Resolving the data ambiguity for periodic crystals

Anonymous Author(s)
Affiliation
Address
`email`

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 [18].

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.

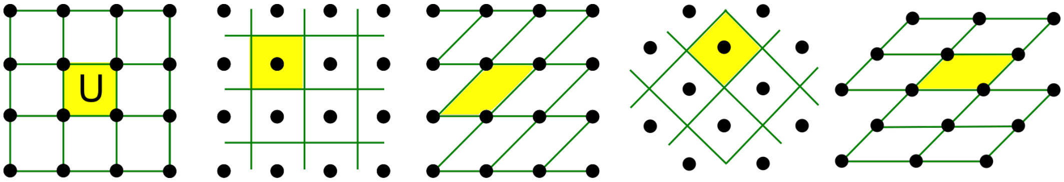

Figure 1: These isometric lattices are given by different cells and motifs. **1st**: $U = \langle(1,0),(0,1)\rangle$, $M = \{(0,0)\}$. **2nd**: $U = \langle(1,0),(0,1)\rangle$, $M = \{(\frac{1}{2},\frac{1}{2})\}$. **3rd**: $U = \langle(1,0),(1,1)\rangle$, $M = \{(0,0)\}$. **4th**: $U = \langle(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}),(-\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}})\rangle$, $M = \{(\frac{1}{2},\frac{1}{2})\}$. **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 [19]. 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.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

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 [17] 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.

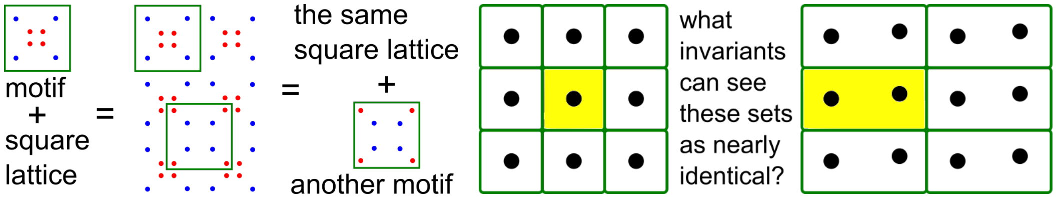

Figure 2: **Left** : even for a fixed cell of a lattice Λ, different motifs M can define isometric periodic sets $\Lambda + M$. **Right**: for almost any perturbation, the symmetry group and (the minimum volume of) any reduced cell discontinuously change, which justifies continuity (1.1d) in Problem 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 [4]. It suffices to classify up to isometry including mirror reflections. As a linear map, an isometry f reverses orientation if the determinant $\det(v_1, \dots, v_n)$ of basis vectors has the same sign as $\det(f(v_1), \dots, f(v_n))$.

The traditional approach to identify a periodic crystal is to use its conventional or reduced cell [31, 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, [17, section 1] and [20, 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 $d(I_1, I_3) \leq d(I_1, I_2) + d(I_2, I_3)$;*

(1.1d) continuity *: $d(I(S), I(Q)) \leq C d_B(S, Q)$ for a fixed constant C and any 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 done 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 [13] or a DNA-type code that also allows an explicit reconstruction for any periodic crystal.

For example, Computer Vision tries to identify humans or 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. Hence the ambiguity challenge exemplified by rigorously stated Problem 1.1 was the major obstacle on the road to an efficient materials design.

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 [19], 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 parameters of a unit cell and a 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 [18]. This 'embarrassment of over-prediction' wastes resources and time to run simulations and then analyze results, often by visual inspection, because there were no fast and reliable tools.

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 $\vec{p}-\vec{q}$. All conditions in Problem 1.1 are not completely fulfilled by the state-of-the art methods even for finite sets in $\mathbb{R}^n$. The non-isometric 4-point sets in Fig. 3 (left) are a counter-example to the completeness of the distance distribution [7].

