POINTWISE DISTANCE DISTRIBUTIONS FOR DETECTING NEAR-DUPLICATES IN LARGE MATERIALS DATABASES*

DANIEL E. WIDDOWSON† AND VITALIY A. KURLIN‡

Abstract. Many real objects are often given as discrete sets of points such as corners or other salient features. For our main applications in chemistry, points represent atomic centers in a molecule or a solid material. 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.

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 the Pointwise Distance Distribution (PDD) for any discrete set and prove 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 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. 74E15, 68U05, 51N20

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

This paper is a substantial extension of the 10-page conference version at NeurIPS 2022 [63]. 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 set in a metric space and rigorously proves the properties above in finite and periodic cases. We also adapt the invariants to 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 continuous and generically complete invariants are motivated by the previously unresolved ambiguity of digital representations of molecules and crystals in terms of atomic coordinates or lattice bases. Fig. 1 (middle) shows that the same periodic set can be obtained by periodically repeating different motifs of points.

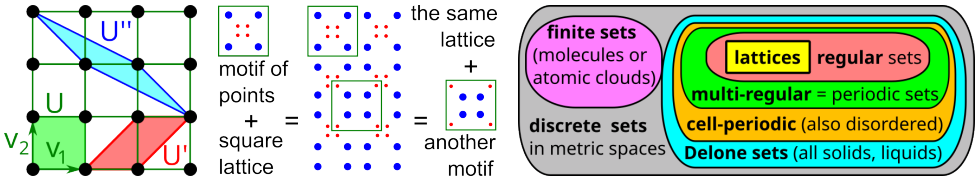

FIG. 1. ***Left**: a lattice can be defined by many primitive bases. **Middle**: a periodic set can be defined by different pairs (basis, motif). **Right**: a hierarchy of discrete sets, which model periodic crystals and amorphous solids with points at atomic centers, see Definitions 1.1, 1.2, 1.5, 3.3.*

*LaTeX2e Standard Macros were used from https://epubs.siam.org/journal-authors#macros

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).

The crucial question "same or different?" was explicitly raised for crystals [54] and makes sense for many other real objects. For a cloud of unordered points in computer vision or chemistry applications, a list of atomic coordinates depends on a given coordinate system and an order of atoms. The independence of coordinate representations is important for identifying rigid structures and rigid conformations of flexible molecules such as proteins whose properties depend on a rigid shape.

Noisy measurements imply that any real objects are at least slightly different. Hence the next practical question is "how much different?" If noise is ignored up to any positive threshold, noisy perturbations of atomic centers can be continued sufficiently long to make any given sets identical. This *sorites paradox* [33] can be resolved by quantifying even tiny differences through a continuous distance metric.

DEFINITION 1.1 (a *discrete set* S in a *metric space* X with a *metric* d_X). *A* metric space *is any set X of objects (called* points*) 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 any points* $a, b, c \in X$. *A set* $S \subset X$ *is called* discrete *if there is a constant* $\varepsilon > 0$ *such that all points of S are ε-separated, so* $d_X(a,b) \geq \varepsilon$ *for any* $a, b \in S$.

An example of a discrete set S is a finite set in $\mathbb{R}^n$ with the Euclidean metric denoted by $|\vec{p} - \vec{q}|$ for any points $p, q \in \mathbb{R}^n$. Here $\vec{p}$ denotes the vector from the origin $0 \in \mathbb{R}^n$ to p. The positivity $d_X(a,b) \geq 0$ follows from other axioms: $2d_X(a,b) = d_X(a,b) + d_X(b,a) \geq d_X(a,a) = 0$. Without the first axiom, d is called a *pseudo-metric* and can be the zero function: $d_X(a,b) = 0$ for all a, b. If the triangle inequality is allowed to fail with any additive error $\varepsilon > 0$, the results of clustering such as k-means and DBSCAN can be predetermined and hence may not be trustworthy [51].

DEFINITION 1.2 (lattice, unit cell, motif, l-periodic set). *Vectors* $\vec{v}_1, \dots, \vec{v}_n \in \mathbb{R}^n$ *form a* basis *if any vector in* $\mathbb{R}^n$ *can be written as* $\vec{v} = \sum_{i=1}^{n} t_i \vec{v}_i$ *for unique* $t_1, \dots, t_n \in \mathbb{R}$. *For any* $1 \leq l \leq n$, *the first l vectors define the* lattice $\Lambda = \{\sum_{i=1}^{l} c_i \vec{v}_i \mid c_1, \dots, c_l \in \mathbb{Z}\}$ *and the* unit cell $U = \{\sum_{i=1}^{n} t_i \vec{v}_i \mid t_1, \dots, t_l \in [0,1), t_{l+1}, \dots, t_n \in \mathbb{R}\} \subset \mathbb{R}^n$. *If* $l = n$, *then U is an n-dimensional parallelepiped. If* $l < n$, *then U is an infinite slab over an l-dimensional parallelepiped on* $\vec{v}_1, \dots, \vec{v}_l$. *For any finite set of points (called a* motif*)* $M \subset U$, *the sum* $S = M + \Lambda = \{\vec{p} + \vec{v} \mid p \in M, v \in \Lambda\}$ *is an* l-periodic point set.

Any unit cell U includes only a partial boundary: we exclude the points with any coefficient $t_i = 1$, $i = 1, \dots, l$, for convenience. Then $\mathbb{R}^n$ for $l = n$ is tiled by the shifted cells $\{U + \vec{v} \mid \vec{v} \in \Lambda\}$ 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 $\bigcup_{p \in M} (\vec{p} + \Lambda)$ of lattices whose origins are shifted to all $p \in M = S \cap U$.

If we double a unit cell in one direction, e.g. by taking the basis $2\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n$, the doubled motif $M \cup (M + \vec{v}_1)$ 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 $S \cap U$ has the smallest size among all unit cells U 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 sets, which differ only by translation.

Finite and periodic point sets represent molecules and periodic crystals at the atomic scale by considering zero-sized points at all atomic centers. 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. So these bonds are not real sticks and only abstractly represent interatomic interactions, while atomic nuclei are real objects. We model all materials at the fundamental level of atoms, which will suffice for all real materials. Because any object 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 $\mathbb{R}^n$, any isometry decomposes into translations, rotations, and reflections, which generate the Euclidean group* $\mathrm{E}(n)$*. If reflections are excluded, orientation-preserving isometries are also called* rigid motions *and form group* $\mathrm{SE}(n)$*.*

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 gives 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 by B. Delone [19] 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 one 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 in 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 [25]. 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 dense gases.

All m-regular sets for $m > 1$ 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 $\vec{q}-\vec{p}$. 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 $\mathbb{R}^2$ modeling graphene is regular, but not a lattice because there are two points in a primitive unit cell. 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. [20, 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 magenta 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 $\frac{1}{m}\sum_{p\in S} p$ is not, so the center of mass cannot reliably distinguish rigid shapes 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 equivalences, e.g. rigid motions, affine or projective 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 any 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 number m of points in a motif M from Definition 1.2 because a Crystallographic Information File (CIF) specifying a basis and atomic coordinates in this basis has a linear length $O(m)$ in the motif size m. Some infinite Delone sets can described in a finite form, e.g. some aperiodic crystals [58] 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.

