POINTWISE DISTANCE DISTRIBUTIONS OF DISCRETE SETS*

DANIEL E. WIDDOWSON† AND VITALIY A. KURLIN‡

Abstract. The basic input of a real object is a discrete set of points such as corners or other salient features. For our main applications in chemistry points represent atomic centers in a molecule or solid materials. We study the problem of classifying discrete (finite and periodic) sets of unordered points under isometry, which is any transformation preserving distances in a metric space.

The experimental noise motivates the new practical requirement to make such invariants Lipschitz continuous so that perturbing every point in its ε-neighborhood changes the invariant up to a constant multiple of ε in a suitable distance satisfying all metric axioms. Because given points are unordered, the key challenge is to compute all invariants and metrics in a near-linear time of the input size.

We define a Pointwise Distance Distribution (PDD) for any discrete set and proves in addition to the properties above the completeness of PDD for all periodic sets in general position. The PDD can compare nearly 1.5 million crystals from the world's four largest databases within hours on a modest desktop computer. The impact is upholding the data integrity in crystallography because the PDD will not allow anyone to claim a 'new' material as a noisy disguise of a known crystal.

Key words. isometry classification, complete invariant, continuous metric, periodic crystal

MSC codes. 68U05, 51N20, 52C07

1. Introduction: motivations, problem statement, and contributions.

This paper is a substantial extension of the 10-page conference version at NeurIPS 2022 [60]. The original paper introduced the Pointwise Distance Distribution (PDD) as an isometry invariant of a periodic set of points in any Euclidean space $\mathbb{R}^n$, and claimed the key properties (Lipschitz continuity, near-linear time computability, and generic completeness) without proofs. This extended version defines PDD for any discrete sets in a metric space and rigorously proves the properties above in the finite and periodic case. We also modify the invariants in a more convenient form, speed up the original implementation almost by an order of magnitude, and report much larger experiments on the world's largest experimental databases of periodic materials.

The practical motivations for a new continuous approach to classifying physical objects will be clear after introducing the basic concepts in Definition 1.1–1.5.

DEFINITION 1.1 (a *discrete set* S in a *metric space* X with a *metric* d_X). *A* metric space *is any (possibly, infinite) set X of any objects with a* distance metric $d : X \times X \to \mathbb{R}$ *satisfying the metric axioms: (1)* coincidence $d_X(a,b) = 0$ *if and only if $a = b$, (2)* symmetry $d_X(a,b) = d_X(b,a)$, *and (3)* triangle inequality $d_X(a,b) + d_X(b,a) \geq d_X(a,c)$ *for $a, b, c \in X$. A set $S \subset X$ is called* discrete *if there is a constant $\varepsilon > 0$ such that all objects of S are ε-separated, so $d_X(a,b) \geq \varepsilon$ for $a, b \in S$.*

An example of a discrete set S is a finite set of points in $\mathbb{R}^n$ with the Euclidean metric denoted by $|p - q|$ for points $p, q \in \mathbb{R}^n$. The positivity of a metric $d(a,b) \geq 0$ follows from the axioms above: $2d(a,b) = d(a,b) + d(b,a) \geq d(a,a) = 0$. Without the first axiom, the distance d is called a *pseudo-metric* and can be even the zero: $d(a,b) = 0$ for all a, b. If the triangle inequality fails with any additive error $\varepsilon > 0$, the results of clustering such as k-means and DBSCAN may not be trustworthy [49].

*Submitted to the editors on October 3, 2024.

Funding: Royal Society APEX fellowship APX/R1/231152, New Horizons grant EP/X018474/1

†Department of Computer Science, Liverpool, UK (D.E.Widdowson@liverpool.ac.uk).

‡Department of Computer Science, Liverpool, UK (vkurlin@liv.ac.uk, http://kurlin.org).

DEFINITION 1.2 (lattice, unit cell, motif, periodic set). *Vectors* $v_1, \dots, v_n \in \mathbb{R}^n$ *form a* basis *if any vector in* $\mathbb{R}^n$ *can be written as* $v = \sum_{i=1}^{n} t_i v_i$ *for unique* $t_1, \dots, t_n \in \mathbb{R}$. *Any basis generates the* lattice $\Lambda = \{\sum_{i=1}^{n} c_i v_i \mid c_1, \dots, c_n \in \mathbb{Z}\}$ *and the* unit cell $U = \{\sum_{i=1}^{n} t_i v_i \mid 0 \le t_i < 1, i = 1, \dots, n\}$. *For any finite set of points (called a motif)* $M \subset U$, *the sum* $S = M + \Lambda = \{p + v \mid p \in M, v \in \Lambda\}$ *is called a* periodic point set.

Any unit cell U is a parallepiped with a partial boundary: we exclude the points with any coefficient $t_i = 1$, $i = 1, \dots, n$, for convenience so that $\mathbb{R}^n$ is tiled by the shifted cells $U + v$ for $v \in \Lambda$ without overlaps. Any lattice is an example of a periodic set with one point in a motif. Any periodic point set $S = M + \Lambda$ can be considered a finite union $\bigcup_{p \in M} (p + \Lambda)$ of lattices whose origins are shifted to all points $p \in M$.

If we double a unit cell in one direction, e.g. by taking the basis $2v_1, v_2, \dots, v_n$, the doubled motif $M \cup (M + v_1)$ with the sublattice on the new basis defines the original periodic point set $S = M + \Lambda$. A basis and its cell U of S are called *primitive* if U has the smallest volume among all unit cells of S. Fig. 1 (left) shows a square lattice in $\mathbb{R}^2$, which (as any lattice) can be generated by infinitely many primitive bases. Even if we fix a basis, Fig. 1 (middle) shows that different motifs in the same primitive cell U define equivalent periodic point sets, which differ only by translation.

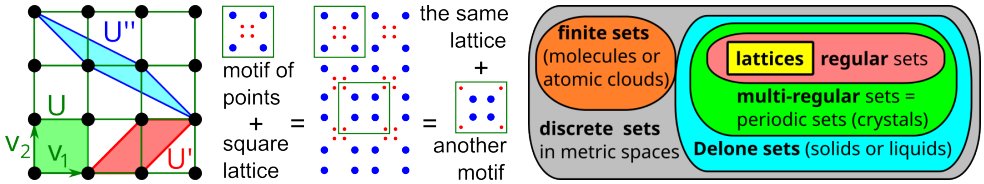


FIG. 1. ***Left**: a lattice can be defined by many primitive bases. **Middle**: a periodic set can be defined by different pairs (basis, motif). **Right**: a hierarchy of discrete sets, which model periodic crystals, amorphous solids, and liquids with points at atomic centers, see Definitions 1.1, 1.2, 1.5.*

Finite and periodic point sets represent molecules and periodic crystals at the atomic scale by considering zero-sized points at all atomic centers. In theory, chemical bonds can be modelled by straight-line edges between atomic centers. However, even the strongest covalent bonds within a molecule depend on various thresholds for distances and angles. In other words, these bonds are not real sticks in space and only abstractly represent inter-atomic interactions, while atomic nuclei are proper physical objects. We model all materials at the fundamental level of atoms, which will turn out to suffice for experimental materials. Because the same object such as a lattice can be defined in many different ways, Definition 1.3 formalizes an equivalence.

DEFINITION 1.3 (equivalence relation). *An* equivalence *is a binary relation (denoted by* $\sim$*) on any kind of objects satisfying the following axioms: (1)* reflexivity*: any objects* S *is equivalent to itself, so* $S \sim S$*; (2)* symmetry*: if* $S \sim Q$*, then* $Q \sim S$*; (3)* transitivity*: if* $S \sim Q$ *and* $Q \sim T$*, then* $S \sim T$*. Any object* S *defines its* equivalence class $[S] = \{Q \mid Q \sim S\}$ *as the full collection of all objects* Q *equivalent to* S.

The transitivity axiom justifies that all equivalence classes are disjoint: if $[S]$ and $[T]$ share a common object Q, then $[S] = [T]$. Any well-defined classification should be based on an equivalence, whose practical examples are considered below.

DEFINITION 1.4 (isometry, rigid motion in $\mathbb{R}^n$). *In a metric space X, an* isometry *is any map $f : X \to X$ that preserves inter-point distances, i.e. $d(f(p), f(q)) = d(p, q)$ for all $p, q \in X$. In $\mathbb{R}^n$, all isometries decompose into translations, rotations, and reflections, which all form the Euclidean group $\mathrm{E}(n)$. If reflections are excluded, orientation-preserving isometries are also called* rigid motions *and form group $\mathrm{SE}(n)$.*

The rigid motion (denoted by $\cong$) is the strongest equivalence for many objects in practice because translations and rotations of a molecule or solid material keep all their properties at least under the same ambient conditions such as temperature and pressure. The isometry (denoted by $\simeq$) is only slightly weaker by allowing reflections. Taking compositions with a uniform scaling in $\mathbb{R}^n$ or including (say) affine transformations, we get weaker equivalences that define smaller spaces of classes.

This paper focuses on isometry as a more general equivalence defined in any metric space. Our main problem will be to continuously parametrize equivalence classes of (various kinds of) discrete sets under isometry. Delone sets were introduced [20] as (r, R)-systems in $\mathbb{R}^n$ and make sense in any metric space X. Let $\bar{B}(p; r) = \{q \in X \mid d(p, q) \le r\}$ be the closed ball with a center $p \in X$ and a radius r.

DEFINITION 1.5 (Delone sets and m-regular sets). *In a metric space X, a* Delone *set S is any subset of X satisfying the following conditions:*

(a) packing*: there is a radius $r > 0$ such that the closed balls $\bar{B}(p; r)$ for all points $p \in S$ are disjoint or, equivalently, all distances between points of S are at least $2r$;*

(b) covering*: there is a radius $R > 0$ such that $\bar{B}(p; R)$ for all $p \in S$ cover X, i.e. $\bigcup_{p \in S} \bar{B}(p; R) = X$, or, equivalently, $\bar{B}(p; R)$ for any $p \in X$ has at least on point of S.*

*A Delone set is called m-*regular *if S splits into m classes under the* global isometry equivalence*: $p \sim q$ if there is an isometry $f : X \to X$ such that $f(S) = S$, $f(p) = q$.*

The packing condition implies that S is a discrete set X by specifying a minimum inter-point distance $\varepsilon = 2r$ and is well-motivated by the fact that real atoms strongly repel each other at very short distances. The covering condition says that X has no unbounded 'empty' balls without any points of S and is also motivated by the absence of infinite round pores in solid materials, liquids, and even some dense gases.

All m-regular sets for $m \ge 1$ are also called multi-regular, while 1-regular sets are often called *regular*. Any lattice $\Lambda \subset \mathbb{R}^n$ is regular because the required isometry $f : \Lambda \to \Lambda$ mapping a point $p \in \Lambda$ to another $q \in \Lambda$ is the translation by the vector $q - p$. Similarly, any periodic point set S is m-regular, where m is upper bounded by the size of a motif M of S. A honeycomb periodic set in $\mathbb{R}^2$, which models graphene, is not a lattice due to two points in a primitive unit cell but is regular. The regularity means that S looks the same when viewed from any point of S. Fig. 1 (middle) shows a 2-regular set whose points split into red and blue classes under the global isometry equivalence. [21, Theorem 1.3] proved that any multi-regular Delone set is periodic.