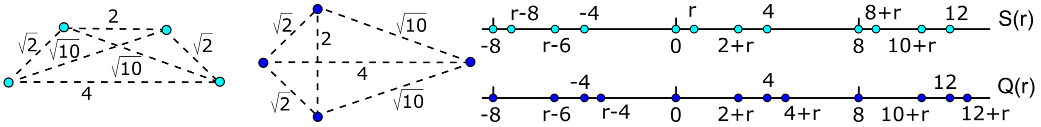

Figure 3: **Left**: point sets $K = \{(\pm 2, 0),\ (\pm 1, 1)\}$ and $T = \{(\pm 2, 0), (-1, \pm 1)\}$ can not be distinguished by their six pairwise distances $\sqrt{2}, \sqrt{2}, 2, \sqrt{10}, \sqrt{10}, 4$. **Right**: 1D periodic sets $S(r) = \{0, r, 2 + r, 4\} + 8\mathbb{Z}$ and $Q(r) = \{0, 2 + r, 4, 4 + r\} + 8\mathbb{Z}$ for $0 < r \leq 1$ have the same Patterson function [38, p. 197, Fig. 2]. All these pairs are distinguished by PDD in section 3.

The existence of an isometry between two m-point sets in $\mathbb{R}^n$ can be checked in time $O(m^{\lceil n/3 \rceil} \log m)$ by [8], which can be improved to $O(m \log m)$ in $\mathbb{R}^4$ [32]. Significant results on matching bounded rigid shapes and registration of finite point sets were obtained in [41, 24, 21, 16]. The research on graph isomorphisms [42, 27] can be potentially used for periodic graphs with fixed edges between points of a periodic set. These methods focused on binary true/false answers without continuously quantifying the similarity. Mémoli's seminal work on *distributions of distances* [35], also known as *shape distributions* [37, 5, 29, 26, 22], 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).

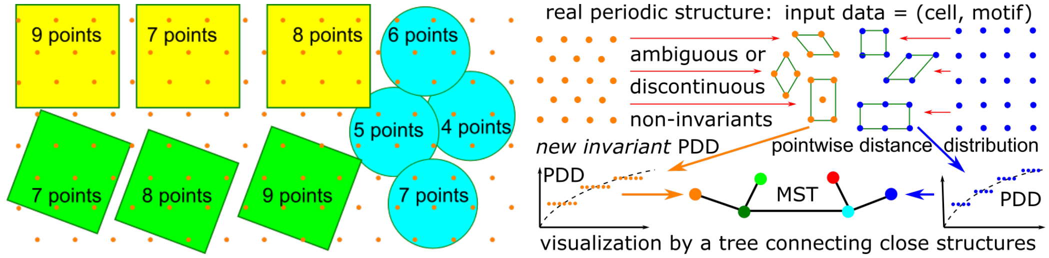

Figure 4: **Left**: hard to choose a finite subset that truly represents an infinite periodic set, discontinuity under perturbations of points or set sizes is similar to Fig. 2 (right). **Right**: an ambiguous input is transformed into the invariant PDD to visualize any dataset as a Minimum Spanning Tree (MST).

The Mercury software visually compares periodic structures [10] 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 RMSD fails the triangle inequality and is too slow for pairwise comparisons, see section 6. One natural similarity is the maximum displacement of atoms under thermal vibrations. This *bottleneck distance* $d_B(S, Q)$ between periodic point sets is the maximum Euclidean distance needed to perturb every point $p \in S$ to its unique match in Q. Since d_B is minimized over infinitely many bijections and points, d_B is computationally intractable. Even worse, $d_B(S, Q) = +\infty$ for the set of integers $S = \mathbb{Z}$ and $Q = (1 + \varepsilon)\mathbb{Z}$ scaled up for any small $\varepsilon > 0$. If we scale given periodic sets S, Q to the same density, the resulting $d_B(S, Q) < +\infty$ is the *wobbling* distance [23], which is discontinuous under perturbations, see the supplementary materials.

The discontinuity under perturbations is the major weakness of many past invariants including Voronoi diagrams, which should be matched via infinitely many rotations [36], space groups and other group-theoretic invariants [28]. 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 [46].

One of the oldest crystal descriptors is the X-ray diffraction pattern whose single crystal form is best for determining a 3D structure of an experimental crystal [12]. 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*,[39], 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.

For any $k \geq 1$, Edelsbrunner et al. [17] introduced the k-th *density function* $\psi_k(t)$ of a periodic point set $S = \Lambda + M \subset \mathbb{R}^3$ as the total volume of the regions within the unit cell U of Λ covered by exactly k balls $B(p; t)$ with a radius $t \geq 0$ 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 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 [17] 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 $Y = \{0, 1, 3\}$ [30, section 4] have the same density functions for all $k \geq 1$ [34, Example 10] but were distinguished in [20, Example 5b].