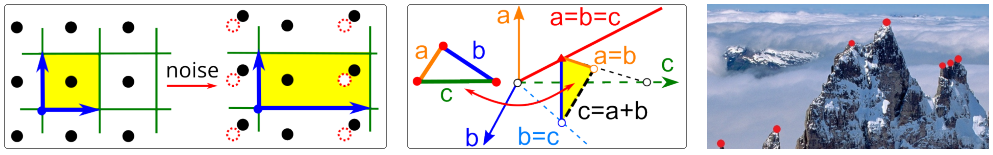

FIG. 2. ***Left****: the symmetry group and a reduced cell discontinuously change under tiny noise.* ***Middle****: the space of 3 points under isometry is parametrized by inter-point distances $0 < a \le b \le c \le a+b$.* ***Right****: energy landscapes of crystals show optimized structures as isolated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

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 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 changing the space group and even the primitive cell volume, and also by replacing some chemical elements to avoid detection by chemical composition [3, section 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 asymptotic for l-periodic sets, rigorous proofs of the Lipschitz continuity (also for adjusted and normalized invariants), 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 millions of artificial structures were claimed 'new' without checking for duplication with known crystals. The decisive advance is closing this discontinuity loophole in crystallography, which is demonstrated for 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 [63, 66]. Finite sets have two subcases: ordered points (easy) and unordered (much harder).

Ordered finite sets. Kendall's shape theory [37] studies ordered points $p_1, \dots, p_m \in \mathbb{R}^n$ whose complete isometry invariant is the distance matrix [57, 38] or the Gram matrix of scalar products $\vec{p}_i \cdot \vec{p}_j$ [62, chapter 2.9], [61]. A brute-force extension to m unordered points requires $m!$ matrices due to $m!$ permutations ruled out by (1.6d).

Unordered finite sets (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 [7]. The genericity condition allows almost all clouds apart from 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.

Problem 1.6 expands the question 'Can we hear the shape of a drum?' [35] which has the negative answer in terms of 2D polygons that are indistinguishable by spectral invariants [28, 29, 52, 17, 47]. Problem 1.6 looks for stronger invariants that can completely 'sense' as in (1.6b), not only 'hear', the rigid shape of any cloud.

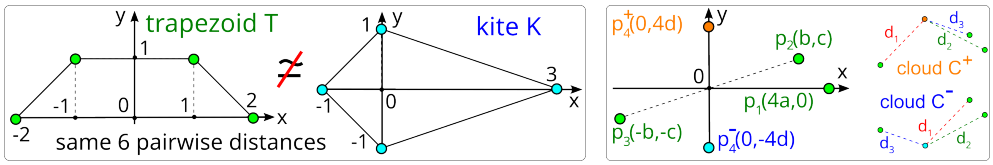

FIG. 3. *Non-isometric clouds of 4 points with the same 6 pairwise distances.* ***Left****: the trapezoid T has points $(\pm 2, 1)$, $(\pm 4, -1)$. The kite K has $(5, 0)$, $(-3, 0)$, $(-1, \pm 2)$.* ***Right****: the infinite family of non-isometric clouds $C^+ \not\cong C^-$ sharing p_1, p_2, p_3 and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

Computational geometry studied earlier versions of Problem 1.6 by developing canonical representations of point clouds [2, 8, 4], which can be considered complete invariants, and also metrics between isometry classes of clouds. For example, any metric between fixed clouds extends to their isometry classes [32, 14, 13] by minimization over infinitely many transformations from the group $\mathrm{E}(n)$. This extension of the Hausdorff distance [31] for m-point clouds in $\mathbb{R}^2$ has time $O(m^5 \log m)$, see [27]. The Gromov-Wasserstein metrics [48, 49] are defined for any 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 [56] and polynomial algorithms for partial cases in [1, 44, 46].

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 full planet and hence a realizable continuous invariant I, which can be considered an analog of coordinates of latitude and longitude.

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 [50]. 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 [39].

For general periodic point sets, the latest advance announced in [63] 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 [48] without studying the conditions of Problem 1.6. For finite clouds in $\mathbb{R}^n$, the complete invariants under rigid motion with Lipschitz continuous metrics were developed in [66, 40]. The high polynomial-time complexity of these latest invariants motivates using the much faster PDD in practice.

3. The Pointwise Distance Distribution and other isometry invariants.

This section introduces the Pointwise Distance Distribution (PDD) for any discrete set S with a finite subset M 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 $D(S, M; k)$ 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 $\mathrm{PDD}(S,M;k)$. *The* Average Minimum Distance AMD_i *is the weighted average of the i-th column in* $\mathrm{PDD}(S,M;k)$ *for each* $i=1,\dots,k$. *Let* $\mathrm{AMD}(S,M;k)$ *denote the vector* $(\mathrm{AMD}_1,\dots,\mathrm{AMD}_k)$.

Definition 3.1 introduced the isometry invariant $\mathrm{PDD}(S,M;k)$ of a pair (S,M) for a finite subset M in any Delone set S. For any l-periodic point set $S\subset\mathbb{R}^n$, Theorem 3.1 will prove that PDD is independent of a motif $M\subset S$. We use the simpler notations $\mathrm{PDD}(S;k),\mathrm{AMD}(S;k)$ in the finite ($S=M$) and periodic cases.

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 by* PDD*s even for* $k=1$.

TABLE 1
Each point of $T,K\subset\mathbb{R}^2$ in Figure 3 (left) has distances to other points in increasing order. After keeping only distances (not neighbors), the resulting PDD*s distinguish $T\not\cong K$, see Example 3.2.*

points of T	dist. to neighbor 1	dist. to neighbor 2	dist. to 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 $(-2,0)$
points of K	dist. to neighbor 1	dist. to neighbor 2	dist. to 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 $(3,0)$

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

Theorem 3.1 extends [63, Theorem 3.2], which was stated for n-periodic sets without proof, to all finite sets, l-periodic sets, and pairs (S,M) from Definition 3.1.

THEOREM 3.1 (invariance of PDD). ***(a)*** *Any isometry $S\to Q$ mapping a finite subset $M\subset S$ of m points to $N\subset Q$, we have* $\mathrm{PDD}(S,M;k)=\mathrm{PDD}(Q,N;k)$ *and* $\mathrm{AMD}(S,M;k)=\mathrm{AMD}(Q,N;k)$ *for any* $1\le k<m$. *Hence, if $S=M$ is a finite space, then* $\mathrm{PDD}(S;k)$ *and* $\mathrm{AMD}(S;k)$ *are well-defined isometry invariants of S.*
(b) *For any l-periodic point set $S\subset\mathbb{R}^n$, where* $1\le l\le n$, $\mathrm{PDD}(S;k)$ *and* $\mathrm{AMD}(S;k)$ *are isometry invariants of S (independent of a motif $M\subset S$) for any* $k\ge 1$.

Proof. **(a)** For any sets $M\subset S$ and their isometric images $N\subset Q$, the invariance follows from the fact that any isometry preserves all inter-point distances.

(b) For any l-periodic point set $S=\Lambda+M\subset\mathbb{R}^n$, we first show that scaling up a cell U and hence the motif $M=S\cap U$ of m points keeps PDD invariant. For any integer $b\ge 1$, a matrix $B\in\mathrm{GL}(l;\mathbb{Z})$ with $|\det B|=b$ acts on the first l vectors $\vec{v}_1,\dots,\vec{v}_l$ that generate the l-dimensional base parallelepiped P of U in Definition 1.2.

