the simulated crystals on the left plus the experimental crystal T2-δ in red, based on $\mathrm{PDD}(S; 100)$. All simulated crystals are colored by lattice energy according to the key in the bottom right corner.

Table 4: The traditional COMPACK algorithm [16] measures similarity by attempting to match up to 15 (by default) molecules from two crystals. The low values 1, 2, 3 (of 15) mean that only a few molecules could be aligned in the two crystals with a small enough Root Mean Square Deviation (RMSD), so this is not considered a match.

	0014	0015	0049	0098	4274	5916	5919	5920	5924
0014	15	15	3	1	1	1	2	2	2
0015	15	15	3	1	1	1	2	2	2
0049	3	3	15	1	10	2	1	1	1
0098	1	1	1	15	1	1	1	1	1
4274	1	1	10	1	15	2	2	2	2
5916	1	1	2	1	2	15	10	10	10
5919	2	2	1	1	2	10	15	15	15
5920	2	2	1	1	2	10	15	15	15
5924	2	2	1	1	2	10	15	15	15

We also compare the new invariant to powder X-ray diffraction patterns (PXRD), a popular way of measuring similarity between structures. We note that unlike Earth mover's distance between PDDs, the 'distance' between PXRDs is not a proper metric on the space of crystals, and so is not appropriate to use it to continuously explore the space. Our experiments show that $1-$PXRD (since a PXRD score of 1 means identical) fails to satisfy the triangle inequality on the T2 dataset over 8% of the time, and the RMSD values given by COMPACK similarly fail to satisfy this axiom, see Table 5.

Our second dataset has all 12576 structures in the Drug Subset [15] of the Cambridge Structural Database (CSD). Though the CSD is the world's largest collection of more than 1.1M crystals, it is effectively a huge folder of Crystallographic Information Files (CIF) with limited search functions, for example by a chemical composition.

Crystal 1	Crystal 2	RMSD
T2-γ	T2-α	2.805
T2-γ	T2-ϵ	0.184
T2-α	T2-ϵ	0.231

Crystal 1	Crystal 2	1−PXRD
T2-0001	T2-0029	0.187
T2-0001	T2-5333	0.110
T2-0029	T2-5333	0.0683

Table 5: Triples of crystals where the commonly used tools RMSD and PXRD fail the traingle inequality axiom required of a proper metric. The CIFs are available by request.

The CSD Drug subset only includes small-molecule crystal structures containing a drug molecule, typically smaller than the T2 molecule. Since larger than necessary values of k don't affect distances to closer neighbors, we tried $k = 100$ and $k = 200$ for the CSD Drug Subset.

After computing $\mathrm{PDD}(S;k)$, we use a standard EMD algorithm for distances between the distributions. A Minimum Spanning Tree can be computed from a smaller graph on PDD invariants not requiring all pairwise distances. Since distances asymptotically approach $\mathrm{PPC}(S)\sqrt[3]{k}$, we used the instantly computable Point Packing Coefficient $\mathrm{PPC}(S)$ to find 3000 neighbors for each of 12576 crystals. For $k = 200$, the computations of 12576×3000 EMD distances (excluding all repetitions) took 7 hours 39 min (1.238 ms per comparison, 2.19 sec per crystal for all 3000 neighbors).

Any MST can be visualized by TreeMap [61] or any other graph drawing algorithm. The continuity and invariance of Pointwise Distance Distributions in Theorems 3.2 and 4.3 guarantee that any crystal data analysis based on these invariants is theoretically justified. Since a Minimum Spanning Tree has no cycles, some close crystals might appear at terminal vertices of distant branches, which could be connected to form a cycle. Drawing a graph with cycles requires more advanced algorithms. However, the new invariants can be used for a quick search of similar crystals without visualization.

We have interactive versions of Minimum Spanning Trees of the T2 dataset and CSD drug subset. In any Javascript-enabled browser one can zoom these TreeMaps, color the vertices by properties, hover over any vertex to get more information. Though aspirin and paracetamol are chemically different, they have similar pharmaceutical properties and their crystal structures (dozens of polymorphs with CSD references ACSALA and HXACAN) were found in close branches.

The 'structure' of a crystal was previously measured by only one continuous isometry invariant: the physical density equal to the weight of atoms within a unit cell, divided by the cell volume. All other descriptors are either non-invariants, e.g. powder diffraction patterns depend on various thresholds and are not always reliable, or are discontinuous under perturbations, e.g. symmetry groups.

The main contribution is replacing the single-value density by much stronger more isometry invariants PDD. Comparing crystals by only their densities was the latest 'state-of-the-art' in [32], where 5 forms of 9 experimental crystals were initially matched to dozens of simulated crystals with close densities. Then hundreds of crystals were visually inspected to select the five matches in Fig. 9.

Now a 'structure' of a crystal can be much better represented by many more numerical invariants in PDD. Fig. 11 is the first justified map of (a discretely sampled) space of T2 crystals, because the underlying distance between PDD invariants satisfies all metric axioms and continuity under perturbations by [63] and Theorem 4.3.

The map in Fig. 11 shows a Minimum Spanning Tree built on 5679+9 crystals. Each crystal is a vertex. The distance between any crystals is the Earth Mover's distance between the first order PDD invariants. The most interesting vertices in the T2 map are experimental crystals shown in red, because their energies are impractical to measure. The map shows experimental crystals with simulated predictions for the first time, because past plots such as in Fig. 9 need all energy values.