The latest advance [3] 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 [15, 6, 14]. The continuous metric on isosets [2, Corollary 35] has only an approximate algorithm, so Problem 1.1 remained open.

3 The Pointwise Distance Distribution $\mathrm{PDD}(S;k)$ of a periodic point set S

Distances to neighbors were considered in [20, Definition 5], though only their average was proved to be invariant under permutations of points. New Definition 3.1 below introduces the weights that make PDD continuous under perturbations in Theorem 4.3. See all proofs in the supplementary materials.

Definition 3.1 (Pointwise Distance Distribution PDD**)** *Let a periodic set $S = \Lambda + M$ have points $p_1, \dots, p_m$ in a unit cell. For $k \geq 1$, 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 $0 < r \leq 1$.

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

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

Any lattice $\Lambda \subset \mathbb{R}^n$ has a 1-point motif $M = \{p\}$, hence $\mathrm{PDD}(S;k)$ consists of a single row of increasing distances from p to all other points $\Lambda - \{p\}$. Fig. 5 (right) shows a honeycomb periodic set S whose motif consists of two symmetric points that have the same distances to all their neighbors, hence two rows of $D(S;k)$ collapse to a single vector $\mathrm{PDD}(S;k)$. 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 $\mathrm{AMD}_k(S(r))$ is independent of $0 < r < 1$, 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 $\mathrm{AMD}(S;k)$.

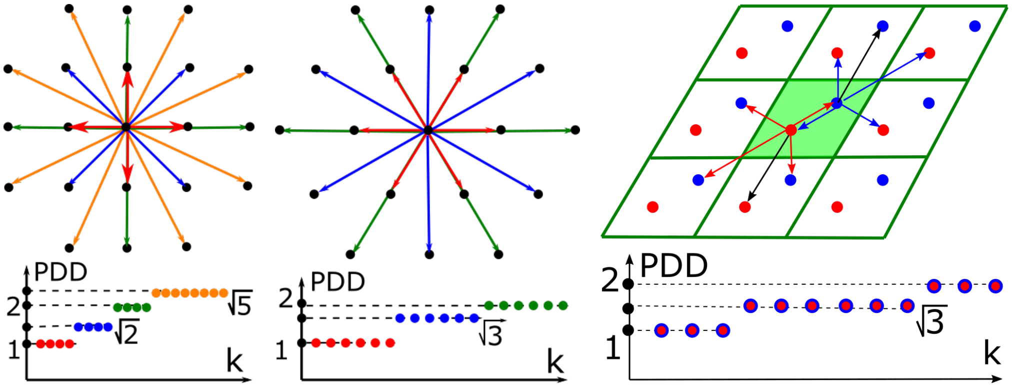

Figure 5: The square lattice (left), hexagonal lattice (middle), and honeycomb periodic set (right) with a minimum inter-point distance of 1 have $\mathrm{PDD}(S;k)$ with a single row of increasing distances.

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 [10], [4] 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. 5 (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 [45] 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 [40], 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* $\sum\limits_{j=1}^{m(Q)} f_{ij} \leq w_i(S)$ *for $i = 1, \dots, m(S)$,* $\sum\limits_{i=1}^{m(S)} f_{ij} \leq w_j(Q)$ *for $j = 1, \dots, m(Q)$,* $\sum\limits_{i=1}^{m(S)} \sum\limits_{j=1}^{m(Q)} f_{ij} = 1$. ■

The first condition $\sum\limits_{j=1}^{m(Q)} f_{ij} \leq w_i(S)$ 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 [40, appendix], needs $O(m^3 \log m)$ time for distributions of a maximum size m and is approximated in $O(m)$ time [43, 25].

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

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 $k \geq 1$, if finite or periodic sets $S, Q \subset \mathbb{R}^n$ satisfy $d_B(S, Q) < r(S)$, then* $\mathrm{EMD}(\mathrm{PDD}(S; k), \mathrm{PDD}(Q; k)) \leq 2 d_B(S, Q)$. ■

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 [20, Theorem 16]. Following the earlier work [17, section 5.1], the supplementary materials 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 [11, 33]. Theorem 4.4 assumes that a lattice Λ is given and reconstructs a periodic point set $S = \Lambda + M$ in any dimension $n \geq 2$.

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 the supplementary materials.

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 [17, 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 world's largest Cambridge Structural Database (CSD) since 1960s. Now the CSD has more 1.17M known periodic 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 [18] 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. 6, see the original files and more details in the supplementary materials. When this pair was compared by another free software Platon [44], 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.