Let $B(U)\subset\mathbb{R}^n$ denote the cell obtained from U by applying B to P and keeping all other basis vectors $v_{l+1},\dots,v_n$ fixed. Then $D(S,S\cap B(U);k)$ from Definition 3.1

has the larger size $bm \times k$ but (due to periodicity of S) splits into m blocks, each corresponding to b points of the scaled motif $S \cap B(U)$ that are obtained from a single point $p \in M$ by translations by vectors of Λ. Since translations preserve distances, each of m blocks has b identical rows of distances to k neighbors in S, the same as in $D(S, M; k)$. Then $\mathrm{PDD}(S, S \cap B(U); k) = \mathrm{PDD}(S, M; k)$ due to collapsing of identical rows in Definition 3.1. So $\mathrm{PDD}(S; k)$ is independent of any motif $M = S \cap U$.

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. □

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 $\mathrm{PDD}(S; k)$ but all previous distances remain the same. Definition 3.3 will help describe the asymptotic of $\mathrm{PDD}(S; k)$ as $k \to +\infty$ in Theorem 3.6, which uses Lemma 3.4 extending [65, Lemma 11] to l-periodic sets $S \subset \mathbb{R}^n$ for any $1 \le l \le n$, see all skipped proofs in SM3.

Definition 3.3 (Point Packing Coefficient PPC of a cell-periodic set S). *For $1 \le l \le n$ and a basis $\vec{v}_1, \dots, \vec{v}_n \in \mathbb{R}^n$, consider the lattice the* lattice $\Lambda = \{\sum_{i=1}^{l} c_i \vec{v}_i \mid c_1, \dots, c_l \in \mathbb{Z}\}$ *and the unit cell $U = \{\sum_{i=1}^{n} t_i \vec{v}_i \mid t_1, \dots, t_l \in [0,1), t_{l+1}, \dots, t_n \in \mathbb{R}\}$. A discrete set $S \subset \mathbb{R}^n$ is* cell-periodic *if S has a fixed number m points in every shifted cell $U + \vec{v}$ for all $\vec{v} \in \Lambda$. If $l < n$, let $R^l \subset \mathbb{R}^n$ be the subspace spanned by $\vec{v}_1, \dots, \vec{v}_l$, then U is an infinite slab based on the l-dimensional parallelepiped of volume* $\mathrm{vol}[U \cap R^l]$ *. The volume of the unit ball in $\mathbb{R}^l$ is $V_l = \dfrac{\pi^{n/2}}{\Gamma(\frac{l}{2}+1)}$, where Euler's Gamma function [18] is $\Gamma(m) = (m-1)!$ and $\Gamma(\frac{m}{2}+1) = \sqrt{\pi}(m-\frac{1}{2})(m-\frac{3}{2})\cdots\frac{1}{2}$ for any integer $m \ge 1$. Define the* Point Packing Coefficient *of S as* $\mathrm{PPC}(S) = \sqrt[l]{\dfrac{\mathrm{vol}[U \cap R^l]}{mV_l}}$.

Any l-periodic set is cell-periodic, but all cell-periodic sets form a wider collection of Delone sets and model disordered solid materials that can have an underlying lattice with atoms at different positions in periodically translated cells $U + \vec{v}$, see Fig. 1.

Lemma 3.4 (bounds on points within a cylinder). *For any $1 \le l \le n$ and a basis $\vec{v}_1, \dots, \vec{v}_n \in \mathbb{R}^n$, let $S \subset \mathbb{R}^n$ be a cell-periodic set with a unit cell U based on the l-dimensional parallelepiped $U \cap R^l$, where $R^l \subset \mathbb{R}^n$ is spanned by $\vec{v}_1, \dots, \vec{v}_l$. Define the* width w *of U as* $\sup_{u,v \in U \cap R^l} |\vec{u} - \vec{v}|$. *For any point $p \in S \cap U$ and a radius r, consider*

the cylinder $C(p;r) = \{\sum_{i=1}^{n} t_i\vec{v}_i$ *such that* $t_1,\dots,t_n \in \mathbb{R}$ *and* $|p - \sum_{i=1}^{l} t_i\vec{v}_i| \le r\} \subset \mathbb{R}^n$,

the lower *union* $U^-(p;r) = \bigcup\{(U+\vec{v})$ *such that* $\vec{v}\in\Lambda, (U+\vec{v}) \subset C(p;r)\} \subset \mathbb{R}^n$,

the upper *union* $U^+(p;r) = \bigcup\{(U+\vec{v})$ *such that* $\vec{v}\in\Lambda, (U+\vec{v}) \cap C(p;r) \neq \emptyset\}$.

Let the unions $U^{\pm}(p;r)$ *contain* $m^{\pm}(p;r)$ *shifted cells of* $U+\vec{v}$ *for some* $\vec{v}\in\Lambda$. *Let* S *have* $m = |S\cap U|$ *points in* U. *Then the number of points from* S *in* $C(p;r)$ *satisfies*

$$\left(\frac{r-w}{\mathrm{PPC}(S)}\right)^l \le m^-(p;r)m \le |S\cap C(p;r)| \le m^+(p;r)m \le \left(\frac{r+w}{\mathrm{PPC}(S)}\right)^l.$$

LEMMA 3.5 (distance bounds). *In the notations of Lemma 3.4, let the subspace* R^{n-l} *be orthogonal to* R^l, *which spanned by the first* l *basis vectors of a cell* U. *Let the* height h *of a cell-periodic set* $S\subset\mathbb{R}^n$ *with the cell* U *be the maximum distance between points in the orthogonal projection of* S *to* R^{n-l}, *so if* $l=n$, *then* $h=0$. *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 the full set* S. *Then* $\mathrm{PPC}(S)\sqrt[l]{k} - w < d_k(S;p) \le \sqrt{(\mathrm{PPC}(S)\sqrt[l]{k}+w)^2+h^2}$, $k\ge 1$.

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

Proof of Theorem 3.6. Lemma 3.5 gives the following bounds for $\delta_k = \dfrac{d_k(S;p)}{\sqrt[l]{k}} - \mathrm{PPC}(S)$. The lower bound is $\delta_k > -u_k$, where $u_k = \dfrac{w}{\sqrt[l]{k}} \to 0$ as $k\to+\infty$ because w is fixed. The upper bound is $\delta_k \le \sqrt{(\mathrm{PPC}(S)+u_k)^2 + (h/\sqrt[l]{k})^2} - \mathrm{PPC}(S) \to 0$ as $k\to+\infty$, because h is fixed. Hence $\delta_k = \dfrac{d_k(S;p)}{\sqrt[l]{k}} - \mathrm{PPC}(S) \to 0$ as $k\to+\infty$. ∎

By Theorem 3.6, $\mathrm{AMD}_k(S)$ and all distances in the last column of $\mathrm{PDD}(S;k)$ asymptotically approach $\mathrm{PPC}(S)\sqrt[l]{k}$ as $k\to+\infty$ and hence are largely determined by $\mathrm{PPC}(S)$ for large k. That is why the most descriptive information is contained in $\mathrm{PDD}(S;k)$ for smaller values of k, e.g. we use $k=100$ atomic neighbors in most experiments on crystals. To neutralize the asymptotic growth, we subtract and also normalize by the term $\mathrm{PPC}(S)\sqrt[l]{k}$ to get simpler invariants under uniform scaling.