Fig. 11 is a substantial advancement in visualizing crystal datasets, because the underlying invariants $\mathrm{PDD}(S;k)$ are much more informative than a single-value density. The nine experimental crystals (red dots) clearly split into five groups marked by T2-α, T2-β, T2-γ, T2-δ, T2-ε. These nine experimental crystals have no known energies and cannot appear in the past energy landscapes. The four red dots at the very top represent slightly different forms of a T2-γ crystal synthesized at different temperatures.

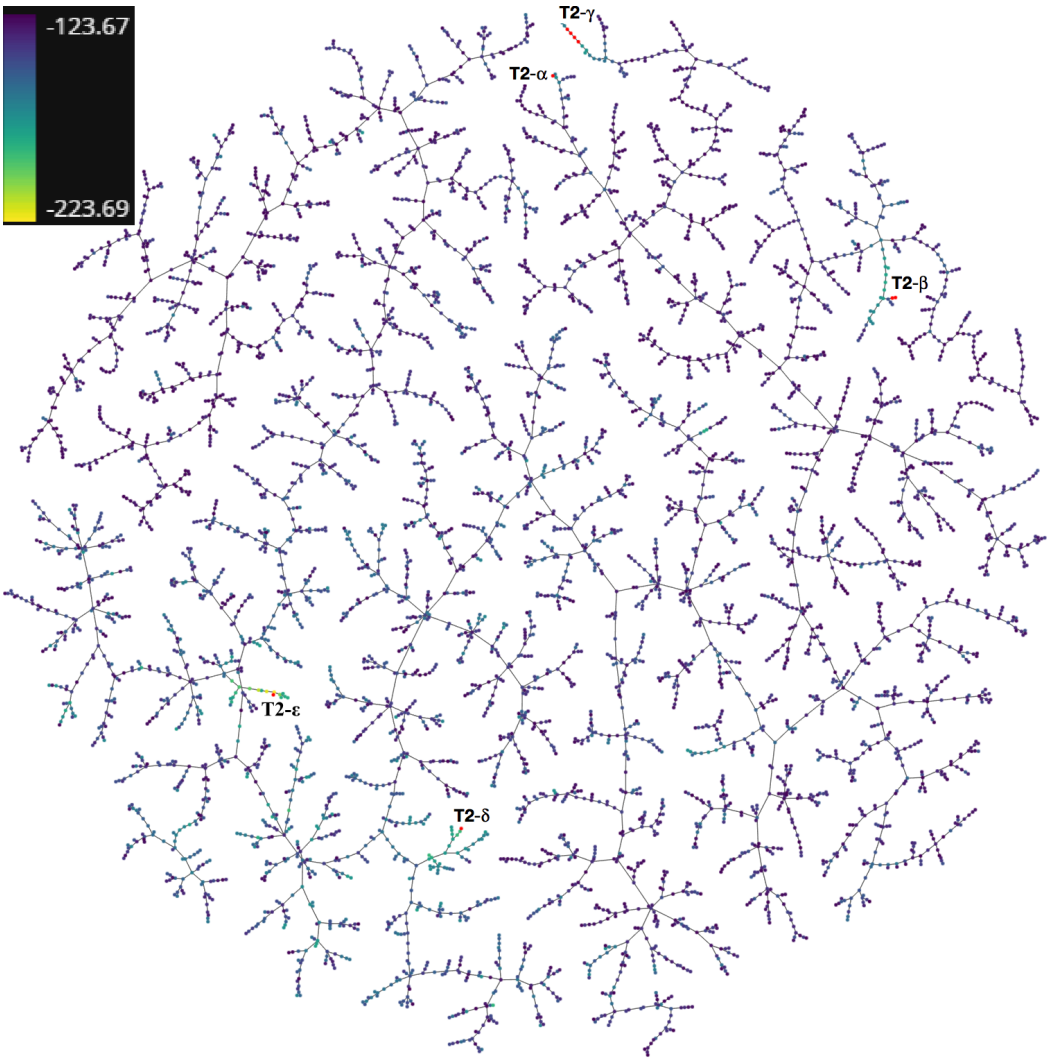

Figure 11: A Minimum Spanning Tree of 9+5679 T2 crystals drawn by TreeMap [61]. The distance between any crystals S, Q is the $\mathrm{EMD}(\mathrm{PDD}(S;125), \mathrm{PDD}(Q;125))$. The color bar in the top left corner shows energy values. The minimum of -223.69 kJ/mol corresponds to 0 in Figure 9. Nine experimental crystals, which form the families T2-α, T2-β, T2-γ, T2-δ, T2-ε, are shown in red, because their energies are not experimentally measured (a crystal is disintegrated only by a blow up).

The map can be interactively explored by opening the file T2L_PDD125_TMap.html (available by request) in any browser in the same folder with the corresponding JavaScript files. In the interactive version one can color vertices by a property, which can be changed in the drop-down menu in the bottom right corner. For example, all crystals with low energies appear in yellow colors.

In Fig. 11 six of nine experimental T2 crystals have the codes DEBXIT01...06, though they split into two groups (unique polymorphs): four nearly identical crystals T2-γ and two nearly identical crystals T2-β. Our final experiment on all 229K molecular organic crystals found thousands of such pairs in the Cambridge Structural Database.