A finite set in $\mathbb{R}^n$ is not a Delone set but any finite subset of a finite metric space is Delone. The latter special case is indicated by cyan and orange regions slightly touching each other in Fig. 1 (middle). All other inclusions are strict, not to scale.

The key tool in classifying under an equivalence is an *invariant* that is a function I taking the same value on all equivalent objects. For a finite set $S \subset \mathbb{R}^n$, the number m of points is an isometry invariant, but the geometric average $\dfrac{1}{m} \sum\limits_{p \in S} p$ is not, so the center of mass cannot be reliably used to distinguish rigid conformations of molecules.

We state the mapping problem for any discrete sets under isometry, though the same conditions make sense for many other objects, e.g. graphs and polygonal meshes, and equivalence, e.g. rigid motions and affine transformations in $\mathbb{R}^n$.

PROBLEM 1.6 (**mapping problem** for spaces of discrete sets under isometry). *For a metric space X with a metric d_X, find a map $I : \{$discrete sets of unordered points in $X\} \to$ a metric space with a metric d satisfying the following conditions.*

(a) ***Completeness****: any sets $S \simeq Q$ are isometric if and only if $I(S) = I(Q)$.*

(b) ***Realizability****: the image $\{I(S) \mid S \subset X\}$ is parametrized so that taking any value of I from this image allows us to reconstruct $S \subset X$ uniquely up to isometry of X.*

(c) ***Lipschitz continuity****: there is a constant λ such that if Q is obtained by perturbing each point of S up to ε in the metric d_X, then $d(I(S), I(Q)) \le \lambda\varepsilon$.*

(d) ***Computability****: the invariant I, the metric d, and the reconstruction of $S \subset X$ from $I(S)$ can be computed in a time that depends polynomially on the input sizes.*

For any finite set $S \subset X$, its input size is the number m of points. For any periodic point set $S \subset \mathbb{R}^n$, its input size is the size of a motif M from Definition 1.2 because a Crystallographic Information File (CIF) specifying for any periodic crystal a basis and atomic coordinates in this basis has a linear length $O(m)$ in the motif size. Some infinite Delone sets can described in a finite form, e.g. quasi-periodic crystals [54] can be obtained as projections of periodic crystals in higher dimensions.

We leave these general cases for future work and will focus on finite and periodic point sets, which already cover many applications where Problem 1.6 was open.

The completeness in (1.6a) implies that the invariant I is a descriptor with *no false negatives* and *no false positives* for all discrete sets, and hence can be considered a DNA-style code that uniquely identifies any isometry class. The realizability in (1.6b) is even stronger and enables us to sample the space of realizable invariants and reconstruct the resulting set S, while a real DNA code is insufficient to grow a living organism. The Lipschitz continuity in (1.6c) is motivated by the ever-present thermal vibrations and experimental noise. Fig. 2 (left) shows that almost any perturbation of points can arbitrarily scale up a primitive cell. This inherent discontinuity of traditional cell-based representations remained a practical loophole in crystallography at least since 1965 [43] and allowed disguising known materials by a slight perturbation abruptly changing the space group and even the primitive cell volume, and also by replacing some chemical elements to avoid detection by chemical composition.

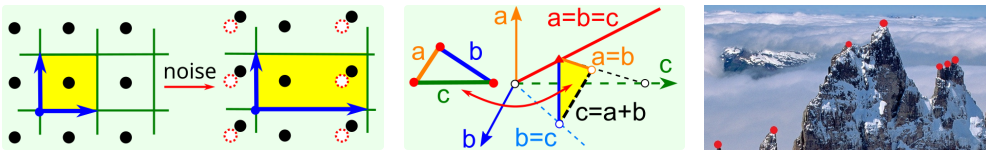


FIG. 2. ***Left****: the symmetry group and a reduced cell discontinuously change under tiny noise.* ***Middle****: the space of 3 points under isometry is parametrized by inter-point distances $0 < a \le b \le c \le a + b$* ***Right****: energy landscapes of crystals show optimized structures as isolated peaks of height$= -$energy. To see beyond the 'fog', we need a map parametrized by invariants in Problem 1.6.*

Fig. 2 (middle) shows a solution of Problem 1.6 for $m = 3$ points saying that any triangle is determined under isometry by 3 ordered inter-point distances. Real or simulated crystals are local optima (mountain peaks) in Fig. 2 (right) on a continuous space of (isometry classes of) periodic point sets, whose 'geography' was unknown.

Contributions. We introduce the Pointwise Distance Distribution for any discrete set in a metric space. This generality is of broad interest to experts in computational geometry and applications to physical objects from molecules to solid or even liquid materials. The previously unpublished aspects are the rigorous proofs of the Lipschitz continuity, near-linear time computability, and generic completeness in the finite and periodic case. The linear-time algorithms and the hierarchical nature of PDD computations have become extremely important for big databases, especially in the last years when thousands of artificial materials were claimed as 'new' without checking for duplication with known crystals. The decisive advance is closing this discontinuity loophole in crystallography, which is demonstrated in the world's largest databases.

2. Review of rigorous approaches to mapping spaces of discrete sets. This section reviews progress in solving Problem 1.6 for finite and periodic point sets by proof-based methods than by experimental studies, which are reviewed in [60, 62]. Finite sets have two subcases: ordered points (easy) and unordered (much harder).

Ordered finite sets. Kendall's shape theory [38] studies m ordered points $p_1, \dots, p_m \in \mathbb{R}^n$ whose complete isometry invariant is the distance matrix [53, 39] or the Gram matrix of scalar products $p_i \cdot p_j$, see [59, chapter 2.9], [58]. A brute-force extension to m unordered points requires $m!$ matrices due to $m!$ permutations ruled out by (1.6d).

Unordered finite sets (also called point clouds). Extending the case of $m = 3$ points in Fig. 2 (middle), Boutin and Kemper proved in 2004 that the unordered distribution of distances between m points uniquely determines a generic m-point cloud $C \subset \mathbb{R}^n$ under isometry [6]. This general position means almost all clouds apart form a measure 0 subspace among all clouds. For any cloud C of m unordered points in a metric space X, writing all distances in increasing order gives the Sorted Distance Vector $\mathrm{SDV}(C)$ of $\frac{m(m-1)}{2}$ values computable in time $O(m^2 \log m)$ The space of 4-point clouds in $\mathbb{R}^2$ has dimension 5 because 6 inter-point distances satisfy one polynomial equation saying that the tetrahedron on these points has volume 0. Fig. 3 shows a 4-parameter family of pairs of non-isometric clouds with the same SDV.

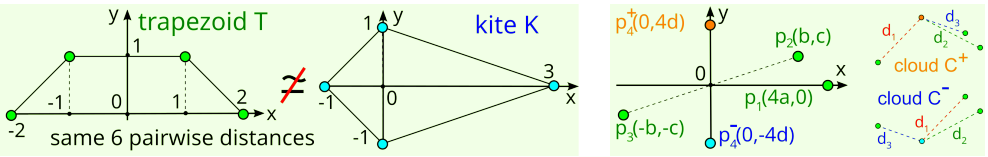


FIG. 3. *Non-isometric clouds of 4 points with the same 6 pairwise distances.* ***Left****: the trapezoid T has points $(\pm 2, 1)$, $(\pm 4, -1)$. The kite K has $(5, 0)$, $(-3, 0)$, $(-1, \pm 2)$.* ***Right****: the infinite family of non-isometric clouds $C^+ \not\simeq C^-$ sharing p_1, p_2, p_3 and depending on parameters $a, b, c, d > 0$.*

Problem 1.6 expands the question 'Can we hear the shape of a drum?' [36] which has has the negative answer in terms of 2D polygons that are indistinguishable by spectral invariants [30, 31, 50, 19, 46]. Problem 1.6 looks for stronger invariants that can completely 'sense' as in (1.6b), not only 'hear', the rigid shape of any cloud.

Computational geometry studied earlier versions of Problem 1.6 by developing canonical representations of point clouds [1, 7, 8, 4], which can be considered complete invariants, or (separately) metrics between isometry classes of clouds. For example, any metric between fixed clouds extends to their isometry classes [34, 15, 13] by minimization over infinitely many transformations from the group $\mathrm{E}(n)$. This extension of the Hausdorff distance [33] for m-point clouds in $\mathbb{R}^2$ has time $O(m^5 \log m)$, see

[14, 29]. The Gromov-Wasserstein metrics [47] are defined for metric-measure spaces also by minimizing over infinitely many correspondences between points, but cannot be approximated with a factor less than 3 in polynomial time unless P=NP, see Corollary 3.8 in [52] and polynomial algorithms for partial cases in [45]. Also, computing a metric between isometry classes of clouds is only a part of Problem 1.6. Indeed, to efficiently navigate on Earth, in addition to distances between cities, we need a satellite-type view of the whole planet and hence a realizable continuous invariant I, which can be considered an analog of the latitude and longitude coordinates.

Classical crystallography studied

Geometric Data Science has gradually developed and solved simpler versions of Problem 1.6 since 2020 when the continuity condition was first stated for lattices [48]. The case of 2D lattices was finished in [41] with a slightly weaker Hölder continuity (because the Lipschitz continuity is impossible under perturbations of a lattice basis) for a stronger relation under rigid motion in $\mathbb{R}^2$, see continuous chiral distances and geographic-style maps in [10, 9]. The case of 3D lattices is being finalized in [40].

For general periodic point sets, the latest advance announced in [60] without proofs is the Pointwise Distance Distribution (PDD), which solves Problem 1.6 for finite and periodic point sets in general position. This PDD previously appeared as a local distribution of distances in the finite case [47] without studying the conditions of Problem 1.6. For finite point clouds in $\mathbb{R}^n$, the complete invariants under rigid motion with Lipschitz continuous metrics were developed in [62]. The high polynomial-time complexity of these latest invariants motivates using the much faster PDD in practice.

3. The Pointwise Distance Distribution in the finite and periodic case. This section introduces the Pointwise Distance Distribution (PDD) for any finite subset M of a discrete set S in a metric space X. If S is finite, we always set $M = S$. If S is periodic, M is a motif of S, but PDD will depend only on S, not on M.