DEFINITION 3.7 (simplified invariants ADA, PDA, AND, PND). *Let* $S\subset\mathbb{R}^n$ *be any* l*-periodic set with an underlying lattice generated by* l *vectors. The* Average Deviation from Asymptotic *is* $\mathrm{ADA}_k(S) = \mathrm{AMD}_k(S) - \mathrm{PPC}(S)\sqrt[l]{k}$ *for* $k\ge 1$. *The* Pointwise Deviation from Asymptotic $\mathrm{PDA}(S;k)$ *is obtained from the matrix* $\mathrm{PDD}(S;k)$ *by subtracting* $\mathrm{PPC}(S)\sqrt[l]{j}$ *from any distance in a row* i *and a column* j *for* $i\ge 1\le j\le k$. *The* Average Normalized Deviation *is* $\mathrm{AND}_k(S) = \mathrm{ADA}_k(S)/(\mathrm{PPC}(S)\sqrt[l]{k})$, $k\ge 1$. *The* Pointwise Normalized Deviation $\mathrm{PND}(S;k)$ *obtained from* $\mathrm{PDA}(S;k)$ *by dividing every element in a row* i *and a column* j *by* $\mathrm{PPC}(S)\sqrt[l]{j}$ *for* $i\ge 1\le j\le k$.

COROLLARY 3.8 (invariance of AND, PND under uniform scaling). *For any* l*-periodic set* $S\subset\mathbb{R}^n$, $\mathrm{AND}_k(S)$ *and* $\mathrm{PND}(S;k)$ *in Definition 3.7 are invariant under isometry and uniform scaling for any* $k\ge 1$. *Moreover,* $\mathrm{AND}_k(S)\to 0$ *as* $k\to+\infty$.

Proof. By Theorem 3.1, $\mathrm{PDD}(S;k)$ and hence all deviations in Definition 3.7 are invariant under isometry. Under uniform scaling $p \mapsto cp$ for a real constant $c \neq 0$, any inter-point distance and $\mathrm{PPC}(S) = \sqrt[l]{\dfrac{\mathrm{vol}[U \cap R^l]}{mV_l}}$ is multiplied by c because $\mathrm{vol}[U \cap R^l]$ is scaled by the factor c^l. Hence $\mathrm{AND}_k(S)$ and $\mathrm{PND}(S;k)$ are invariant under both isometry and uniform scaling. To prove that $\mathrm{AND}_k(S) \to 0$ as $k \to +\infty$, use Theorem 3.6: $\mathrm{AND}_k(S) = \dfrac{\mathrm{ADA}_k(S)}{\mathrm{PPC}(S)\sqrt[l]{k}} = \dfrac{\mathrm{AMD}_k(S)}{\mathrm{PPC}(S)\sqrt[l]{k}} - 1 \to \dfrac{\mathrm{PPC}(S)}{\mathrm{PPC}(S)} - 1 = 0$. □

We conjecture that $\mathrm{ADA}_k(S) \to 0$ as $k \to +\infty$ without the extra division by $\sqrt[l]{k}$ for $l \geq 2$, which is confirmed by experiments on crystals and holds for $S = \mathbb{Z}^n$ in SM3.

The key input sizes for computing $\mathrm{PDD}(S;k)$ of any l-periodic point set $S \subset \mathbb{R}^n$ are the number m of points in a unit cell U and the number k of neighbors. The full input consists of k, a basis of U 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 on both 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 each row before collapsing in Definition 3.1, see a Python code in SM2 of supplementary materials.

THEOREM 3.9 (PDD complexity). *Let $S \subset \mathbb{R}^n$ be any l-periodic set with a minimum inter-point distance $d_{\min}$ and a unit cell $U = P \times R^{n-l}$, where $P \subset \mathbb{R}^l$ is a parallelepiped in the l-dimensional subspace R^l with the orthogonal subspace R^{n-l} in $\mathbb{R}^n$. Consider the* width $w = \sup\limits_{u,v \in P} |\vec{u} - \vec{v}|$ *and the* height h *equal to the maximum distance between points in the orthogonal projection of S to R^{n-l}. If the motif $M = S \cap U$ consists of m points, then $\mathrm{PDD}(S;k)$ can be computed for any $k \geq 1$ in time*

$$O(km(2^{4n}\log k + \log m) + 2^{12n} m \log^2 k + (2^{8n}/l)k \log k + a^l bk),$$

where $a = 1 + \dfrac{2.5w + 2h}{\mathrm{PPC}(S)}$ and $b = \log(2\mathrm{PPC}(S) + 3w + 5h) - \log d_{\min}$. The complexity of $\mathrm{AMD}(S;k)$ and invariants $\mathrm{PDA}(S;k), \mathrm{PND}(S;k)$ from Definition 3.7 is the same as of $\mathrm{PDD}(S;k)$ because the extra computations can be done in time $O(km)$.

Proof of Theorem 3.9. In the notations of Lemma 3.4, we have integers $1 \leq l \leq n$ and a basis $\vec{v}_1, \dots, \vec{v}_n$ of $\mathbb{R}^n$. The first l basis vectors $\vec{v}_1, \dots, \vec{v}_l$ generate the subspace $\mathbb{R}^l \subset \mathbb{R}^n$ and the lattice $\Lambda \subset \mathbb{R}^l$. Fix the origin $0 \in \mathbb{R}^n$ be at the center of the parallelepiped $U \cap R^l$. Then any point $p \in M = S \cap U$ is covered by the closed ball $\bar{B}(0;r)$ for the radius $r = \sqrt{(0.5w)^2 + h^2} \leq 0.5w + h$. By Lemma 3.5, all k neighbors of p are covered by the closed cylinder $C(0;R)$ of the radius $R = r + \sqrt{(\mathrm{PPC}(S)\sqrt[l]{k} + w)^2 + h^2} \leq \mathrm{PPC}(S)\sqrt[l]{k} + 1.5w + 2h$. To generate all Λ-translates of M within $C(0;R)$, we gradually extend U in cylindrical layers by adding more shifted cells $U + \vec{v}$ for vectors $v \in \Lambda$ until we get the upper union $U^+(0;R)$ covering the cylinder $C(0;R)$. The upper union $U^+(0;R)$ includes k neighbors of each motif point and has the size $\mu = |S \cap U^+(0;R)| = m^+(0;R)m$ estimated by Lemma 3.4:

$$\mu \leq \left(\frac{R+w}{\mathrm{PPC}(S)}\right)^l \leq \left(\frac{\mathrm{PPC}(S)\sqrt[l]{k} + 2.5w + 2h}{\mathrm{PPC}(S)}\right)^l = \left(\sqrt[l]{k} + \frac{2.5w + 2h}{\mathrm{PPC}(S)}\right)^l =$$

$$= k\left(1+\frac{2.5w+2h}{\mathrm{PPC}(S)\sqrt[l]{k}}\right)^l \le k\left(1+\frac{2.5w+2h}{\mathrm{PPC}(S)}\right)^l = a^l k, \text{ where } a = 1+\frac{2.5w+2h}{\mathrm{PPC}(S)}.$$