In addition to the five pairs of isometrically identical crystals discussed in section 6, a search on the CSD using PDD on molecular centers instead of atomic centers revealed nine pairs of structures which were almost identical, listed in table 6. The supplementary code includes a script to compare the previously mentioned five pairs by PDD, however the code to compute molecular centers is private, so it was not possible to include a script reproducing the results below. Instead, the .CIFs for

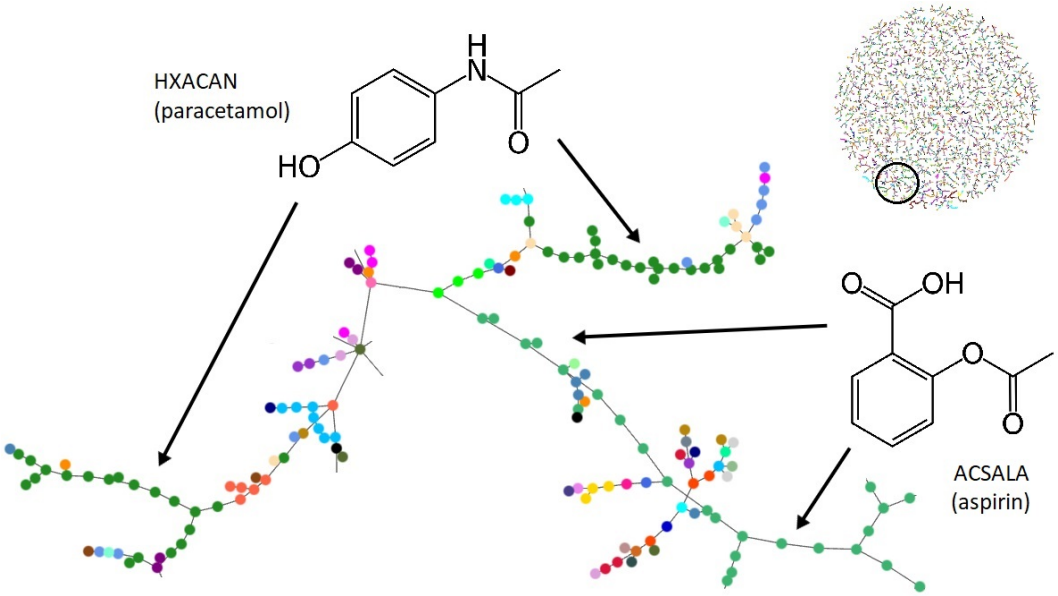

Figure 12: **Top right**: a Minimum Spanning Tree (MST) of 12576 structures in the Drug Subset of the Cambridge Structural Database. The distance between any crystals S, Q is $\mathrm{EMD}(\mathrm{PDD}(S;100), \mathrm{PDD}(Q;100))$. All structures from the same family are shown in one random color. **The zoomed-in part** shows the polymorphs of paracetamol (HXACAN) and aspirin (ACSALA), whose crystals have slightly different molecules but can be found on the same shelf in a typical pharmacy due to their similar pharmaceutical properties. All their polymorphs were found in close branches of this MST based on pure geometric invariants without using any chemistry.

each of these crystals is included in the folder `identical_by_mol_centers` and can be manually inspected (or opened in visualization software) to see the similarity.

CSD ID 1	CSD ID 2	EMD ($k = 100$)
ADESAG	REWPOB	1.45×10^{-5}
ROWBIQ	TUHMOA	9.00×10^{-4}
BESNOC01	XILXIA	3.93×10^{-3}
GIMHOA	XORCOX	5.63×10^{-3}
FUQMAG	GAVFEP	6.73×10^{-3}
COXJIJ	DUPZIY	6.79×10^{-3}
HAPNOE	HAPNUK	7.15×10^{-3}
LETYAM	UKUWUT	8.66×10^{-3}
DUCPUN	XUWMEI01	9.88×10^{-3}

Table 6: Selection of pairs of structures from the CSD with exactly zero EMD when compared by molecular centers. The closest pair ADESAG and REWPOB have identical fractional coordinates except for one difference at the fifth decimal place, a difference 100 times smaller than the typical threshold for considering two sites identical (0.001Å).

B Appendix B: examples and instructions for the PDD code and data files

B.1 Pseudocode for computing Pointwise Distance Distributions (PDD)

The algorithm accepts any periodic point set $S \subset \mathbb{R}^n$ in the form of a unit cell U and a motif $M \subset S$. The cell is given as a square $n \times n$ matrix with basis vectors in the columns, and the motif points in Cartesian form lying inside the unit cell. For dimension 3, the typical Crystallographic Information File (CIF) with six unit cell parameters and motif points in terms of the cell basis is easily converted to this format. Otherwise, the unit cell and motif points can be given directly, in any dimension.

Specifically, the PDD function's interface is as follows:

Input:

- `motif`: array shape (m, n). Coordinates of motif points in Cartesian form.
- `cell`: array shape (n, n). Represents the unit cell in Cartesian form.
- `k`: `int` > 0. Number of columns to return in PDD$(S; k)$.

Output:

- `pdd`: array with $k + 1$ columns.

Before giving the pseudocode, we outline some of the key objects and functions in use:

- A generator `g`, which creates points from the periodic set S to find distances to,
- KDTrees (canonically k is the dimension here, in our case it's denoted n), data structures designed for fast nearest-neighbour lookup in n-dimensional space.

Once `g` is constructed, `next(g)` is called to get new points from the infinite set S. The first call returns all points in the given unit cell (i.e. the motif), and successive calls returns points from unit cells further from the origin in a spherical fashion.