Definition 3.1 (PDD and AMD invariants). *Let $M = \{p_1, \dots, p_m\}$ be a finite subset of a discrete set S in a metric space X. Fix an integer $k \geq 1$. For every point $p_i \in M$, let $d_1(p) \leq \dots \leq d_k(p)$ be the distances from p to its k nearest neighbors within the full set S (not restricted to M). The matrix $D(S;k)$ has m rows consisting of the distances $d_1(p_i), \dots, d_k(p_i)$ for $i = 1, \dots, m$. If any $l \geq 1$ rows coincide, we collapse them into a single row and assign the weight l/m to this row. The resulting matrix of maximum m rows and $k+1$ columns including the extra (say, 0-th) column of weights is the* Pointwise Distance Distribution $\mathrm{PDD}(S;k)$. *The* Average Minimum Distance AMD_i *is the weighted average of the i-th column in $\mathrm{PDD}(S;k)$ for each $i = 1, \dots, k$. Let $\mathrm{AMD}(S;k)$ denote the vector $(\mathrm{AMD}_1, \dots, \mathrm{AMD}_k)$.*

Example 3.2 (4-point clouds T, K in Fig. 3 (left)). *Table 1 shows the 4×3 matrices $D(S;3)$ from Definition 3.1. 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 T, K are distinguished already by PDD for $k = 1$.*

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

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

Theorem 3.1 (invariance of PDD). *For any finite set S in a metric space X*

TABLE 1
Each point in $T, K \subset \mathbb{R}^2$ from Figure 3 (left) has distances to other points in increasing order. After keeping only distances (not neighbors), the resulting PDD invariants distinguish $T \not\cong K$.

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

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

or a periodic point set $S \subset \mathbb{R}^n$, the Pointwise Distance Distribution $\mathrm{PDD}(S;k)$ *from Definition 3.1 and* $\mathrm{AMD}(S;k)$ *are isometry invariants of S for any $k \geq 1$.*

Proof. For any finite set $S \subset X$, the isometry invariance follows from the fact that any isometry preserves all inter-point distances by Definition 1.4.

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. Then $\mathrm{PDD}(S;k)$ remains invariant under all isometries due to the arguments below.

We will show below that the matrix $D(S;k)$ and 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 of $\mathbb{R}^n$. Any primitive cell U of S is bijectively mapped by f to the unit cell $f(U)$ of $Q = f(S)$, which should be also primitive. Indeed, if Q is preserved by a translation along a vector 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 $f^{-1}(v)$, 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 periodic sets S, Q, because 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. □

Because PDD has ordered columns (by the index k of neighbors) and unordered rows (representing points in a motif), all such matrices even with different numbers of rows can be compared by Earth Mover's Distance, or other metrics on weighted distributions, see Definition 4.1. We can convert any PDD into a fixed-size matrix, which can be flattened into a vector for easy comparisons, while keeping the conti-

nuity and almost all invariant data. Any distribution of m unordered values can be reconstructed from its m moments defined below. When all weights w_i are rational as in our case, the distribution can be expanded to equal-weighted values $a_1, \dots, a_m$. The m moments can recover all $a_1, \dots, a_m$ as roots of a polynomial of degree m whose coefficients are expressed via the m moments [44]. For example, any reals a, b are the roots of $t^2 - (a+b)t + ab$, where $ab = \frac{1}{2}((a+b)^2 - (a^2+b^2))$.

Let A be any unordered set of real numbers $a_1, \dots, a_m$ with weights $w_1, \dots, w_m$, respectively, such that $\sum_{i=1}^{m} w_i = 1$. For any integer $l \geq 1$, the l-th *moment* [37, section 2.7] is $\mu_l(A) = \sqrt[l]{m^{1-l} \sum_{i=1}^{m} w_i a_i^l}$, so $\mu_1(A) = \sum_{i=1}^{m} w_i a_i$ is the usual average. For $l \geq 2$, we avoid subtracting μ_1 from a_i's, which would convert μ_2 into the standard deviation σ, and normalize by the factor $m^{(1/l)-1}$ to guarantee the continuity of all moments with the Lipschitz constant $\lambda = 2$.

Definition 3.3 (Pointwise Distance Moment PDM[l]). *Fix integers $l, h \geq 1$. For a column A of the Pointwise Distance Distribution* PDD$(S;k)$*, which consists of unordered numbers $a_1, \dots, a_m$ with weights from Definition 3.1, write the new column $(\mu_1(A), \dots, \mu_l(A))$. The new $l \times k$) matrix is the* Pointwise Distance Moment PDM[l]$(S;k)$.

PDM[1]$(S;k)$ is called the vector of *Average Minimum Distances* $\mathrm{AMD}(S;k) = (\mathrm{AMD}_1, \dots, \mathrm{AMD}_k)$. The matrix PDM[$l$] has ordered rows and columns but is a bit weaker than PDD (with the same $h, k_1, \dots, k_h$) because each column is reconstructable from its moments (for large enough l) only up to permutation, but PDM[l] more quickly filters distant crystals. We can flatten any matrix PDM[l] with indexed entries to a vector. Vectors $u, v \in \mathbb{R}^m$ of distances are compared by $L_\infty(u,v) = \max\limits_{i=1,\dots,m} |u_i - v_i|$ which controllably changes under perturbations of interatomic distances.

The number k of neighbors is considered not a parameter that affects the invariant but as a degree of approximation like the number of decimal places on a calculator.

If we increase k, more columns with larger values are added to PDD$(S;k)$ but all previous distances remain the same. We will describe the asymptotic of PDD$(S;k)$.

Definition 3.4 (Point Packing Coefficient PPC). *The volume of the unit ball in $\mathbb{R}^n$ is $V_n = \dfrac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)}$, where $\Gamma(m) = (m-1)!$ and $\Gamma(\frac{m}{2}+1) = \sqrt{\pi}(m-\frac{1}{2})(m-\frac{3}{2})\cdots\frac{1}{2}$ for integer $m \geq 1$. Let a periodic point set $S \subset \mathbb{R}^n$ have a unit cell U of volume* $\mathrm{vol}(U)$ *The* Point Packing Coefficient *is* $\mathrm{PPC}(S) = \sqrt[n]{\dfrac{\mathrm{vol}(U)}{mV_n}}$.

Lemma 3.5 is a slightly variation of [61, Lemma 12], see its proof in appendix A.

Lemma 3.5 (distance bounds). *Let $S \subset \mathbb{R}^n$ be a periodic point set with a unit cell U of a diameter d. For any point $p \in S \cap U$, let $d_k(S;p)$ be the distance from p to its k-th nearest neighbor in S. Then $c(S)\sqrt[n]{k} - d < d_k(S;p) \leq c(S)\sqrt[n]{k} + d$, $k \geq 1$.*

Theorem 3.6 (asymptotic of PDD$(S;k)$ as $k \to +\infty$). *For any point p in a periodic point set $S \subset \mathbb{R}^n$, let $d_k(S)$ be the distance from p to its k-th nearest neighbor in S. Then* $\lim\limits_{k\to+\infty} \dfrac{d_k(S)}{\sqrt[n]{k}} = \mathrm{PPC}(S)$.

Proof. The proof follows from Lemma 3.5 after dividing the proved upper bound $|d_k(S) - c(S)\sqrt[n]{k}| \leq d$ by $\sqrt[n]{k}$ and taking the limit as $k \to +\infty$. ∎

4. Lipschitz continuous Earth Mover's Distance on PDD invariants. The continuity of $\mathrm{PDD}(S;k)$ under perturbations of S will be measured by the Earth Mover's Distance [51], which applies 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 $\sum_{i=1}^{m(S)} w_i(S) = 1$.

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 the matrix $\mathrm{PDD}(S;k)$.

For any vectors $R_i = (r_{i1}, \dots, r_{ik})$ and $R_j = (r_{j1}, \dots, r_{jk})$ of k values, we use the L_∞-distance $|R_i - R_j|_\infty = \max_{l=1,\dots,k} |r_{il} - r_{jl}|_\infty$ to get the Lipschitz constant $\lambda = 2$. Other Minkowski metrics L_q for $q \geq 1$ will lead to higher Lipschitz constants, also depending on k, while L_∞ is motivated by bounded vibrations of atoms.

Definition 4.1 (Earth Mover's Distance EMD). *Let finite or periodic sets S, Q in $\mathbb{R}^n$ 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 $\mathrm{EMD}(I(S), I(Q)) = \sum_{i=1}^{m(S)} \sum_{j=1}^{m(Q)} f_{ij} |R_i(S) - R_j(Q)|$ *for $f_{ij} \in [0,1]$ subject to $\sum_{j=1}^{m(Q)} f_{ij} \leq w_i(S)$ for $i = 1, \dots, m(S)$, $\sum_{i=1}^{m(S)} f_{ij} \leq w_j(Q)$ for $j = 1, \dots, m(Q)$, $\sum_{i=1}^{m(S)} \sum_{j=1}^{m(Q)} f_{ij} = 1$.*

The first condition $\sum_{j=1}^{m(Q)} f_{ij} \leq w_i(S)$ 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} for $j = 1, \dots, m(Q)$. The second condition $\sum_{i=1}^{m(S)} f_{ij} = w_j(Q)$ 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 $\sum_{i=1}^{m(S)} \sum_{j=1}^{m(Q)} f_{ij} = 1$ forces to 'flow' all rows $R_i(S)$ to all rows $R_j(Q)$.

The EMD satisfies all metric axioms [51, appendix], needs $O(m^3 \log m)$ time for distributions of a maximum size m and is approximated in $O(m)$ time [55, 27].

Theorem 4.1 (lower bound of EMD). *For finite or periodic point sets $S, Q \subset \mathbb{R}^n$, we have $\mathrm{EMD}(\mathrm{PDD}(S;k), \mathrm{PDD}(Q;k)) \geq ||\mathrm{AMD}(S;k) - \mathrm{AMD}(Q;k))||_\infty$, $k \geq 1$.*

Proof. 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]. □

Theorem 4.2 uses bounded perturbations of points up to ε in the metric d_X of an ambient space X. Because atoms are not considered outliers or noise, such perturbations can be formalized as the *bottleneck distance* $d_B(S,Q) = \inf_{g:S \to Q} \sup_{p \in S} |p - g(p)|$ minimized over all bijections $g : S \to Q$ between infinite sets. [60, Example 2.1] shows that the 1-dimensional lattices $\mathbb{Z}$ and $(1+\delta)\mathbb{Z}$ have $d_B = +\infty$ for any $\delta > 0$.

If any lattices have equal density (or unit cell volume), they have a finite bottleneck distance d_B by [22, Theorem 1(iii)]. If we consider only periodic point sets $S, Q \subset \mathbb{R}^n$ with the same density (or unit cells of the same volume), $d_B(S,Q)$ becomes

a well-defined wobbling distance [11], which is still discontinuous under perturbations by [60, Example 2.2], see related results for non-periodic sets in [42].

Recall that 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 4.2 substantially generalizes the fact that shifting any points up to ε changes the distance between them up to 2ε.

THEOREM 4.2 (continuity of PDD). *Let Q be obtained from a finite or periodic set S by perturbing every point of S up to ε in the metric d_X of an ambient metric space X. In the case of a periodic point set $S \subset \mathbb{R}^n$, we also assume that $r(S) > \varepsilon$. Then $\mathrm{EMD}(\mathrm{PDD}(S;k), \mathrm{PDD}(Q;k)) \le 2\varepsilon$ for any $k \ge 1$.*

Theorem 4.2 will be proved by using Lemmas 4.2, 4.3, 4.4 below. The first two lemmas appeared in similar forms in [26, Lemma 2] and [61, Lemma 8], respectively, so appendix A includes their more detailed proofs for completeness.