For a nearest neighbor search [23], we can build a compressed cover tree on μ points of $T = S \cap U^+(0;R)$ in time $O(\mu c_{\min}^8 \log \frac{2R+h}{d_{\min}})$ by [24, Theorem 3.7], where $c_{\min} \le 2^n$ is the minimized expansion constant of T, and $\frac{2R+h}{d_{\min}}$ is the upper bound for the ratio of max/min inter-point distances. Then $R \le \mathrm{PPC}(S)\sqrt[l]{k} + 1.5w + 2h$ gives

$$\log(2R+h) \le \log(\sqrt[l]{k}(2\mathrm{PPC}(S)+3w+5h)) = \log(2\mathrm{PPC}(S)+3w+5h) + (\log k)/l,$$

$$\text{so } \log\frac{2R+h}{d_{\min}} = b + \frac{1}{l}\log k, \text{ where } b = \log(2\mathrm{PPC}(S)+3w+5h) - \log d_{\min}.$$

By [24, Theorem 4.9], using a compressed cover tree on T, we can find k neighbors of m points from $S \cap U$ among μ points of T in time $O(mc^2 \log k(c_{\min}^{10} \log \mu + ck))$, where $c \le 2^n$ is the expansion constant of T. Because $\log \mu \le \log k + l \log a$, we can compute all distances from each of m points to their k nearest neighbors in T in time

$$O(\mu(b+(\log k)/l)c_{\min}^8) + O(mc^2 \log k(c_{\min}^{10}\log\mu + ck)) \le$$
$$O(a^l k(b+(\log k)/l)2^{8n}) + O(m2^{2n}\log k(2^{10n}(\log k + l\log a) + 2^{2n}k)) \le$$
$$O(a^l bk + (2^{8n}/l)k\log k) + O(2^{4n}m(k\log k + 2^{8n}(\log^2 k + l\log a \log k)) \le$$
$$O(2^{4n}(m+2^{4n}/l)k\log k + 2^{12n}m\log^2 k + a^l bk), \text{ where we used } l\log a \le O(\log k).$$

The ordered lists of distances from points $p \in S \cap U$ to their k nearest neighbors in T are the rows of the matrix $D(S;k)$. It remains to lexicographically sort m lists of ordered distances, which needs time $O(km \log m)$, because a comparison of ordered lists of the length k takes $O(k)$ time. The total time for $\mathrm{PDD}(S;k)$ is

$$O(2^{4n}(m+2^{4n}/l)k\log k + 2^{12n}m\log^2 k + a^l bk) + O(km\log m) =$$
$$O(km(2^{4n}\log k + \log m) + 2^{12n}m\log^2 k + (2^{8n}/l)k\log k + a^l bk). \qquad \square$$

The worst-case estimate in Theorem 3.9 is conservative due to the upper bound 2^n for the expansion constants $c_{\min}, c$ from [24, Definition 1.4]. We conjecture that this upper bound can be reduced to 2^l for any l-periodic point set $S \subset \mathbb{R}^n$.

For any fixed dimensions $l \le n$, if we ignore the parameters $a, b, d_{\min}$, and $\mathrm{PPC}(S)$, then the complexity in Theorem 3.9 becomes $O(km(\log k + \log m))$, which is near-linear in both k, m. For the most practical dimensions $l = n = 3$, experiments in section 6 will report running times in minutes on a modest desktop computer for about 1.5 million real crystals from the world's largest experimental databases.

4. Lipschitz continuous Earth Mover's Distance on invariants. This section proves the continuity of the vectorial invariants AMD, ADA, AND, matrix invariants PDD, PDA, PND, and their moments. For matrix invariants, we will use the Earth Mover's Distance (EMD) [53], which is well-defined for any weighted distributions of different sizes. Definition 4.1 of EMD makes sense for any matrix invariant $I(S)$ that is an unordered collection of row vectors $\vec{R}_i(S)$ with weights $w_i(S) \in (0,1]$ satisfying $\sum_{i=1}^{m(S)} w_i(S) = 1$. Each row $\vec{R}_i(S)$ should have a size independent of i. This size can be the number k of neighbors for $\mathrm{PDD}(S;k)$. For any vectors $\vec{R}_i = (r_{i1}, \dots, r_{ik})$ and $\vec{R}_j = (r_{j1}, \dots, r_{jk})$, the *Minkowski* distance is $L_q(\vec{R}_i, \vec{R}_j) = \left(\sum_{l=1}^{k} |r_{il} - r_{jl}|^q\right)^{1/q}$ for any real $q \ge 1$ and $L_{+\infty}(\vec{R}_i, \vec{R}_j) = \max_{l=1,\dots,k} |r_{il} - r_{jl}|$.

DEFINITION 4.1 (Earth Mover's Distance EMD_q). *Let discrete sets S, Q in a metric space have weighted distributions $I(S), I(Q)$ as above. A* flow *from $I(S)$ to $I(Q)$ is an $m(S) \times m(Q)$ matrix whose element $f_{ij} \in [0,1]$ is a partial* flow *from $\vec{R}_i(S)$ to $\vec{R}_j(Q)$. For any real $q \geq 1$, the* Earth Mover's Distance *is the minimum* cost $\mathrm{EMD}_q(I(S), I(Q)) = \sum\limits_{i=1}^{m(S)} \sum\limits_{j=1}^{m(Q)} f_{ij} L_q(\vec{R}_i(S), \vec{R}_j(Q))$ *subject to* $\sum\limits_{j=1}^{m(Q)} f_{ij} = w_i(S)$ *for* $i = 1, \dots, m(S)$, $\sum\limits_{i=1}^{m(S)} f_{ij} = w_j(Q)$ *for* $j = 1, \dots, m(Q)$, $\sum\limits_{i=1}^{m(S)} w_i(S) = 1 = \sum\limits_{j=1}^{m(Q)} w_j(Q)$.

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 vector $\vec{R}_i(S)$ 'flows' into all vectors $\vec{R}_j(Q)$ via partial flows $f_{ij} \in [0,1]$ for $j = 1, \dots, m(Q)$. The second condition $\sum\limits_{i=1}^{m(S)} f_{ij} = w_j(Q)$ means that all 'flows' f_{ij} from $\vec{R}_i(S)$ for $i = 1, \dots, m(S)$ 'flow' into $\vec{R}_j(Q)$ up to the maximum weight $w_j(Q)$. The last condition forces all vectors $\vec{R}_i(S)$ to 'flow' to all vectors $\vec{R}_j(Q)$.

The EMD satisfies all metric axioms [53, appendix], needs $O(m^3 \log m)$ time for distributions of a maximum size m and can be approximated in $O(m)$ time [59, 55].

The Lipschitz continuity of invariants in EMD will use bounded perturbations of points up to ε in the metric d_X of an ambient space X. Because atoms are not outliers or noise, such perturbations can be formalized as the *bottleneck distance* $d_B(S, Q) = \inf\limits_{g: S \to Q} \sup\limits_{p \in S} d_X(g(p), p)$ minimized over all bijections $g : S \to Q$ between (possibly infinite) sets. This definition is computationally intractable even for finite sets due to exponentially many $m!$ bijections between sets of m points. [63, 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 S, Q are lattices of equal density (equal unit cell volume), they have a finite bottleneck distance d_B by [21, 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 [63, 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 (Lipschitz continuity). *Let M be a finite subset of a discrete set S in a space X with a metric d_X. Let Q and its finite subset T be obtained from S and M, respectively, by perturbing every point of S up to ε in the metric d_X. Fix any real $q \in [1, +\infty]$ and an integer $k \geq 1$. Interpret $\sqrt[q]{k}$ as 1 in the limit case $q = +\infty$.*

(a) Then $\mathrm{EMD}_q(\mathrm{PDD}(S, M; k), \mathrm{PDD}(Q, T; k)) \leq 2\varepsilon \sqrt[q]{k}$.