A KDTree is constructed with a point set T, then queried with another Q, returning a matrix with distances from all points in Q to their nearest neighbors (up to some given number, k below) in T, as well as the indices of these neighbors in T.

The functions `collapse_equal_rows` and `lexsort_rows`, which perform the collapsing and lexicographical sorting steps of computing PDD (Definition 3.1) respectively, are assumed to be implemented elsewhere.

The following pseudocode finds PDD$(S; k)$ for a periodic set S described by `motif` and `cell`:

```
def PDD(motif, cell, k):

    cloud = [] # contains points from S
    g = point_generator(motif, cell)

    # at least k points will be needed
    while len(cloud) < k:
        points = next(g)
        cloud.extend(points)

    # first distance query
    tree = KDTree(cloud)
    D_, inds = tree.query(motif, k)
    D = zeros_like(D_)

    # repeat until distances don't change,
    # then all nearest neighbors are found
    while not D == D_:
        D = D_
        cloud.extend(next(g))
        tree = KDTree(cloud)
        D_, inds = tree.query(motif, k)

    pdd = collapse_equal_rows(D_)
    pdd = lexsort_rows(pdd)
    return pdd
```

B.2 Instructions for the attached PDD code and specific examples

A Python script implementing Pointwise Distance Distributions along with examples can be found in the zip archive included in this submission. Python 3.7 or greater is required. The dependency packages are NumPy (< 1.22), SciPy ($\geq 1.6.1$), numba ($\geq 0.55.0$) and ase ($\geq 3.22.0$); if you do not wish to affect any currently installed versions on your machine, create and activate a virtual environment before the following.

Unzip the archive and in a terminal navigate to the unzipped folder. Install the requirements by running `pip install -r requirements.txt`. Then run `python` followed by the example script of choice, and then any arguments (outlined below), e.g.

```
$ python kite_trapezium_example.py

trapezium: [(0, 0), (1, 1), (3, 1), (4, 0)]
PDD:
[[0.5        1.41421356 2.         3.16227766]
 [0.5        1.41421356 3.16227766 4.        ]]

kite: [(0, 0), (1, 1), (1, -1), (4, 0)]
PDD:
[[0.25       1.41421356 1.41421356 4.        ]
 [0.5        1.41421356 2.         3.16227766]
 [0.25       3.16227766 3.16227766 4.        ]]

EMD between trapezium and kite: 0.874032
```

List of included example scripts and their parameters:

- `kite_trapezium_example.py` prints the PDDs of the finite 4-point sets K (kite) and T (trapezium) in Fig. 3, along with their Earth mover's distance.
- `1D_sets_example.py` shows that the 1D periodic sets in Fig. 3 are distinguished by their PDDs for any parameter $0 < r \leq 1$. This script requires the parameter r to be passed after the file name, e.g. '`python 1D_sets_example.py 0.5`'.
- `T2_14_15_example.py` compares the crystals shown in Fig. 7, whose original .CIFs are included. This optionally accepts the parameter k controlling the number of columns in the computed PDD, e.g. '`python T2_14_15_example.py --k 50`' compares by PDD with $k = 50$. If not included, $k = 100$ is used as the default.
- `CSD_duplicates_example.py` computes and compares the PDDs of the 5 pairs of isometric crystals from the CSD discussed in section 6, giving distances of exactly zero. This optionally accepts the parameter k controlling the number of columns in the computed PDD, in the same way as `T2_14_15_example.py` above.

If you wish to run the code on your own sets or CIF files, you can use the functions exposed in the main script `pdd.py`. Use `pdd.read_cif()` to parse a cif and return a crystal, or define one manually as a tuple `(motif, cell)` with NumPy arrays. Pass this as the first argument to `pdd.pdd()` with an integer `k` as the second to compute the PDD. Pass two PDDs to `pdd.emd()` to calculate the Earth mover's distance between them. For finite sets, the function `pdd.pdd_finite()` accepts just one argument, an array containing the points, and returns the PDD.

C Appendix C: rigorous proofs of Theorems 3.2, 4.2, 4.3, 4.4, 5.1.

Proof of Theorem 3.2. For any periodic point set $S \subset \mathbb{R}^n$, we first show that scaling up a cell U to a non-primitive cell keeps $\mathrm{PDD}(S;k)$ invariant. It suffices to scale up a cell U by a factor of l, say along the first basis vector $\vec{v}_1$ of U, then the number m of motif points of S is multiplied by l.

Then $D(S;k)$ from Definition 3.1 has the larger size $lm \times k$ but (due to periodicity) consists of l-tuples of identical rows of distances from points $p + i\vec{v}_1$, $i = 0, \dots, l-1$, to their k neighbors within S. Hence $\mathrm{PDD}(S;k)$ remains invariant, under all isometries due to the result below.

Now we show that the matrix $D(S;k)$ from Definition 3.1, hence $\mathrm{PDD}(S;k)$, is independent of a primitive unit cell. Let U, U' be primitive cells of a periodic set $S \subset \mathbb{R}^n$ with a lattice Λ. Any point $q \in S \cap U'$ can be translated by some $\vec{v} \in \Lambda$ to a point $p \in S \cap U$ and vice versa. These translations establish a bijection between the motifs $S \cap U \leftrightarrow S \cap U'$ and preserve distances. So $\mathrm{PDD}(S;k)$ is the same for both U, U'.