LEMMA 4.2 (common lattice). *Let periodic point sets $S, Q \subset \mathbb{R}^n$ have a bottleneck distance $d_B(S,Q) < r(Q)$, where $r(Q)$ is the packing radius. Then S, Q have a common lattice Λ with a unit cell U such that $S = \Lambda + (U \cap S)$ and $Q = \Lambda + (U \cap Q)$.*

LEMMA 4.3 (perturbed distances). *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. $||a - a_i| - |b - b_i|| \le 2\varepsilon$.*

LEMMA 4.4 (perturbed distance vectors). *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 4.3 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 4.2. 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 Lemma 4.2, 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 $\varepsilon = d_B(S,Q) < r(S)$, we can define a bijection g from every point $a \in S$ to its 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 $i \neq j$, 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 $\mathrm{PDD}(Q;k)$, e.g. $R_j(Q)$ in the j-th row of $\mathrm{PDD}(Q;k)$. Then $\mathrm{EMD}(\mathrm{PDD}(S;k), \mathrm{PDD}(Q;k)) \leq \frac{1}{m} \sum\limits_{i=1}^{m} |\vec{R}_i(S) - \vec{R}_i(Q)|_\infty$, because EMD minimizes the cost over all flows in Definition 4.1. Because $|\vec{R}_i(S) - \vec{R}_i(Q)| \leq 2(d_B(S,Q) + \delta)$ by Lemma 4.4, we get $\mathrm{EMD}(\mathrm{PDD}(S;k), \mathrm{PDD}(Q;k)) \leq \frac{1}{m} \sum\limits_{i=1}^{m} 2(d_B(S,Q) + \delta) = 2(d_B(S,Q) + \delta)$. Since the last inequality holds for any small $\delta > 0$, we get $\mathrm{EMD}(\mathrm{PDD}(S;k), \mathrm{PDD}(Q;k)) \leq 2d_B(S,Q)$. ∎

5. Generic completeness of Pointwise Distance Distributions. We prove the generic completeness in both finite (easy) and periodic (much harder) cases.

THEOREM 5.1. *Any cloud $C \subset \mathbb{R}^n$ of m unordered points with distinct inter-point distances. can be reconstructed from* $\mathrm{PDD}(C; m-1)$*, uniquely up to isometry.*

Proof. Under the given condition of general position, every inter-point distance $|p - q|$ between points $p, q \in C$ appears twice in $\mathrm{PDD}(C; m-1)$: once in the row of p and once in the row of q. After choosing an arbitrary order on the points, $\mathrm{PDD}(C; m-1)$ suffices to reconstruct the classical distance matrix on ordered points. This distance matrix enables a uniquely reconstruction of C up to isometry [53, 39]. ∎

For a periodic point set $S \subset \mathbb{R}^n$, the generic completeness of PDD is much harder because infinitely many distances between points of S are repeated due to periodicity. We introduce a few auxiliary concepts to define *distance-generic* periodic sets later.

For any point p in a lattice $\Lambda \subset \mathbb{R}^n$, the open *Voronoi domain* $V(\Lambda; p) = \{q \in \mathbb{R}^n \mid |q - p| < |q - p'| \text{ for any } p' \in \Lambda - p\}$ 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 [23].

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 $\bar{V}(\Lambda; p) \cap \bar{V}(\Lambda, p') \neq \emptyset$. 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 5.2 (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 $d_1(q) \leq \dots \leq d_n(q)$ *from the point q over all n-tuples $(p_1, \dots, p_n)$ described above.*

The lattice $\mathbb{Z}^2$ has the neighbor set $N(\mathbb{Z}^2) = \{(\pm 1, 0), (0, \pm 1)\}$. If Λ is generated by $(2,0), (0,1)$, the neighbor set $N(\Lambda) \subset \Lambda$ includes the 3rd neighbors $(0, \pm 2)$ of the points $(0, \pm 0.4) \in V(\Lambda; 0)$. Indeed, if Definition 5.2 has a radius $r < 2$, then $\Lambda \cap \bar{B}(0; r) - \{0\} = \{(0, \pm 1)\}$ 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 $((0, \pm 1), (\pm 2, 0))$, 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 $p_2 = (\pm 2, 0)$ of q.

LEMMA 5.3. *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 5.3. 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 5.2 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 5.4 (a distance-generic set). *A periodic point set $S = \Lambda + M \subset \mathbb{R}^n$ with the origin $0 \in \Lambda \subset S$ is* distance-generic *if the following conditions hold.*

(5.4a) The vectors $\vec{p}, \vec{q}$ are not orthogonal for any points $p, q \in S \cap V(\Lambda; 0)$.

(5.4b) 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 $\vec{v} \in \Lambda$.

(5.4c) 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 points $p_1, \dots, p_n$ 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 (5.4b) 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 (5.4b) for $|\vec{v}| \le 2\text{diam}[U]$ 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 (5.4b) allows us to recognize *lattice distances* from any point $p \in M$ to its lattice translates $\Lambda + p$ in the row of $\text{PDD}(S; k)$ representing p. Indeed, only a lattice distance d appears in the row together with $2d$ (and possibly with higher multiples) by condition (5.4b). Any lattice distance d and its multiple are repeated twice in every row, because any lattice is centrally symmetric.

All conditions of Definition 5.4 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 (5.4c) 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.

The number m of points in a unit cell U is an isometry invariant because any isometry maps U to another cell of the same size. In dimensions $n = 2, 3$, any lattice Λ can be reconstructed from its isometry invariants [18, 41, 40].

Theorem 5.1 assumes that a lattice Λ is given and reconstructs a periodic point set $S = \Lambda + M$ in any dimension $n \ge 2$.

THEOREM 5.1 (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, $\mathrm{PDD}(S;k)$.*

Proof. 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 (5.4b) 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 neighbor 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 (5.4b). By Lemma 5.3 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 (5.4c) guarantees that these $n+1$ spheres can intersect at a single point in the open Voronoi domain $V(\Lambda;0)$ only if all the conditions of Definition 5.4 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 $d_0 < d_1 < \cdots < d_n$. By making a choice to resolve 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 (5.4a). Hence p and any further point of $S \cap V(\Lambda; 0)$ is uniquely determined by q and Λ. ∎

6. Near-linear time algorithm for Pointwise Distance Distributions. The key input sizes for computing $\mathrm{PDD}(S;k)$ are the number m of points in a unit cell U and the number k of neighbors. The full input consists of k and a periodic point set $S \subset \mathbb{R}^n$ given by a cell basis and a motif of m points with coordinates in this basis as described in Definition 1.2. For a fixed dimension n and other parameters, the asymptotic complexity of $\mathrm{PDD}(S;k)$ will depend near linearly in each of k, m.

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 in Definition 3.1, see the detailed code in appendix B. For any unit cell U, consider the diameter $\mathrm{diam}(U) = \sup\limits_{p,q \in U} |p-q|$ and the *skewness* $\nu = \frac{\mathrm{diam}(U)}{\sqrt[n]{\mathrm{vol}(U)}}$.

THEOREM 6.1 (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 $\mathbb{R}^n$.*

Appendix A has a proof of Lemma 6.1, which appeared in [61, Lemma 11].

LEMMA 6.1 (bounds on points within a ball). *Let $S \subset \mathbb{R}^n$ be any periodic point set with a unit cell U, which generates a lattice Λ and has a diameter $d = \mathrm{diam}(U)$. For any point $p \in S \cap U$ and a radius r, consider the* lower *union $U'(p;r) = \bigcup\{(U+\vec{v})$ such that $\vec{v} \in \Lambda, (U+\vec{v}) \subset \bar{B}(p;r)\}$ and the* upper *union $U''(p;r) = \bigcup\{(U+\vec{v})$ such that $\vec{v} \in \Lambda, (U+\vec{v}) \cap \bar{B}(p;r) \neq \emptyset\}$. Then the number of points from S in the closed ball $\bar{B}(p;r)$ with center p and radius r has the bounds*
$$\left(\frac{r-d}{c(S)}\right)^n \le m\frac{\mathrm{vol}[U'(p;r)]}{\mathrm{vol}[U]} \le |S \cap \bar{B}(p;r)| \le m\frac{\mathrm{vol}[U''(p;r)]}{\mathrm{vol}[U]} \le \left(\frac{r+d}{c(S)}\right)^n.$$

Proof of Theorem 6.1. Let the origin $0 \in \mathbb{R}^n$ be in the center of the unit cell U. If d is the diameter of U, any point $p \in M = S \cap U$ is covered by the closed ball $\bar{B}(0, 0.5d)$. By Lemma 3.5, all k neighbors of p are covered by the ball $\bar{B}(0;r)$ of radius $r = c(S)\sqrt[n]{k} + 1.5d$. To generate all Λ-translates of M within $\bar{B}(0;r)$, we gradually extend U in spherical layers by adding more shifted cells until we get the upper union $U''(0;r) \supset \bar{B}(0;r)$. By Lemma 6.1 the union $U''(0;r)$ includes k neighbors of motif points and has at most $\mu \le m\frac{\mathrm{vol}[U''(0;r)]}{\mathrm{vol}[U]} \le$

$$\le \left(\frac{c(S)\sqrt[n]{k} + 2.5d}{c(S)}\right)^n = \left(\sqrt[n]{k} + \frac{2.5d}{c(S)}\right)^n = O(2^n(k + m(2.5\nu)^n V_n)) \text{ points.}$$

To get the last expression, we use the rough estimate $(a+b)^n \le 2^n(a^n+b^n)$ with $a^n = k$, $b^n = \left(\frac{2.5d}{c(S)}\right)^n = \frac{(2.5d)^n}{\mathrm{vol}[U]} m V_n = m(2.5\nu)^n V_n$ for $\nu = \frac{d}{\sqrt[n]{\mathrm{vol}[U]}}$.

A cover tree on μ points can be built in time $O(\mu \log \mu)$, where hidden parameters were recently revisited in [24, 25] by correcting mistakes in proofs and pseudo-code for cover trees. The ordered lists of distances are the rows of the matrix $D(S;k)$. 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. Using $\log \mu = n \log 2 + O(\log(k + m(2.5\nu)^n V_n)) = nO(\log(km))$, the total time is

$$O(\mu \log \mu + mk \log^2 k) = O(2^n(k + m(2.5\nu)^n V_n) n \log(km) + mk \log^2 k) =$$

$$= O((5\nu)^n V_n km \log^2(km)), \text{ which is near linear in both key inputs } k, m. \quad \square$$

7. Upholding the data integrity of the world's largest databases. This section reports thousands of previously unknown (near-)duplicates in the world's five largest databases [56, 32, 63, 35]. The sizes in Table 2 below are the numbers of all periodic crystals (no disorder and full geometric data) in September 2024 (total number 1,433,650, nearly 1.5 million), see details of experiments in appendix C.