(b) If S, Q are l-periodic and $\min\{r(S), r(Q)\} > \varepsilon$, *then* $\mathrm{PPC}(S) = \mathrm{PPC}(Q)$, *and* $\mathrm{EMD}_q(\mathrm{PDA}(S; k), \mathrm{PDA}(Q; k)) \leq 2\varepsilon \sqrt[q]{k}$, $\mathrm{EMD}_q(\mathrm{PND}(S; k), \mathrm{PND}(Q; k)) \leq \dfrac{2\varepsilon \sqrt[q]{k}}{\mathrm{PPC}(S)}$.

Theorem 4.2 is proved in SM3 of supplementary materials similar to [65, Lemma 8] for $q = +\infty$. All columns of PDD, PDA, PND are ordered by the index k of neighbors. Though their rows are unordered (as points of a motif M), all such matrices even with different numbers of rows can be compared by Earth Mover's Distance, or by

any other metrics on weighted distributions, see Definition 4.1. We can simplify any PDD into a fixed-size matrix, which can be flattened into a vector, while keeping the continuity and almost all invariant data. Any distribution of m unordered values can be reconstructed from its m moments 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 degree m polynomial whose coefficients are expressed via the m moments [45], e.g. any $a,b\in\mathbb{R}$ can be found from $a+b, a^2+b^2$ as 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 $b\ge 1$, the b-th *moment* [36, section 2.7] is $\mu_b(A)=\sqrt[b]{m^{1-b}\sum_{i=1}^{m} w_i a_i^b}$, so $\mu_1(A)=\sum_{i=1}^{m} w_i a_i$ is the usual average.

For any integer $b\ge 2$, we avoid subtracting μ_1 from the numbers $a_1,\dots,a_m$, which would convert μ_2 into the standard deviation σ, and normalize by the factor $m^{(1/b)-1}$ to guarantee the continuity of moments with the Lipschitz constant $\lambda=2$.

DEFINITION 4.3 (b-moments matrix $\mu^{(b)}$). *Fix any integer $b\ge 1$. Let $I(S)$ be a matrix invariant of a cell-periodic set S. For every column A of $I(S)$, consisting of unordered numbers with weights, write the column $(\mu_1(A),\dots,\mu_b(A))$. All new columns form the* b-moments matrix $\mu^{(b)}[I(S)]$*, which has b canonically ordered rows.*

For $b=1$, the $1\times k$ matrix $\mu^{(1)}[\mathrm{PDD}(S;k)]$ appeared in Definition 3.1 as the vector $\mathrm{AMD}(S;k)=(\mathrm{AMD}_1,\dots,\mathrm{AMD}_k)$. All rows and columns of the b-moments matrix $\mu^{(b)}[I(S)]$ are ordered but this matrix is a bit weaker than $I(S)$ because each column can be reconstructed from its moments (for a large enough b) only up to permutation. We can flatten any moments matrix $\mu^{(b)}[I(S)]$ with indexed entries to a vector and use this vector for machine learning on discrete sets S [6, 5].

Theorem 4.4 substantially extends [63, Theorem 4.2] to other isometry invariants of any finite and l-periodic sets for a Minkowski metric L_q with any real $q\ge 1$.

THEOREM 4.4 (lower bounds of EMD). *For finite or l-periodic sets $S,Q\subset\mathbb{R}^n$,*

(a) $\mathrm{EMD}_q(\mathrm{PDD}(S;k),\mathrm{PDD}(Q;k))\ge L_q(\mathrm{AMD}(S;k),\mathrm{AMD}(Q;k))$*;*

(b) $\mathrm{EMD}_q(\mathrm{PDA}(S;k),\mathrm{PDA}(Q;k))\ge L_q(\mathrm{ADA}(S;k),\mathrm{ADA}(Q;k))$*;*

(c) $\mathrm{EMD}_q(\mathrm{PND}(S;k),\mathrm{PND}(Q;k))\ge L_q(\mathrm{AND}(S;k),\mathrm{AND}(Q;k))$ *for any $q,k\ge 1$.*

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 of Theorem 5.1. 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 of 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 [57, 38].□

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* $V(\Lambda;p) = \{q \in \mathbb{R}^n$ such that $|q-p| < |q-p'|$ for any $p' \in \Lambda - p\}$ is the neighborhood of all points $q \in \mathbb{R}^n$ that are strictly closer to p than to all other points p' of the lattice Λ [22].

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 example, for a generic lattice $\Lambda \subset \mathbb{R}^2$, the domain $V(\Lambda;p)$ 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 $N(\Lambda) = \Lambda \cap \bar{B}(0;r) \setminus \{0\}$ 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 $q \in V(\Lambda;0)$.*

Consider all unordered points $p_1, \dots, p_n \in N(\Lambda)$ that are linearly independent*, i.e. the vectors $\vec{p}_1, \dots, \vec{p}_n$ form a linear basis of $\mathbb{R}^n$. For any point $q \in V(\Lambda;0)$, a lexicographically smallest list of distances $d_1(q) \le \dots \le d_n(q)$ from q to all linearly independent points $p_1, \dots, p_n \in N(\Lambda)$ is called the list of* basis distances *of q.*

The linear independence of vectors $\vec{p}_1, \dots, \vec{p}_n$ in Definition 5.2 guarantees that any point q is uniquely determined in $\mathbb{R}^n$ by its distances $|q|, d_1(q), \dots, d_n(q)$ to $n+1$ neighbors $0, p_1, \dots, p_n$, which are not in the same $(n-1)$-dimensional subspace.

Let Λ be generated by $(2,0), (0,1)$. The Voronoi domain $V(\Lambda;0)$ is the rectangle $(-1,1) \times (-0.5, 0, 5)$. 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 in Definition 5.2 Λ has a radius $r < 2$, then $\Lambda \cap \bar{B}(0;r) \setminus \{0\} = \{(0, \pm 1)\}$ is in the 1-dimensional subspace (y-axis) of $\mathbb{R}^2$. For $q = (0, 0.4)$, considering all pairs $(\vec{p}_1, \vec{p}_2)$ that generate $\mathbb{R}^2$ among the four possibilities $((0, \pm 1), (\pm 2, 0))$, we find the basis distances $d_1(q) = 0.6 < d_2(q) = \sqrt{0.4^2 + 2^2} \approx 2.04$ 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* covering 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 $\bar{V}(\Lambda;0)$ of the Voronoi domain has $n+1$ lattice neighbors (within Λ) among the origin $0 \in \Lambda$ and at least $2(2^n - 1)$ Voronoi neighbors of 0 [16]. In $\mathbb{R}^n$, any vertex of the boundary of $V(\Lambda;0)$ is equidistant to at least $n+1$ points of Λ (the origin 0 and its n Voronoi neighbors). The longest of these distances to Voronoi neighbors is the covering radius $R(\Lambda)$. The ball $\bar{B}(0; 2R(\Lambda))$ covers all Voronoi neighbors of 0 and hence the whole neighbor set $N(\Lambda)$. □

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

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

(5.4b) For 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 $q \in S \cap V(\Lambda;0)$, let $d_0 = |q|$ be its distance to the closest neighbor $p_0 = 0$ in Λ. Take any linearly independent points $p_1, \dots, p_n \in N(\Lambda)$ and any distances $d_1 \le \dots \le d_n$ from q to some points in $S \cap \bar{B}(0; 2R(\Lambda))$. The $n+1$ spheres $\partial B(p_i; d_i)$ can meet at a single point of $S \cap V(\Lambda;0)$ only if $d_1 \le \dots \le d_n$ are the basis distances of q and only for two tuples $p_1, \dots, p_n \in N(\Lambda)$ related by $\vec{v} \mapsto -\vec{v}$.