Now we prove that $\mathrm{PDD}(S;k)$ is preserved by any isometry $f : S \to Q$. Any primitive cell U of S is bijectively mapped by f to the unit cell $f(U)$ of Q, which should be also primitive. Indeed, if Q is preserved by a translation along a vector $\vec{v}$ that doesn't have all integer coefficients in the basis of $f(U)$, then $S = f^{-1}(Q)$ is preserved by the translation along $f^{-1}(\vec{v})$, which doesn't have all integer coefficients in the basis of U, so U was non-primitive. Since U and $f(U)$ have the same number of points from S and $Q = f(S)$, the isometry f gives a bijection between the motifs of S, Q.

For any finite or periodic sets S, Q, since f maintains distances, every list of ordered distances from $p_i \in S \cap U$ to its first k nearest neighbors in S, coincides with the list of the ordered distances from $f(p_i)$ to its first k neighbors in Q. These coincidences of distance lists give $\mathrm{PDD}(S;k) = \mathrm{PDD}(Q;k)$ after collapsing identical rows. □

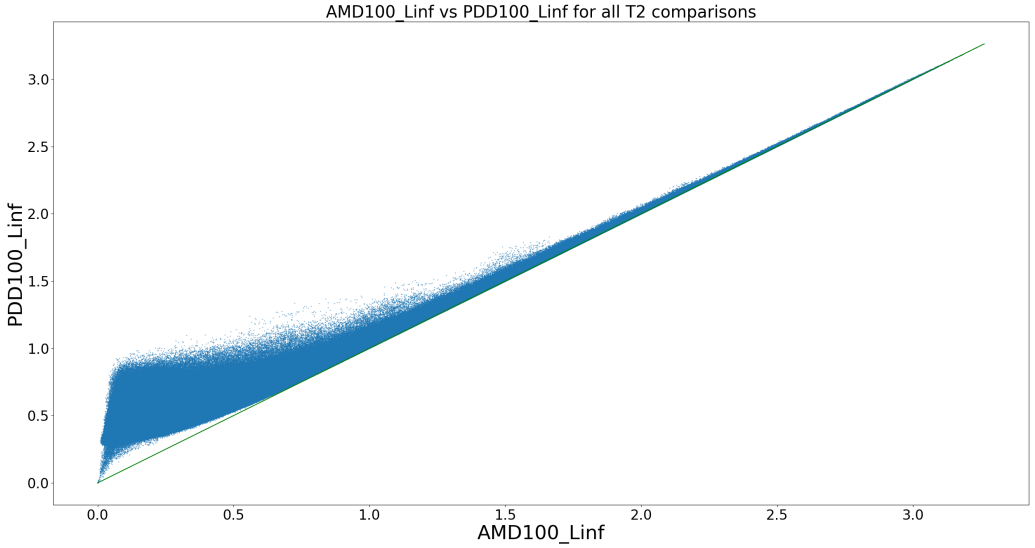

Figure 13: Illustration of Theorem 4.2. The coordinates of each point are the L_∞-based distances between $\mathrm{AMD}(S;100), \mathrm{AMD}(Q;100)$ and $\mathrm{PDD}(S;100), \mathrm{PDD}(Q;100)$ for all pairs S, Q of 5679 T2 crystals reported in [32].

Proof of Theorem 4.2. Considering $\mathrm{PDD}(S;k)$ as a weighted distribution of rows, $\mathrm{AMD}(S;k)$ is its centroid from [17, section 3]. The lower bound follows from [17, Theorem 1]. □

Lemma C.1. *For some $\varepsilon > 0$, let $g : S \to Q$ be a bijection between finite or periodic sets such that $|a - g(a)| \le \varepsilon$ for all $a \in S$. Then, for any $i \ge 1$, let $a_i \in S$ and $b_i \in Q$ be the i-th nearest neighbors of points $a \in S$ and $b = g(a) \in Q$, respectively. Then the Euclidean distances from the points a, b to their i-th neighbors a_i, b_i are 2ε-close to each other, i.e. $||a - a_i| - |b - b_i|| \le 2\varepsilon$.* ■

Proof. Shifting the point $g(a)$ back to a, assume that $a = g(a)$ is fixed and all other points change their positions by at most 2ε. Assume by contradiction that the distance from a to its new i-th neighbor b_i is less than $|a - a_i| - 2\varepsilon$. Then all first new i neighbors $b_1, \dots, b_i \in Q$ of a belong to the open ball with the center a and the radius $|a - a_i| - 2\varepsilon$.

Since the bijection g shifted every $b_1, \dots, b_i$ by at most 2ε, their preimages $g^{-1}(b_1), \dots, g^{-1}(b_i)$ belong to the open ball with the center a and the radius $|a - a_i|$. Then the i-th neighbor of a within S is among these i preimages, i.e. the distance from a to its i-th nearest neighbor should be strictly less

than the assumed value $|a - a_i|$. We get a contradiction assuming that the distance from a to its new i-th neighbor b_i is more than $|a - a_i| + 2\varepsilon$. □

Lemma C.2. *For $\varepsilon > 0$, let $g : S \to Q$ be a bijection between finite or periodic sets so that $|a - g(a)| \le \varepsilon$ for all $a \in S$. Then g changes the vector $\vec{R}_a(S) = (|a - a_1|, \dots, |a - a_k|)$ of the first k minimum distances from any point $a \in S$ to its k nearest neighbors $a_1, \dots, a_k \in S$ by at most 2ε in the L_∞-distance. So if $b_1, \dots, b_k \in Q$ are the k nearest neighbors of $b = g(a)$ within Q and $\vec{R}_b(S) = (|b - b_1|, \dots, |b - b_k|)$ is the vector of the first k minimum distances from $b = g(a)$, then the L_∞-distance $|\vec{R}_a(S) - \vec{R}_b(Q)|_\infty \le 2\varepsilon$.* ■