TABLE 2
Links and sizes (numbers of pure periodic crystals) of the world's five largest databases.

database and web address	crystals
CSD : Cambridge Structural Database http://ccdc.cam.ac.uk/solutions/software/csd	831,126
COD : Crystallography Open Database http://www.crystallography.net/cod	344,127
ICSD : Inorganic Crystal Struct. Database icsd.products.fiz-karlsruhe.de/en	105,162
MP : Materials Project by the Berkeley lab http://next-gen.materialsproject.org	153,235

To neutralize the effect of increasing distances AMD_k with respect to k, Theorem 3.6 motivated us to subtract the asymptotic $\mathrm{PPC}(S)\sqrt[3]{k}$ in Definition 7.1.

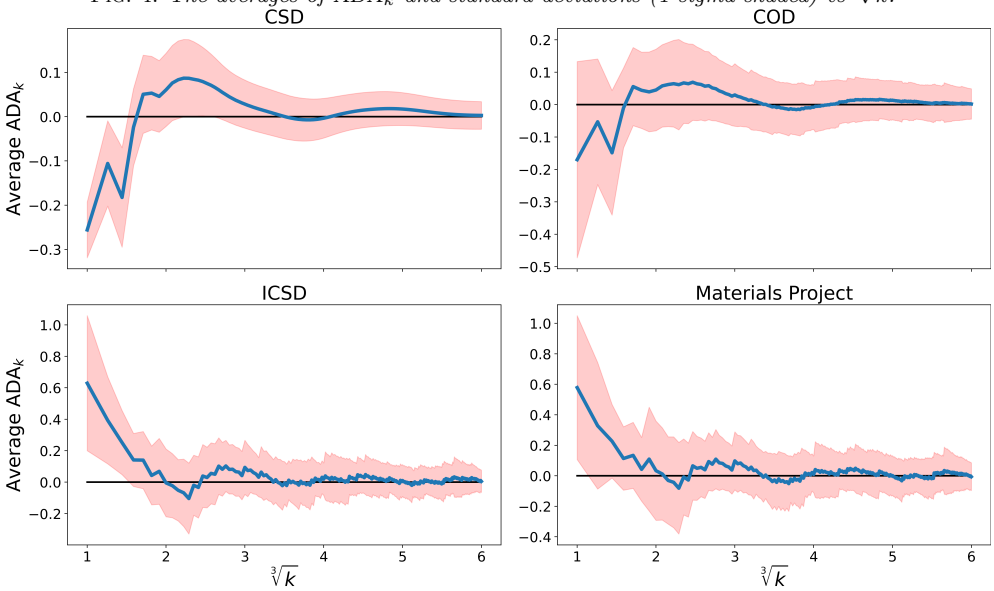
DEFINITION 7.1 (Average/Pointwise Deviations from Asymptotic: ADA, PDA). *Distances in* $\mathrm{PDD}(S;k)$ *are increasing in* k *by Theorem 3.6, to avoid the dominance by the largest value of* k*, the vector* $\mathrm{ADA}(S;k)$ *and matrix* $\mathrm{PDA}(S;k)$ *are obtained from* $\mathrm{AMD}(S;k), \mathrm{PDD}(S;k)$ *by subtracting* $\mathrm{PPC}(S)\sqrt[n]{i}$ *from each* i*-th coordinate/column, respectively, for all* $i = 1, \dots, k$.

While $\mathrm{AMD}_k(S)$ monotonically increases in k, the invariants $\mathrm{ADA}_k(S)$ can be positive or negative as deviations around the asymptotic $\mathrm{PPC}(S)\sqrt[3]{k}$. Fig. 4 reveals geometric differences between the mainly organic databases CSD and Crystallography Open Database (COD) versus the inorganic databases ICSD and MP. In all cases, ADA_k decreases to 0 as $k \to +\infty$ justifying our computations up to $k = 100$ below.

We first used the vector $\mathrm{ADA}(S;100)$ to find nearest neighbors across all databases by kd-trees [28] up to $L_\infty \le 0.01$Å. Since the smallest inter-atomic distances are about $1\text{Å} = 10^{-10}$m, atomic displacements up to 0.01Å are considered experimental noise. For the closest pairs found by $\mathrm{ADA}(S;100)$, the stronger $\mathrm{PDA}(S;100)$ can have only larger distances $\mathrm{EMD} \ge L_\infty$ by [17, section 3]. The CSD, COD, ICSD are expected to have only experimental structures. MP is obtained from ICSD by extra simulations.

Table 3 shows that the well-curated 59-year-old CSD has 0.9% near-duplicate crystals, while more than a third of the ICSD consists of near-duplicates that are geometrically almost identical so that all atoms can be matched by an average perturbation up to 0.01Å. [3, section 6] described thousands of more embarrassing exact duplicates, where chemical elements were replaced while keeping all coordinates fixed. These replacements are physically impossible without more substantial perturbations, so several journals are investigating integrity [12], see more examples in Appendix C.

FIG. 4. *The averages of* ADA_k *and standard deviations (1 sigma shaded) vs* $\sqrt[3]{k}$.



The bold numbers in Table 3 count near-duplicates within each database, which should be filtered out for any analysis or machine learning else the ground truth data becomes skewed, see the percentages for different thresholds in Fig. 3 (right). Other numbers are matches across different databases.

TABLE 3

Count and percentage of all pure periodic crystals in each database (left) found to have a near-duplicate in other databases (top) by the distance EMD < 0.01Å *on matrices* PDA$(S;100)$.

databases	CSD		COD		ICSD		MP	
	count	%	count	%	count	%	count	%
CSD	**7687**	**0.9**	272649	32.8	4649	0.6	21	0.0
COD	276328	80.3	**19231**	**5.6**	36553	10.6	5239	1.52
ICSD	4736	4.5	48899	46.5	**35189**	**33.5**	16386	15.6
MP	64	0.0	11989	7.82	14312	9.3	**19177**	**12.5**

In the past, the (near-)duplicates were impossible to detect at scale, because the traditional comparison through iterative alignment of 15 (by default) molecules by the COMPACK algorithm [16] is too slow for all-vs-all comparisons. Tables 4 and 5 compare the running times: **hours** of PDA$(S;100)$ vs **years** of RMSD, extrapolated for the same machine from the median time 117 ms (average 582 ms) on 500 random pairs in the CSD. On the same 500 pairs, PDA$(S;100)$ for two crystals per pair and distance EMD took only 7.48 milliseconds on average. All experiments were done on a typical desktop (AMD Ryzen 5 5600X 6-core, 32GB RAM).

8. Conclusions, limitations, future work and growing impact. For more than 100 years, crystals were classified almost exclusively by discrete invariants such as space groups or by using reduced cells, which 3D structures from diffraction patterns. Fig. 2 (left) showed that any known crystal can also be disguised by changing a unit

TABLE 4

Times in seconds (less than 8.5 hours in total) to find near-duplicates in Table 3 with EMD $\leq$ 0.01Å *on* PDA$(S;100)$ *across five major databases, compare with years in Table 5.*

databases	CSD	COD	ICSD	MP	sum of times, hrs:min:sec
CSD	403.6	1979.3	42.9	6.2	0:40:32
COD	1979.3	609.7	2249.8	1525.4	1:46:05
ICSD	42.9	2249.8	3362.1	4428.1	2:35:78
MP	6.2	1525.4	4428.1	4431.8	2:53:21

TABLE 5

These times for all comparisons by COMPACK [16] are extrapolated from the median time of 117 ms on 500 random pairs from the CSD on the same typical desktop, which completed Table 3 of near-duplicates across all five databases within 8.5 hours.

database	periodic crystals	all unordered pairs	time, seconds	years
CSD	831,126	345,384,798,375	4.04×10^{10}	1280.5
COD	344,127	59,211,524,001	6.93×10^{9}	219.7
ICSD	105,162	5,529,470,541	6.47×10^{8}	20.5
MP	153,235	11,740,405,995	2.75×10^{9}	87.1

cell, shifting atoms a bit, changing chemical elements, then claimed as 'new', see appendix C. Such artificially generated structures threaten the integrity of experimental databases [12], which are already skewed by previously undetectable near-duplicates.

These challenges motivated the stronger questions "how much different?" and "what is behind a code?", which were formalized in Problem **??** aiming for a continuous parametrization of the space of crystals . One limitation is that PDD is not proved to be complete and a random PDD may not be realizable by a real crystal because inter-atomic distances cannot be arbitrary, which we plan to improve in future work for a full solution of Problem 1.6 in the periodic case. However, these invariants already parametrize the 'universe' containing all known crystals as 'shiny stars' and all not yet discovered crystals hidden in empty spots on the same map, see Fig. 5, 6, 7.

Descriptor	Invariant	Continuity	Complete	Reconstruction	Time
primitive cell	×	×	×	×	✓
reduced cell	✓	×	×	×	✓
space group	✓	×	×	×	✓
PDF [57]	✓	✓	×	×	✓
SOAP [5]	✓	✓	×	×	✓
densities [26]	✓	✓	✓*	×	✓*
isosets [2]	✓	✓	✓	✓	✓*
AMD	✓	✓	×	×	✓
PDD	✓	✓	✓*	✓*	✓

TABLE 6

Comparison of crystal descriptors with regards to the requirements of Problem **??**. *PDD and AMD are introduced in this thesis.* ✓ * *in the 'Computable' column indicates that only an approximate algorithm exists to compute distances, and* ✓ * *in the 'Complete' and 'Reconstruction' columns means that the condition holds in* general position.

The key impact is the efficient barrier for noisy disguises of known crystals because the invariants can quickly find all nearest neighbors of any newly claimed material in the existing databases. We thank all reviewers for supporting scientific integrity, now guaranteed by the proposed invariants.

Appendix A. Detailed proofs of auxiliary results.

Proof of Lemma 4.2. Choose the origin $0 \in \mathbb{R}^n$ at a point of S. Applying translations, we can assume that primitive unit cells $U(S), U(Q)$ of the given periodic sets S, Q have a vertex at the origin 0. Then $S = \Lambda(S) + (U(S) \cap S)$ and $Q = \Lambda(Q) + (U(Q) \cap Q)$, where $\Lambda(S), \Lambda(Q)$ are lattices of S, Q, respectively.

We are given that every point of Q is $d_B(S,Q)$-close to a point of S, where the bottleneck distance $d_B(S,Q)$ is strictly less than the packing radius $r(Q)$.