Condition (5.4b) means that all inter-point distances are distinct apart from necessary exceptions due to periodicity. Since any periodic set $S = M + \Lambda \subset \mathbb{R}^n$ is invariant under translations along all vectors of Λ, condition (5.4b) for $|\vec{v}| \le 2R(\Lambda)$ can be checked only for vectors from all points of S in the original Voronoi domain $V(\Lambda;0)$ to all points in the domain $3V(\Lambda;0)$ extended by factor 3.

Condition (5.4b) implies that S has no points on the boundary $\partial V(\Lambda;0)$, because any such point is equidistant to points $0, v \in \Lambda$ and hence should belong to Λ.

Let a *lattice distance* be the Euclidean distance from any point $p \in M = S \cap V(\Lambda;0)$ to its lattice translate $\vec{p} + \vec{v}$ for all $\vec{v} \in \Lambda$. Condition (5.4a) guarantees that only a lattice distance d appears together with $2d$ (and possibly with higher multiples) in a row of $\mathrm{PDD}(S;k)$. Any lattice distance d and its multiples are repeated twice in every row, because any lattice is centrally symmetric.

LEMMA 5.5 (almost any periodic set is distance-generic). *Let $S = M + \Lambda \subset \mathbb{R}^n$ be any periodic point set. For any $\varepsilon > 0$, one can perturb coordinates of a basis of Λ and of points from M up to ε such that the resulting perturbation S' of S is a distance-generic periodic point set in the sense of Definition 5.4.*

Proof. We can assume that the motif M of S is a subset of the open Voronoi domain $V(\Lambda;0)$ and include the origin 0. We show below that conditions (5.4a,b) define a codimension 1 *discriminant* (singular subspace) in the space of all parameters P that are coordinates of points of M and of basis vectors of Λ. In condition (5.4a), for any points $p, q \in V(\Lambda;0)$, the orthogonality is expressed as $f_a(p,q) = \vec{p} \cdot \vec{q} = \sum_{i=1}^{n} p_i q_i = 0$. In condition (5.4b), for any vectors $\vec{u}, \vec{v}$ that join points of S, have a maximum length $2R(\Lambda)$, and satisfy $u \ne \pm l\vec{v}$ for $l = 1, 2$, the equality $|\vec{u}| = l|\vec{v}|$ can be written as $f_b(u,v) = \sum_{i=1}^{n} u_i^2 - l^2 \sum_{i=1}^{n} v_i^2 = 0$. So condition (5.4a) forbids a codimension 1 subspace defined by finitely many equations $f_b(u,v) = 0$ for all u, v above.

Similarly, condition (5.4c) can be written via polynomial equations in point coordinates. For any fixed radii $d_0, \dots, d_n$, almost all $n+1$ spheres in $\mathbb{R}^n$, whose centers are not in any $(n-1)$-dimensional affine subspace, have no common points. Hence condition (5.4c) also forbids a codimension 1 subspace. All involved functions in equations above are continuous in the coordinates of points and basis vectors. Then a motif $M = S \cap V(\Lambda;0)$ and a basis of Λ can be slightly perturbed to move S to S' outside the union of all finitely many codimension 1 subspaces above. Hence any periodic point set S can be made distance-generic by a small enough perturbation. □

The number m of points in a unit cell U is an isometry invariant because any isometry maps U to another cell where the motif $S \cap U$ has the same size. In dimensions $n = 2, 3$, any lattice Λ can be reconstructed from its isometry invariants [41, 39].

Theorem 5.6 reconstructs a periodic point set $S = M + \Lambda \subset \mathbb{R}^n$ in any dimension $n \ge 2$ from $\mathrm{PDD}(S;k)$ assuming that an n-dimensional lattice Λ of S is given. Complete isometry invariants of lattices in dimensions $n = 2, 3$ appeared in [41, 39].

THEOREM 5.6 (generic completeness of PDD). *Let $S = M + \Lambda \subset \mathbb{R}^n$ be any distance-generic periodic set whose motif M has m points. Let $R(\Lambda)$ be the smallest radius R such that all closed balls with centers $p \in \Lambda$ and radius R cover $\mathbb{R}^n$. For any k such that all distances in the last column of $\mathrm{PDD}(S;k)$ are larger than $2R(\Lambda)$, the set S can be reconstructed from Λ, m and $\mathrm{PDD}(S;k)$, uniquely up to isometry in $\mathbb{R}^n$.*

Proof. 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 $\mathrm{PDD}(S;k)$ by replacing every row of a weight w with the integer number mw of identical rows having the same 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 the periodic point set 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 $\vec{p}-\vec{q} \notin \Lambda$. Take a unique point $q \in S \cap V(\Lambda;0) \setminus \{0\}$ that has the smallest distance $d_0 = |q|$ to the origin and hence uniquely determined in the row of q in $\mathrm{PDD}(S;k)$. Then we will look for n basis distances $d_1 < \cdots < d_n$ from q to its further n lattice neighbors $p_1,\dots,p_n \in N(\Lambda) \subset \Lambda - 0$ such that $\vec{p}_1,\dots,\vec{p}_n$ form a linear basis of $\mathbb{R}^n$. All basis distances $d_0,\dots,d_n$ 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 the shortest distance $d_0 = |q|$.

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 $\mathrm{PDD}(S;k)$, we will try all subsequences $d_1 < \cdots < d_n$ of distinct distances shared by both rows. Similarly, we cannot be sure that n closest neighbors of q in $S \setminus \{0\}$ define linearly independent vectors of Λ.

Hence we try all linearly independent points $p_1,\dots,p_n \in N(\Lambda)$. For all finitely many choices, we check if the $n+1$ spheres $\partial B(p_i;d_i)$ meet at a single point in $V(\Lambda;0)$, which will be the required point q. These $(n-1)$-dimensional spheres are 1D circles for $n=2$ and 2D spheres for $n=3$. Condition (5.4c) will guarantee below a reconstruction of q as a single intersection of these $n+1$ spheres of dimension $n-1$.

The basis distances $d_1 < \cdots < d_n$ of q should form the lexicographically smallest list among all lists of distances from q to points $p_1,\dots,p_n \in N(\Lambda)$. This smallest list emerges for at most two tuples of linearly independent points $p_1,\dots,p_n \in N(\Lambda)$ related by the isometry $\vec{v} \mapsto -\vec{v}$, which preserves Λ. For a first reconstruction outside Λ, we choose any of these tuples and find the intersection point $q = \cap_{i=0}^{n} \partial B(p_i;d_i)$.

Any other point $p \in (S \setminus \{0,q\}) \cap V(\Lambda;0)$ is uniquely determined similarly to the point q above by 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 is uniquely determined by the already fixed point q and lattice Λ. ∎

6. Detecting near-duplicates in the world's largest databases. This section reports thousands of previously unknown (near-)duplicates in the world's largest databases [60, 30, 67, 34]. The sizes in Table 2 below are the numbers of all periodic crystals (with no disorder and full geometric data) in September 2024 (total number is 1,433,650, nearly 1.5 million), see all experimental details in SM1.