Proof. By Lemma C.1 every coordinate of the vector $\vec{R}_a(S)$ changes by at most 2ε. Hence the L_∞-distance from $\vec{R}_a(S)$ to the perturbed vector $\vec{R}_b(Q)$ is at most 2ε. □

Proof of Theorem 4.3. The bottleneck distance is $d_B(S, Q) = \inf\limits_{g:S\to Q} \sup\limits_{a\in S} |a - g(a)|$ between point sets S, Q. Then for any $\delta > 0$ there is a bijection $g : S \to Q$ such that $\sup\limits_{a\in S} |a - g(a)| \le d_B(S, Q) + \delta$. If the given sets S, Q are finite, one can set $\delta = 0$. Indeed, there are only finitely many bijections $S \to Q$, hence the infimum in the definition above is achieved for one of them.

By [24, Lemma 4.1], if the sets S, Q are periodic, they have a common lattice Λ. Any primitive cell U of Λ is a unit cell of S, Q, i.e. $S = \Lambda + (S \cap U)$ and $Q = \Lambda + (Q \cap U)$. Since the bottleneck distance $\varepsilon = d_B(S, Q) < \dfrac{r(S)}{4}$, we can define a bijection g from every point $a \in S$ to a closest point $g(a) \in Q$. If U is a non-primitive unit cell of S, the distance matrix $D(S; k)$ can be constructed as in Definition 3.1, but each row will be repeated $n(S)$ times, where $n(S)$ is $\mathrm{Vol}[U]$ divided by the volume of a primitive unit cell of S. So we can assume that S, Q share a unit cell U and have in U the same number $m(S) = m(Q)$, say both are equal to m. For any $k \ge 1$, we first define the simple 1-1 flow from the rows of $D(S; k)$ to the rows of $D(Q; k)$ by setting $f_{ii} = \frac{1}{m}$ and $f_{ij} = 0$ for $i \ne j$, where $i, j = 1, \dots, m$. Recall that Definition 3.1 collapses all rows of $D(S; k)$ that are identical to each other to a single row, similar for $D(Q; k)$. By summing up weights of collapsed rows, the above flow induces a flow from all distance vectors in $\mathrm{PDD}(S; k)$, e.g. $R_i(S)$ in the i-th row of $\mathrm{PDD}(S; k)$, to all distance vectors in $\mathrm{PDD}(Q; k)$.

Then $\mathrm{EMD}(\mathrm{PDD}(S; k), \mathrm{PDD}(Q; k)) \le \frac{1}{m} \sum\limits_{i=1}^{m} |\vec{R}_i(S) - \vec{R}_i(Q)|_\infty$, because EMD minimizes the cost over all flows in Definition 4.1. Since $|\vec{R}_i(S) - \vec{R}_i(Q)| \le 2(d_B(S, Q) + \delta)$ by Lemma C.2, we get $\mathrm{EMD}(\mathrm{PDD}(S; k), \mathrm{PDD}(Q; k)) \le \frac{1}{m} \sum\limits_{i=1}^{m} 2(d_B(S, Q) + \delta) = 2(d_B(S, Q) + \delta)$. Since the last inequality holds for any small $\delta > 0$, we get $\mathrm{EMD}(\mathrm{PDD}(S; k), \mathrm{PDD}(Q; k)) \le 2d_B(S, Q)$. □

For any point p in a lattice $\Lambda \subset \mathbb{R}^n$, the open *Voronoi domain* $V(\Lambda; p) = \{q \in \mathbb{R}^n \mid |q - p| < |q - p'|$ for any $p' \in \Lambda - p\}$ is the neighborhood of all points $q \in \mathbb{R}^n$ that are strictly close to p than to all other points p' of the lattice Λ [25].

The Voronoi domains $V(\Lambda; p)$ of different points $p \in \Lambda$ are disjoint translation copies of each other and their closures tile $\mathbb{R}^n$, so $\cup_{p\in\Lambda} \bar{V}(\Lambda; p) = \mathbb{R}^n$.

For a generic lattice $\Lambda \subset \mathbb{R}^2$, $V(\Lambda; p)$ is a centrally symmetric hexagon. Points $p, p' \in \Lambda$ are *Voronoi neighbors* if their Voronoi domains share a boundary point, so $\bar{V}(\Lambda; p) \cap \bar{V}(\Lambda, p') \ne \emptyset$.

Below we always assume that any lattice Λ is shifted to contain the origin 0, also any periodic point set $S = \Lambda + M$ has a point at 0.