Assume by contradiction that S, Q have no common lattice. Then there is a point $p \in \Lambda(S)$ whose all integer multiples $kp \in \Lambda(S)$ do not belong to $\Lambda(Q)$ for $k \in \mathbb{Z} - \{0\}$. Any such multiple kp can be translated by a vector of $\Lambda(Q)$ to a point $q(k)$ in the unit cell $U(Q)$ so that $kp \equiv q(k) \pmod{\Lambda(Q)}$. Since the cell $U(Q)$ contains infinitely many points $q(k)$, one can find a pair $q(i) \neq q(j)$ at a distance less than $\delta = r(Q) - d_B(S,Q) > 0$. For any $m \in \mathbb{Z}$, the following points are equivalent modulo (translations along the vectors of) the lattice $\Lambda(Q)$.

$$q(i + m(j-i)) \equiv (i + m(j-i))p = ip + m(jp - ip) \equiv q(i) + m(q(j) - q(i)).$$

These points for $m \in \mathbb{Z}$ lie in a straight line with gaps $|q(j) - q(i)| < \delta$. The open balls with the packing radius $r(Q)$ and centers at all points of Q do not overlap. Hence all closed balls with the radius $d_B(S,Q) < r(Q)$ and the same centers are at least 2δ away from each other. Due to $|q(j) - q(i)| < \delta = r(Q) - d_B(S,Q)$, there is $m \in \mathbb{Z}$ such that $q(i) + m(q(j) - q(i))$ is outside the union $Q + \bar{B}(0; d_B(S,Q))$ of all these smaller balls. Then $q(i) + m(q(j) - q(i))$ has a distance more than $d_B(S,Q)$ from any point of Q. The translations along all vectors of the lattice $\Lambda(Q)$ preserve the union of balls $Q + \bar{B}(0; d_B(S,Q))$. Then the point $(i + m(j-i))p \in S$, which is equivalent to $q(i) + m(q(j) - q(i))$ modulo $\Lambda(Q)$, has a distance more than $d_B(S,Q)$ from any point of Q. This conclusion contradicts the definition of $d_B(S,Q)$. ∎

Proof of Lemma 4.3. 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 $|a - a_i| - 2\varepsilon$. Because 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 $|a - a_i| + 2\varepsilon$. ∎

Proof of Lemma 6.1. Intersect the three regions $U'(p;r) \subset \bar{B}(p;r) \subset U''(p;r)$ with S in $\mathbb{R}^n$ and count resulting points: $|S \cap U'(p;r)| \leq |S \cap \bar{B}(p;r)| \leq |S \cap U''(p;r)|$.

The union $U'(p;r)$ consists of $\dfrac{\mathrm{vol}[U'(p;r)]}{\mathrm{vol}[U]}$ cells, which all have the same volume $\mathrm{vol}[U]$. Since $|S \cap U| = m$, we now get $|S \cap U'(p;r)| = m\dfrac{\mathrm{vol}[U'(p;r)]}{\mathrm{vol}[U]}$. Similarly we

count the points in the upper union: $|S \cap U''(p;r)| = m\dfrac{\mathrm{vol}[U''(p;r)]}{\mathrm{vol}[U]}$. The bounds of $|S \cap \bar{B}(p;r)|$ become

$$m\frac{\mathrm{vol}[U'(p;r)]}{\mathrm{vol}[U]} \le |S \cap \bar{B}(p;r)| \le m\frac{\mathrm{vol}[U''(p;r)]}{\mathrm{vol}[U]},$$

$$\mathrm{vol}[U'(p;r)] \le \frac{\mathrm{vol}[U]}{m}|S \cap \bar{B}(p;r)| \le \mathrm{vol}[U''(p;r)].$$

For the diameter d of the unit cell U, the smaller ball $\bar{B}(p;r-d)$ is completely contained within the lower union $U'(p;r)$. Indeed, if $|\vec{q}-\vec{p}| \le r-d$, then $q \in U+\vec{v}$ for some $\vec{v} \in \Lambda$. Then $(U+\vec{v})$ is covered by the ball $\bar{B}(q;d)$, hence by $\bar{B}(p;r)$ due to the triangle inequality. The inclusion $\bar{B}(p;r-d) \subset U'(p;r)$ implies the lower bound for the volumes:

$$V_n(r-d)^n = \mathrm{vol}[\bar{B}(p;r-d)] \le \mathrm{vol}[U'(p;r)], \text{ where}$$

V_n is the unit ball volume in $\mathbb{R}^n$. The inclusion $U''(p;r) \subset \bar{B}(p;r+d)$ gives

$$\mathrm{vol}[U''(p;r)] \le \mathrm{vol}[\bar{B}(p;r+d)] = V_n(r+d)^n,$$

$$V_n(r-d)^n \le \frac{\mathrm{vol}[U]}{m}|S \cap B(p;r)| \le V_n(r+d)^n,$$

$\dfrac{mV_n}{\mathrm{vol}[U]}(r-d)^n \le |S \cap B(p;r)| \le \dfrac{mV_n}{\mathrm{vol}[U]}(r+d)^n$, which implies the result. ∎

Proof of Lemma 3.5. The closed ball $\bar{B}(p;r)$ of the radius $r = d_k(S;p)$ has more than k points (including p) from S. The upper bound of Lemma 6.1 for $r = d_k(S;p)$ implies that $k < |S \cap \bar{B}(p;r)| \le \dfrac{(r+d)^n}{(c(S))^n}$. Taking the n-th roots, we get $\sqrt[n]{k} < \dfrac{r+d}{c(S)}$, so $r = d_k(S;p) > c(S)\sqrt[n]{k} - d$.

For any smaller radius $r < d_k(S;p)$, the closed ball $\bar{B}(p;r)$ contains at most k points (including p) from S. The lower bound of Lemma 6.1 for $r < d_k(S;p)$ implies that $\dfrac{(r-d)^n}{c(S)^n} \le |S \cap \bar{B}(p;r)| \le k$. Since $\dfrac{(r-d)^n}{c(S)^n} \le k$ holds for the constant upper bound k and any radius $r < d_k(S;p)$, the same inequality holds for the radius $r = d_k(S;p)$. Similarly to the upper bound, we get $\dfrac{r-d}{c(S)} \le \sqrt[n]{k}$, $r = d_k(S;p) \le c(S)\sqrt[n]{k} + d$. Combine the two bounds above as follows: $c(S)\sqrt[n]{k} - d < d_k(S;p) \le c(S)\sqrt[n]{k} + d$. ∎

Appendix B. Examples and instructions for the PDD code and data.

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 $\mathrm{PDD}(S; k)$.

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 $\mathrm{PDD}(S; k)$ 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. **??**, along with their Earth mover's distance.
- `1D_sets_example.py` shows that the 1D periodic sets in Fig. **??** 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. **??**, 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 **??**, 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.

Appendix C. Details of experiments on the largest databases.

This appendix describes the main experiments in more detail. Some entries in the CSD and COD are incomplete or disordered (not periodic). After removing such

entries, we were left with 831,126 CSD structures and 344,127 COD structures.

Firstly, we computed PDM[10](S; 100) for all entries, taking 27 min 33 sec for the CSD and 12 mins 15 sec for COD (2 ms per structure on average). To find exact matches between databases by PDM, we make use of the k-d tree data structure, designed for fast nearest neighbor lookup. A k-d tree can be constructed from any collection of vectors, which can then be queried for a number of nearest neighbors of a new vector, using a binary tree style algorithm with logarithmic search time.

We flattened each PDM[10](S; 100) matrix to a vector with 1000 dimensions, constructed a k-d tree for both CSD and COD, then queried the 10 nearest neighbors for each item in the other. If the most distant neighbor for any entry is closer than the threshold 10^{-13}Å (within floating point error), we extend the search and find more neighbors until all pairs within the threshold are found. We were left with a total of 270,669 matches; an overlap between the databases of one third of the CSD and almost 80% of COD.

CSD refcode	COD ID	Notes
LAVFAP	2001334	Mixed types in original CIF
ZAYRUM	2003941	Mixed types in original CIF
FONGAQ01	2005101	Mixed types in original CIF
TIPYOG	2005914	Mixed types in original CIF
HABTAF	2001740	Mixed types in original CIF
AJIRAM01	2100097	Mixed types in original CIF
LABSAI	2001822	Mixed types in COD CIF
DECTAI	4065524	Mixed types in COD CIF
WATMIO	4309447	Mixed types in COD CIF
NAJQUK	4323901	Mixed types in COD CIF
PIHJUL	4030494	Mixed types in COD CIF
ELOJOE	4314231	CSD remarks replaced atom
MARSIH	4321045	CSD remarks replaced atom
KUTWUU	7126770	CSD remarks replaced atom
XAVDEF	4103386	CSD remarks replaced atom
JEMLAP	4101489	CSD remarks replaced atom
QUCXAP	7117360	CSD remarks replaced atom
PIBTAW	1505325	CSD remarks replaced atom
UKAXUB	7234657	CSD remarks replaced atom
POCLOK	2220314	COLYEI is a duplicate
COLYEI	8102533	POCLOK is a duplicate
JEPLIA	2213484	HIFCAB is a duplicate
LALNET	8102594	POPCAA is a duplicate
SELHAU	4027023	One entry is mistaken
PINHUP	1558382	One entry is mistaken
KABHOL	4113866	One entry is mistaken

TABLE 7

List of 26 matches between the CSD and COD found to have identical geometry but different chemical compositions.

Of particular interest are the 26 pairs which have different compositions, as the impossibility of complex organic structures sharing the exact same geometry but not composition implies an error or labeling issue. The pairs were confirmed as geometric

duplicates by checking their CIFs and found to have different compositions for the reasons in Table 7.

- The original Crystallographic Information File (CIF) has atoms simultaneously labeled as two types or disagreement with what is reported in the published paper (6 pairs),
- Atoms are labeled as two types in the COD CIF (5 pairs),
- Geometric duplicates known to the CSD gave a match with different compositions (4 pairs),
- A remark in the CSD entry explains that atoms were replaced in the curation process because the deposited CIF was incorrect (8 pairs),
- The COD and CSD entries disagree for an unknown reason (3 pairs).

In addition to cross-comparing the CSD and COD, we included the ICSD and Materials Project database (MP) and compared them all pairwise, as well as searching for duplicates within each. Tables 8 and **??** below show how many matches were found, and how many also shared the same composition.

databases	matches	same composition
CSD vs COD	270,669	270,583
CSD vs ICSD	3,913	3,913
COD vs ICSD	35,051	31,918
COD vs MP	2	2
ICSD vs MP	17	7

TABLE 8
Number of exact matches (PDM within 10^{-13}Å) between four databases.

Acknowledgments. The authors thank Andy Cooper FRS (the director of the Materials Innovation Factory) as the PhD co-supervisor of the first author, and the group in Data Science Theory and Applications led by the second author.