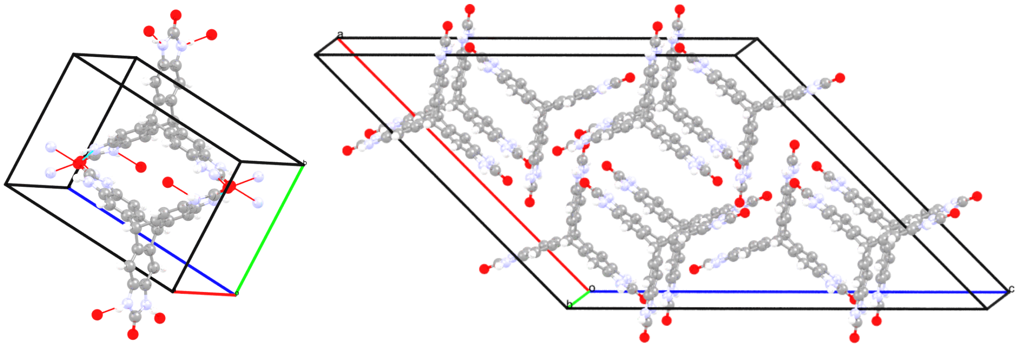

Figure 6: Crystals 14 and 15 based on the T2 molecule have very different Crystallographic Information Files (with different motifs in unit cells of distinct shapes) but are nearly identical up to isometry.

For example, a rough sampling of the density functions $\psi_k(t)$ from [17] 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 [10] 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 [10] 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

6 Conclusions and a discussion of remaining limitations and societal impact

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 $|\mathrm{AMD}(S;100) - \mathrm{AMD}(Q;100)|_\infty \leq 0.01$. 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 [20, 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.

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 [20, 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 [9]. 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*, 2021.

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

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

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

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

[7] 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.

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

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

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

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

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

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

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

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

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

[17] 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.

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

[19] C Groom et al. The cambridge structural database. *Acta Cryst B*, 72(2):171–179, 2016.

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

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

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

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

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

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

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

[27] Y Aflalo et al. On convex relaxation of graph isomorphism. *Proceedings of the National Academy of Sciences*, 112(10):2942–2947, 2015.

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

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

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

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

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

[33] V Kurlin. Mathematics of 2-dimensional lattices, 2022.

[34] Vitaliy Kurlin. Density functions of periodic sequences. *arxiv:2205.02226*, 2022.

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

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

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

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

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

[40] 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.

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

[42] N Shervashidze. Weisfeiler-lehman graph kernels. *J. Machine Learning Research*, 12(9), 2011.

[43] 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.

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

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

[46] 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.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The main theoretical results are stated Theorems 3.2, 4.3, 5.1 , 4.4 and are proved in the supplementary materials. The key experiments are described in sections 5 and 6 with more details in the supplementary materials.
   (b) Did you describe the limitations of your work? [Yes] Yes, discussed in section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Potential fraud and possible retractions of several papers are discussed in section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] In particular, the paper was properyl anonymized.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Yes, all formal theorems are completely stated. For Theorem 4.4, the definition of a distance-generic periodic point set couldn't fit the page limit and is now in the supplementary materials.
   (b) Did you include complete proofs of all theoretical results? [Yes] Yes, in the supplementary materials.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Yes, in the supplementary materials.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] The PDD algorithm used only one parameter $k = 100$ (the number of atomic neighbors) in addition to a typical input (Crystallographic Information File). However, increasing k only adds more invariants to $PDD(S;k)$ without changing the previous values.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Yes, the threshold of 10^{-12}meters (atomic scale) was used to identify and further investigate (near-)duplicates, however exact duplicates were reported at distance 0.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Yes, the specifications of the modest desktop computer appear at the beginning of section 5.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Yes, the Cambridge Structural Database [19] and the database of T2 crystals reported in [18].
   (b) Did you mention the license of the assets? [Yes] Individual structures can be freely downloaded from the Cambridge Structural Database (CSD) by their 6-letter codes given in the paper and supplementary materials.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] For convenience, the supplementary materials include the pairs of duplicate structures and other important crystals mentioned in the paper.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] All used data is freely available.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] All data is non-personal.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] No crowdsourcing, no research with human subjects.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] No crowdsourcing, no human subjects.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] No crowdsourcing, no human subjects.