Definition C.3 (neighbor set $N(\Lambda)$ and basis distances)**.** *For any lattice $\Lambda \subset \mathbb{R}^n$, the* neighbor set *(of the origin 0) is $N(\Lambda) = \Lambda \cap \bar{B}(0; r) - \{0\}$ for a minimum radius r such that $N(\Lambda)$ is not contained in any affine $(n-1)$-dimensional subspace of $\mathbb{R}^n$ and $N(\Lambda)$ includes all $n + 1$ nearest neighbors (within Λ) of any point $q \in V(\Lambda; p)$.*

For any point $q \in V(\Lambda;0)$, consider all n-tuples $(p_1,\dots,p_n)$ of points $p_i \in N(\Lambda)$ such that the vectors $\vec{p}_1,\dots,\vec{p}_n$, form a linear basis of $\mathbb{R}^n$. Order $p_1,\dots,p_n$ by their distances to q. Choose a lexicographically smallest list of basis distances $d_1(q) \le \dots \le d_n(q)$ *from the point q over all n-tuples $(p_1,\dots,p_n)$ described above.* ■

The lattice $\mathbb{Z}^2$ has the neighbor set $N(\mathbb{Z}^2) = \{(\pm1,0),(0,\pm1)\}$. If Λ is generated by $(2,0),(0,1)$, the neighbor set $N(\Lambda) \subset \Lambda$ includes the 3rd neighbors $(0,\pm2)$ of the points $(0,\pm0.4) \in V(\Lambda;0)$.

Indeed, if Definition C.3 has a radius $r < 2$, then $\Lambda \cap \bar{B}(0;r) - \{0\} = \{(0,\pm1)\}$ but the y-axis does not generate $\mathbb{R}^2$. For $q = (0,0.4)$, considering all pairs (p_1,p_2) among the four possibilities $((0,\pm1),(\pm2,0))$, we find the basis distances $d_1(q) = 0.6 < d_2(q) = \sqrt{0.4^2+2^2}$ for the 2nd and 3rd lattice neighbors $p_1 = (0,1)$ and $p_2 = (\pm2,0)$ of q.

Lemma C.4 ($N(\Lambda)$ bounds)**.** *The neighbor set $N(\Lambda)$ of any lattice Λ is covered by $\bar{B}(0;2R(\Lambda))$, where the* packing radius $R(\Lambda)$ *is the minimum $R > 0$ such that $\cup_{p\in\Lambda}\bar{B}(p;R) = \mathbb{R}^n$.* ■

Proof of Lemma C.4. Any point p in the closure $\bar{V}(\Lambda;0)$ has $n+1$ lattice neighbors (within Λ) among the origin $0 \in \Lambda$ and at least $2(2^n-1)$ Voronoi neighbors of 0.

In $\mathbb{R}^n$, any vertex of the boundary of $\bar{V}(\Lambda;0)$ is equidistant to at least $n+1$ points of Λ (the origin 0 and its n Voronoi neighbors). The longest of these distances is the covering radius $R(\Lambda)$. The closed ball $\bar{B}(0;2R(\Lambda))$ covers all Voronoi neighbors of 0, hence all points of $N(\Lambda)$. □

The condition of a linear basis in Definition C.3 guarantees that $n+1$ linearly independent vectors $\vec{p}_1,\dots,\vec{p}_n$ uniquely identify a point q by their basis distances $d_1(q),\dots,d_n(q)$.

Definition C.5 (a distance-generic periodic point set)**.** *A periodic point set $S = \Lambda + M \subset \mathbb{R}^n$ with the origin $0 \in \Lambda \subset S$ is called* distance-generic *if the following conditions hold.*

(C.5a) $\vec{p},\vec{q}$ are not orthogonal for any points $p,q \in S \cap V(\Lambda;0)$.

(C.5b) For any vectors $\vec{u},\vec{v}$ between any two pairs of points in S, if $|\vec{u}| = l|\vec{v}| \le 2R(\Lambda)$ for $l = 1,2$, then $\vec{u} = \pm l\vec{v}$ and $\vec{v} \in \Lambda$.

(C.5c) For any point $q \in V(\Lambda;0)$, let d_0 be the distance from q to its closest neighbor $p_0 = 0$ within Λ. Take any $p_1,\dots,p_n$ in the neighbor set $N(\Lambda)$ with distances $d_1 \le \dots \le d_n$ to q.

The $n+1$ spheres $C(p_i;d_i)$ with the centers p_i and radii d_i, $i = 0,\dots,n$, can meet at the single point $q \in V(\Lambda;0)$ only if $d_1 \le \dots \le d_n$ are the basis distances of q, hence $\vec{p}_1,\dots,\vec{p}_n$ form a linear basis of $\mathbb{R}^n$, only for at most two tuples $p_1,\dots,p_n \in N(\Lambda)$ symmetric in 0. ■

Condition (C.5b) means that all inter-point distances are distinct apart from necessary exceptions due to periodicity. Since any periodic set $S = \Lambda + M \subset \mathbb{R}^n$ is invariant under translations along vectors of its lattices Λ, condition (C.5b) for $|\vec{v}| \le 2\mathrm{diam}[U]$ can be checked only for vectors from all points in the Voronoi domain $V(\Lambda;0)$ to all points in the extended domain $3V(\Lambda;0)$.

Condition (C.5b) allows us to recognize *lattice distances* from any point $p \in M$ to its lattice translates $\Lambda + p$ in the row of $\mathrm{PDD}(S;k)$ representing p. Indeed, only a lattice distance d appears in the row together with $2d$ (and possibly with higher multiples) by condition (C.5b). Any lattice distance d and its multiple are repeated twice in every row, because any lattice is centrally symmetric.

All conditions of Definition C.5 can be written as algebraic equations via coordinates of motif points and basis vectors of a unit cell. Almost all $n+1$ spheres in $\mathbb{R}^n$ have no common points, so condition (C.5c) forbids very singular situations, which can be practically checked since the neighbor set $N(\Lambda)$ is finite for any lattice Λ containing 0. Hence any periodic point set can be made distance-generic by almost any perturbation of points and lattice basis.