REFERENCES

[1] H. Alt, K. Mehlhorn, H. Wagener, and E. Welzl, *Congruence, similarity, and symmetries of geometric objects*, Discrete and Computational Geometry, 3 (1988), pp. 237–256.
[2] O. Anosova and V. Kurlin, *An isometry classification of periodic point sets*, in Proceedings of Discrete Geometry and Mathematical Morphology, 2021, pp. 229–241.
[3] O. Anosova, V. Kurlin, and M. Senechal, *The importance of definitions in crystallography*, IUCrJ (the International Union of Crystallography Journal), 11 (2024), https://doi.org/10.1107/S2052252524004056.
[4] V. Arvind and G. Rattan, *The parameterized complexity of geometric graph isomorphism*, Algorithmica, 75 (2016), pp. 258–276.
[5] A. P. Bartók, R. Kondor, and G. Csányi, *On representing chemical environments*, Physical Review B—Condensed Matter and Materials Physics, 87 (2013), p. 184115.
[6] M. Boutin and G. Kemper, *On reconstructing n-point configurations from the distribution of distances or areas*, Advances in Applied Mathematics, 32 (2004), pp. 709–735.
[7] P. Brass and C. Knauer, *Testing the congruence of d-dimensional point sets*, in Proceedings of SoCG, 2000, pp. 310–314.
[8] P. Brass and C. Knauer, *Testing congruence and symmetry for general 3-dimensional objects*, Computational Geometry, 27 (2004), pp. 3–11.
[9] M. Bright, A. Cooper, and V. Kurlin, *Geographic-style maps for 2-dimensional lattices*, Acta Cryst A, 79 (2023), pp. 1–13.
[10] M. J. Bright, A. I. Cooper, and V. A. Kurlin, *Continuous chiral distances for 2-dimensional lattices*, Chirality, 35 (2023), pp. 920–936.
[11] H.-G. Carstens et al., *Geometrical bijections in discrete lattices*, Combinatorics, Probability and Computing, 8 (1999), pp. 109–129.

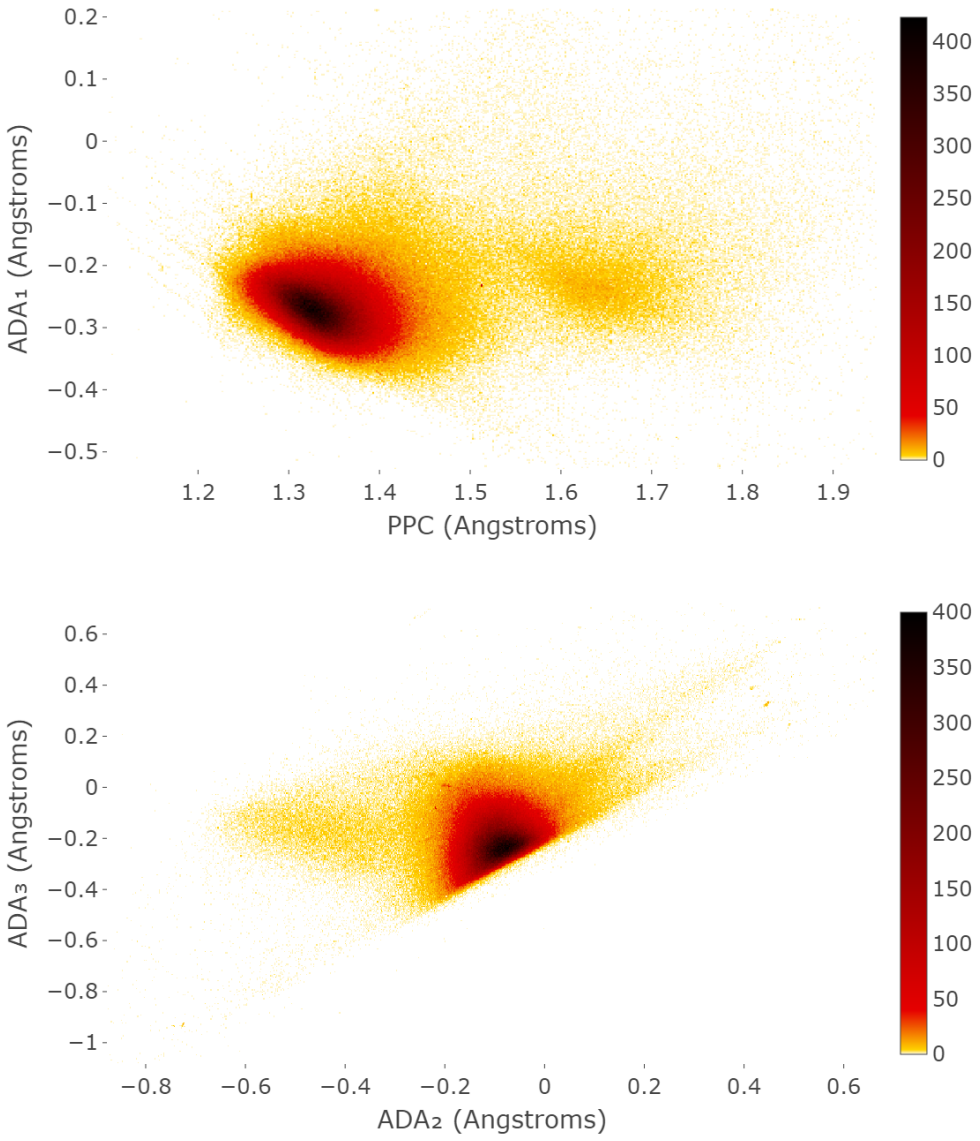


FIG. 5. *The projections of the CSD in the invariants* $\mathrm{PPC}, \mathrm{ADA}_1, \mathrm{ADA}_2, \mathrm{ADA}_3$.

[12] D. S. CHAWLA, *Crystallography databases hunt for fraudulent structures*. https://cen.acs.org/research-integrity/Crystallography-databases-hunt-fraudulent-structures/102/i8, 2024.
[13] P. CHEW, D. DOR, A. EFRAT, AND K. KEDEM, *Geometric pattern matching in d-dimensional space*, Discrete Comp. Geometry, 21 (1999), pp. 257–274.
[14] P. CHEW, M. GOODRICH, D. HUTTENLOCHER, K. KEDEM, J. KLEINBERG, AND D. KRAVETS, *Geometric pattern matching under Euclidean motion*, Computational Geometry, 7 (1997), pp. 113–124.
[15] P. CHEW AND K. KEDEM, *Improvements on geometric pattern matching problems*, in Scandinavian Workshop on Algorithm Theory, 1992, pp. 318–325.
[16] J. CHISHOLM AND S. MOTHERWELL, *Compack: a program for identifying crystal structure sim-*

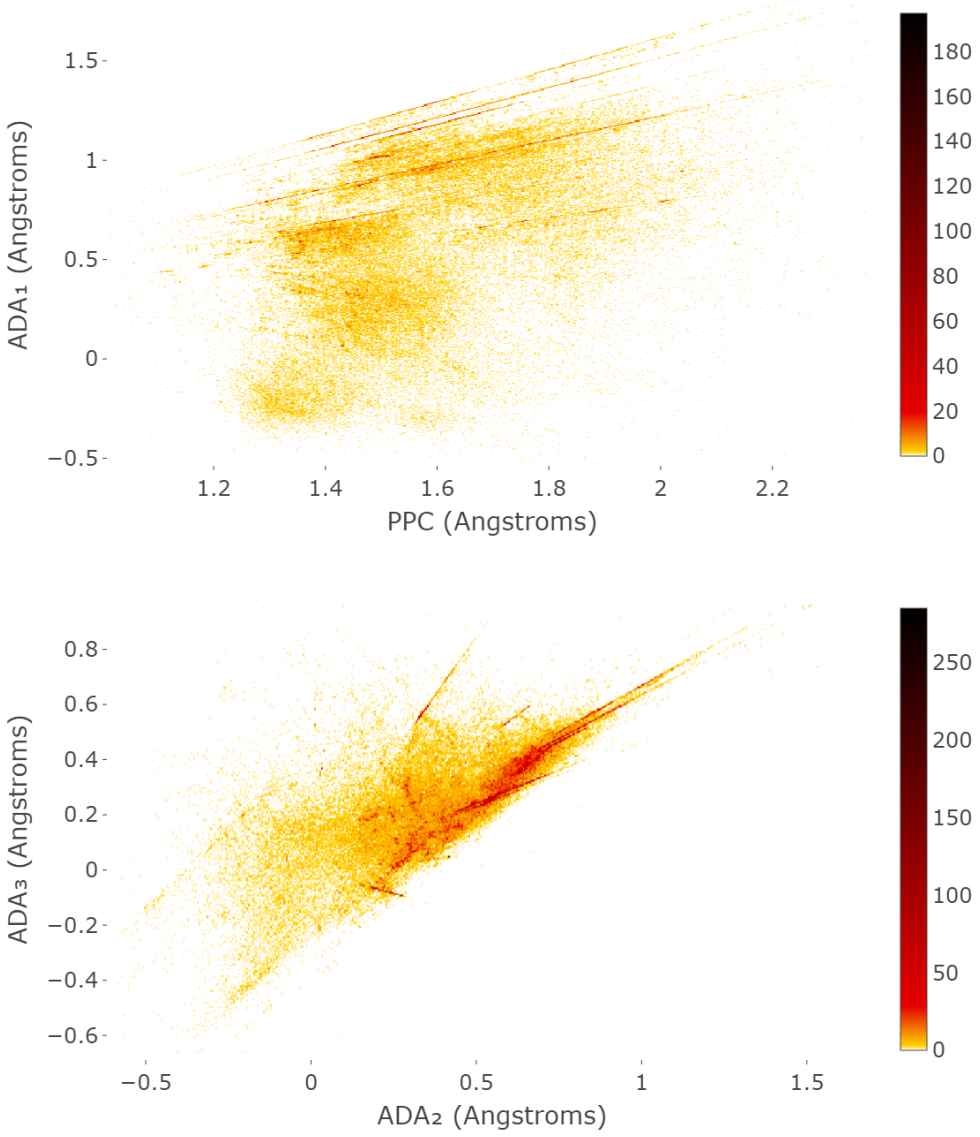


FIG. 6. *The projections of the ICSD in the invariants* $\mathrm{PPC}, \mathrm{ADA}_1, \mathrm{ADA}_2, \mathrm{ADA}_3$.

ilarity using distances, J. Applied Crystal., 38 (2005), pp. 228–231.

[17] S. COHEN AND L. GUIBAS, *The earth mover's distance: Lower bounds and invariance under translation*, tech. 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 (1992), pp. 55–68.

[19] L. COSMO, M. PANINE, A. RAMPINI, M. OVSJANIKOV, M. M. BRONSTEIN, AND E. RODOLA, *Isospectralization, or how to hear shape, style, and correspondence*, in Proceedings of CVPR, 2019, pp. 7529–7538.

[20] B. N. DELONE, N. P. DOLBILIN, M. I. SHTOGRIN, AND R. V. GALIULIN, *A local criterion for regularity of a system of points*, in Dokl. Akad. Nauk SSSR, vol. 227, 1976, pp. 19–21.