Proof of Theorem 4.4. Assuming that $\mathrm{PDD}(S;k)$ is realizable by a periodic point set $S = \Lambda + M$, we will reconstruct all motif points $p \in V(\Lambda;0)$, uniquely up to the central symmetry of $\mathbb{R}^n$ with respect to 0. The given number m of points in a unit cell U of S is a common multiple of all

denominators in rational weights of the rows in the given matrix $\mathrm{PDD}(S;k)$. Enlarge $\mathrm{PDD}(S;k)$ replacing every row of a weight w by mw identical rows of weight $\frac{1}{m}$.

One can assume that the origin $0 \in \Lambda$ belongs to the motif M of S and is represented by the first row of $\mathrm{PDD}(S;k)$. If $\mathrm{PDD}(S;k)$ has $m \geq 2$ rows, we will reconstruct all other $m-1$ points of S within the open Voronoi domain $V(\Lambda;0)$. No points of S can be on the boundary of $V(\Lambda;0)$ due to condition (C.5b) on distinct distances.

Remove from each row of $\mathrm{PDD}(S;k)$ all *lattice distances* between any points of Λ. Then every remaining distance is between only points $p,q \in S$ such that $p-q \notin \Lambda$. Any point $q \in S \cap V(\Lambda;0) - \{0\}$ has its first lattice neighbour 0 at the distance $d_0 = |q|$ and a lexicographically smallest list of basis distances $d_1(q) < \cdots < d_n(q)$ from q to its further n lattice neighbors $p_1,\dots,p_n \in N(\Lambda) \subset \Lambda - 0$ such that the vectors $\vec{p}_1,\dots,\vec{p}_n$ form a basis of $\mathbb{R}^n$. All basis distances are distinct due to (C.5b). By Lemma C.4 they appear once in both rows of the points $0,q \in S$ in $\mathrm{PDD}(S;k)$ after $d_0 = |q|$.

Though the basis distances of q may not be the n smallest values appearing after $d_0 = |q|$ in the first and second rows of 0 and q, we will try all n-distance subsequences $d'_1 < \cdots < d'_n$ shared by both rows. Similarly, we cannot be sure that $n+1$ closest neighbors of q in Λ form an affine basis of $\mathbb{R}^n$. Hence we try all n-tuples of points $p_1,\dots,p_n \in N(\Lambda;0)$ whose vectors form a linear basis of $\mathbb{R}^n$.

For all finitely many choices above, we check if the $n+1$ spheres $S(p_i;d'_i)$, which are 1D circles (for $n=2$) or 2D spheres (for $n=3$), meet at a single point in $V(\Lambda;0)$, which is the reconstructed q.

Condition (C.5c) guarantees that these $n+1$ spheres can intersect at a single point in the open Voronoi domain $V(\Lambda;0)$ only if the three conditions of Definition C.5 hold. Firstly, the vectors $\vec{p}_1,\dots,\vec{p}_n$ should form a linear basis of $\mathbb{R}^n$. Secondly, if some distances $d_1 < \cdots < d_n$ are the basis distances from q to $p_1,\dots,p_n$ only if this list is the lexicographically smallest over all tuples $\{p_1,\dots,p_n\} \subset N(\Lambda)$ that form a linear basis. Thirdly, the single-point intersection happens only for two subsets $\{p_1,\dots,p_n\} \subset N(\Lambda)$ related by the central symmetry with respect to 0. This symmetry is an isometry preserving the lattice Λ and the distances $d_0 < d_1 < \cdots < d_n$. Making a choice by this inevitable ambiguity, we uniquely identify a point $q \in S \cap V(\Lambda;0) - \{0\}$ relative to the fixed Λ.

If the matrix $\mathrm{PDD}(S;k)$ has $m \geq 3$ rows, any further point $p \in (S - \{0,q\}) \cap V(\Lambda;0)$ will be uniquely determined as follows. Similarly to the point q above, we determine a position of p using its basis distances $d_0(p) < d_1(p) < \cdots < d_n(p)$ to points $0 = p_0, p_1, \dots, p_n \in N(\Lambda)$. At the end of reconstruction, we have a final choice between $\pm p$ symmetric with respect to the origin 0.

Since the second point q is already fixed, the third point p is also restricted by the distance $|p-q|$ appearing once only in the second and third rows of $\mathrm{PDD}(S;k)$. The distance $|p-q|$ doesn't help to resolve the ambiguity between $\pm p$ only if q belongs to the bisector of points equidistant to $\pm p$. In this case, $p,0,q$ form a right-angle triangle, which is forbidden by condition (C.5a). Hence p and any further point of $S \cap V(\Lambda;0)$ is uniquely determined by q and Λ. □

Proof of Theorem 5.1. It follows from [34, Theorem 14], which finds k nearest neighbors of m points in (a motif of) S before averaging these k-th distances for $\mathrm{AMD}(S;k)$. The near-linear time is due to the fast k-nearest neighbor search, whose parameterized complexity was recently revisited [28–31] by correcting mistakes in proofs and pseudo-code for cover trees. Instead of averaging, to get $\mathrm{PDD}(S;k)$, it remains only to lexicographically sort m lists of ordered distances in time $O(km \log m)$. Indeed, a comparison of ordered lists of the length k takes $O(k)$ time. □