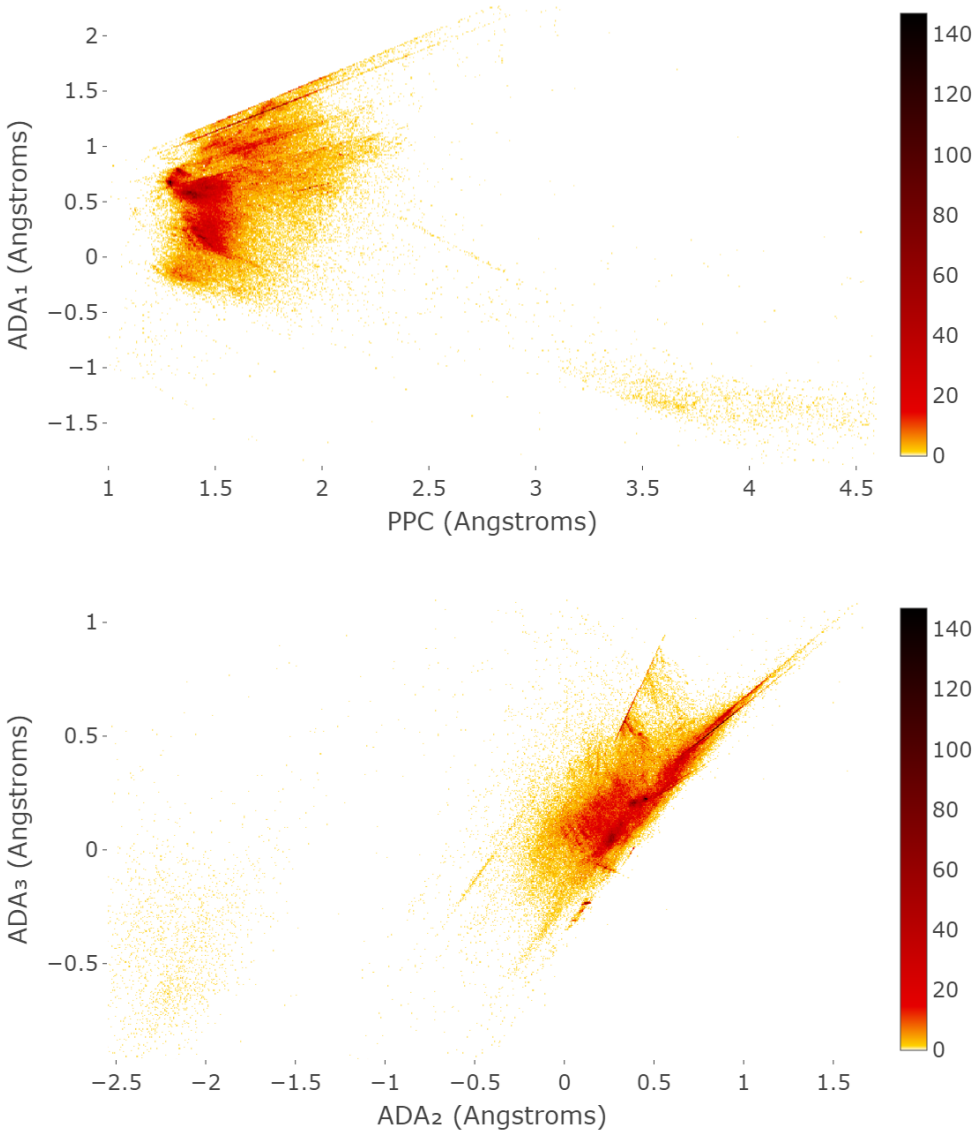


FIG. 7. *The projections of the MP in the invariants* $\mathrm{PPC}, \mathrm{ADA}_1, \mathrm{ADA}_2, \mathrm{ADA}_3$.

[21] N. DOLBILIN, J. LAGARIAS, AND M. SENECHAL, *Multiregular point systems*, Discrete & Computational Geometry, 20 (1998), pp. 477–498.
[22] M. DUNEAU AND C. OGUEY, *Bounded interpolations between lattices*, Journal of Physics A: Mathematical and General, 24 (1991), p. 461.
[23] H. EDELSBRUNNER AND R. SEIDEL, *Voronoi diagrams and arrangements*, Discrete & Computational Geometry, 1 (1986), pp. 25–44.
[24] Y. ELKIN AND V. KURLIN, *Counterexamples expose gaps in the proof of time complexity for cover trees introduced in 2006*, in Topological Data Analysis and Visualization (TopoInVis), 2022, pp. 9–17.
[25] Y. ELKIN AND V. KURLIN, *A new near-linear time algorithm for k-nearest neighbor search*

using a compressed cover tree, in International Conference on Machine Learning (ICML), 2023, pp. 9267–9311.
[26] H. E. et al, *The density fingerprint of a periodic point set*, in Proceedings of SoCG, 2021.
[27] R. S. et al, *Fast and robust comparison of probability measures in heterogeneous spaces*, arXiv:2002.01615, (2020).
[28] F. Gieseke, J. Heinermann, C. Oancea, and C. Igel, *Buffer kd trees: processing massive nearest neighbor queries on gpus*, in International Conference on Machine Learning, PMLR, 2014, pp. 172–180.
[29] M. T. Goodrich, J. S. Mitchell, and M. W. Orletsky, *Approximate geometric pattern matching under rigid motions*, Transactions on Pattern Analysis and Machine Intelligence, 21 (1999), pp. 371–379.
[30] C. Gordon, D. Webb, and S. Wolpert, *Isospectral plane domains and surfaces via riemannian orbifolds*, Inventiones mathematicae, 110 (1992), pp. 1–22.
[31] C. Gordon, D. L. Webb, and S. Wolpert, *One cannot hear the shape of a drum*, Bulletin of the American Mathematical Society, 27 (1992), pp. 134–138.
[32] S. Gražulis, D. Chateigner, R. T. Downs, A. Yokochi, M. Quirós, L. Lutterotti, E. Manakova, J. Butkus, P. Moeck, and A. Le Bail, *Crystallography open database–an open-access collection of crystal structures*, Journal of applied crystallography, 42 (2009), pp. 726–729.
[33] F. Hausdorff, *Dimension und äußeres maß*, Mathematische Annalen, 79 (1919), pp. 157–179.
[34] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, *Comparing images using the Hausdorff distance*, Transactions on pattern analysis and machine intelligence, 15 (1993), pp. 850–863.
[35] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, et al., *Commentary: The materials project: A materials genome approach to accelerating materials innovation*, APL materials, 1 (2013).
[36] M. Kac, *Can one hear the shape of a drum?*, The american mathematical monthly, 73 (1966), pp. 1–23.
[37] E. S. Keeping, *Introduction to statistical inference*, Courier Corporation, 1995.
[38] D. G. Kendall, D. Barden, T. K. Carne, and H. Le, *Shape and shape theory*, John Wiley & Sons, 2009.
[39] J. B. Kruskal and M. Wish, *Multidimensional scaling*, no. 11, Sage, 1978.
[40] V. Kurlin, *A complete isometry classification of 3-dimensional lattices*, arxiv:2201.10543, (2022).
[41] V. A. Kurlin, *Mathematics of 2-dimensional lattices*, Foundations of Computational Mathematics, 24 (2024), p. 805–863, https://doi.org/10.1007/s10208-022-09601-8.
[42] M. Laczkovich, *Uniformly spread discrete sets in R^d*, Journal of the London Mathematical Society, 2 (1992), pp. 39–57.
[43] S. Lawton and R. Jacobson, *The reduced cell and its crystallographic applications*, tech. report, Ames Lab, Iowa State University, 1965.
[44] I. G. Macdonald, *Symmetric functions and Hall polynomials*, Oxford University Press, 1998.
[45] S. Majhi, J. Vitter, and C. Wenk, *Approximating gromov-hausdorff distance in euclidean space*, Computational Geometry, 116 (2024), p. 102034.
[46] R. Marin, A. Rampini, U. Castellani, E. Rodolà, M. Ovsjanikov, and S. Melzi, *Spectral shape recovery and analysis via data-driven connections*, International journal of computer vision, 129 (2021), pp. 2745–2760.
[47] F. Mémoli, *Gromov–Wasserstein distances and the metric approach to object matching*, Foundations of Computational Mathematics, 11 (2011), pp. 417–487.
[48] M. Mosca and V. Kurlin, *Voronoi-based similarity distances between arbitrary crystal lattices*, Crystal Research and Technology, 55 (2020), p. 1900197.
[49] S. Rass, S. König, S. Ahmad, and M. Goman, *Metricizing the euclidean space towards desired distance relations in point clouds*, IEEE Transactions on Information Forensics and Security, (2024).
[50] M. Reuter, F.-E. Wolter, and N. Peinecke, *Laplace–beltrami spectra as 'shape-dna'of surfaces and solids*, Computer-Aided Design, 38 (2006), pp. 342–366.
[51] Y. Rubner, C. Tomasi, and L. Guibas, *The earth mover's distance as a metric for image retrieval*, Int. J Computer Vision, 40 (2000), pp. 99–121.
[52] F. Schmiedl, *Computational aspects of the Gromov–Hausdorff distance and its application in non-rigid shape matching*, Discrete and Computational Geometry, 57 (2017), pp. 854–880.
[53] I. Schoenberg, *Remarks to Maurice Frechet's article "Sur la definition axiomatique d'une classe d'espace distances vectoriellement applicable sur l'espace de Hilbert*, Annals of Mathematics, (1935), pp. 724–732.

[54] M. Senechal, *Quasicrystals and geometry*, CUP Archive, 1996.
[55] S. Shirdhonkar and D. Jacobs, *Approximate earth mover's distance in linear time*, in Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[56] R. Taylor and P. A. Wood, *A million crystal structures: The whole is greater than the sum of its parts*, Chemical reviews, 119 (2019), pp. 9427–9477.
[57] M. W. Terban and S. J. Billinge, *Structural analysis of molecular materials using the pair distribution function*, Chemical Reviews, 122 (2021), pp. 1208–1272.
[58] S. Villar, D. W. Hogg, K. Storey-Fisher, W. Yao, and B. Blum-Smith, *Scalars are universal: equivariant machine learning, structured like classical physics*, Advances in Neural Information Processing Systems, 34 (2021), pp. 28848–28863.
[59] H. Weyl, *The classical groups: their invariants and representations*, no. 1, Princeton university press, 1946.
[60] D. Widdowson and V. Kurlin, *Resolving the data ambiguity for periodic crystals*, Advances in Neural Information Processing Systems, 35 (2022), pp. 24625–24638.
[61] D. Widdowson, M. M. Mosca, A. Pulido, A. I. Cooper, and V. Kurlin, *Average minimum distances of periodic point sets - foundational invariants for mapping all periodic crystals*, MATCH Commun. Math. Comput. Chem., 87 (2022), pp. 529–559.
[62] D. E. Widdowson and V. A. Kurlin, *Recognizing rigid patterns of unlabeled point clouds by complete and continuous isometry invariants with no false negatives and no false positives*, in Computer Vision and Pattern Recognition, 2023, pp. 1275–1284.
[63] D. Zagorac, H. Müller, S. Ruehl, J. Zagorac, and S. Rehme, *Recent developments in the inorganic crystal structure database: theoretical crystal structure data and related features*, Journal of applied crystallography, 52 (2019), pp. 918–925.