We first used the vector $\mathrm{ADA}(S;100)$ to find nearest neighbors across all databases by k-d trees [26] up to $L_\infty \leq 0.01$Å. Since the smallest inter-atomic distances are about 1Å $= 10^{-10}$m, atomic displacements up to 0.01Å are considered experimental

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

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

noise. For the closest pairs found by ADA$(S;100)$, the stronger PDA$(S;100)$ can have only equal or larger EMD $\geq L_\infty$ by Theorem 4.4. The CSD, COD, ICSD should contain experimental structures. MP is obtained from ICSD by extra optimization.

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. Table 1 in [3, section 6] reported many thousands of exact duplicates, where chemical elements were replaced while keeping all coordinates fixed. These replacements are physically impossible without more substantial perturbations. Five journals are investigating integrity [12], see details in appendix SM1.

The bold numbers in Table 3 count near-duplicates and their percentages within each database, which should be filtered out else the ground truth data becomes skewed. Other numbers are counts and percentages 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	
near-duplicates	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 [15] 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 milliseconds (582 ms on average) for 500 random pairs in the CSD. On the same 500 pairs, PDA$(S;100)$ for two crystals and their distance EMD together took only 7.48 ms on average. All experiments were done on a typical desktop computer (AMD Ryzen 5 5600X 6-core, 32GB RAM).

7. Discussion. For hundreds of years, crystals were classified almost exclusively by discrete tools such as space groups or by using reduced cells, which are unique in theory. Fig. 2 (left) showed that any known crystal can be disguised by changing a unit cell, shifting atoms a bit, changing chemical elements, then claimed as 'new', see SM1. Such artificially generated structures threaten the integrity of experimental databases [12], which are skewed by previously undetectable near-duplicates.

These challenges motivated the stronger questions "how much different?" and "can I get a structure from its code?", which were formalized in Problem 1.6 aiming

TABLE 4
Running times in seconds (less than 8.5 hours in total) to find all near-duplicates in Table 3 with EMD ≤ 0.01Å *on* PDA$(S;100)$ *across all 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 [15] are extrapolated on the same machine, which completed Table 3 of near-duplicates across all the major 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

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 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. Appendix SM1 shows these geographic-style maps of all four databases in our invariant coordinates.

The key impact is the efficient barrier for noisy disguises of known structures because the invariants quickly find nearest neighbors of newly claimed materials in the existing databases, as shown for all crystals from GNoME [3] and A-lab [64].

Acknowledgments. This work was supported by the Royal Academy of Engineering Fellowship IF2122/186, the EPSRC New Horizons grant EP/X018474/1, the Royal Society APEX fellowship APX/R1/231152, and the AI for Chemistry hub (EP/Y028775/1). We thank all reviewers for their valuable time and suggestions.

REFERENCES

[1] P. K. Agarwal, K. Fox, A. Nath, A. Sidiropoulos, and Y. Wang, *Computing the Gromov-Hausdorff distance for metric trees*, Transactions on Algorithms, 14 (2018), pp. 1–20.
[2] 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.
[3] O. Anosova, V. Kurlin, and M. Senechal, *The importance of definitions in crystallography*, International Union of Crystallography Journal, 11 (2024), pp. 453–463.
[4] V. Arvind and G. Rattan, *The parameterized complexity of geometric graph isomorphism*, Algorithmica, 75 (2016), pp. 258–276.
[5] J. Balasingham, V. Zamaraev, and V. Kurlin, *Accelerating material property prediction using generically complete isometry invariants*, Scientific Reports, 14 (2024), p. 10132.
[6] J. Balasingham, V. Zamaraev, and V. Kurlin, *Material property prediction using graphs based on generically complete isometry invariants*, Integrating Materials and Manufacturing Innovation, 13 (2024), pp. 555–568.
[7] 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.
[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.
[12] D. S. Chawla, *Crystallography databases hunt for fraudulent structures*, ACS Central Science, 9 (2023), p. 1853–1855.
[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 and K. Kedem, *Improvements on geometric pattern matching problems*, in Scandinavian Workshop on Algorithm Theory, 1992, pp. 318–325.
[15] J. Chisholm and S. Motherwell, *Compack: a program for identifying crystal structure similarity using distances*, J. Applied Crystal., 38 (2005), pp. 228–231.
[16] J. Conway and N. Sloane, *Low-dimensional lattices. VI. Voronoi reduction of three-dimensional lattices*, Proceedings Royal Society A, 436 (1992), pp. 55–68.
[17] L. Cosmo, M. Panine, A. Rampini, M. Ovsjanikov, M. M. Bronstein, and E. Rodola, *Isospectralization, or how to hear shape, style, and correspondence*, in Computer Vision and Pattern Recognition, 2019, pp. 7529–7538.
[18] P. J. Davis, *Leonhard Euler's integral: A historical profile of the gamma function*, The American Mathematical Monthly, 66 (1959), pp. 849–869.
[19] 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.
[20] N. Dolbilin, J. Lagarias, and M. Senechal, *Multiregular point systems*, Discrete & Computational Geometry, 20 (1998), pp. 477–498.
[21] M. Duneau and C. Oguey, *Bounded interpolations between lattices*, Journal of Physics A: Mathematical and General, 24 (1991), p. 461.
[22] H. Edelsbrunner and R. Seidel, *Voronoi diagrams and arrangements*, Discrete & Computational Geometry, 1 (1986), pp. 25–44.
[23] Y. Elkin and V. Kurlin, *Counterexamples expose gaps in the proof of time complexity for cover trees introduced in 2006*, in Top. Data Analysis and Visualization, 2022, pp. 9–17.
[24] Y. Elkin and V. Kurlin, *A new near-linear time algorithm for k-nearest neighbor search using a compressed cover tree*, in Intern. Conference on Machine Learning), 2023, pp. 9267–9311.
[25] R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman lectures on physics: the new millennium edition*, vol. 1, 2011.
[26] F. Gieseke, J. Heinermann, C. Oancea, and C. Igel, *Buffer kd trees: processing massive nearest neighbor queries on GPUs*, in Intern. Conf. Machine Learning, 2014, pp. 172–180.
[27] M. Goodrich, J. S. Mitchell, and M. Orletsky, *Approximate geometric pattern matching under rigid motions*, Trans. Pattern Analysis and Machine Intel., 21 (1999), pp. 371–379.
[28] C. Gordon, D. Webb, and S. Wolpert, *Isospectral plane domains and surfaces via riemannian orbifolds*, Inventiones mathematicae, 110 (1992), pp. 1–22.
[29] 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.
[30] S. Gražulis, D. Chateigner, R. 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*, J Appl. Crystallography, 42 (2009), pp. 726–729.
[31] F. Hausdorff, *Dimension und äußeres maß*, Mathematische Annalen, 79 (1919), pp. 157–179.
[32] D. Huttenlocher, G. Klanderman, and W. Rucklidge, *Comparing images using the Hausdorff distance*, Trans. Pattern analysis and machine intelligence, 15 (1993), pp. 850–863.
[33] D. Hyde, *The sorites paradox*, in Vagueness: A guide, Springer, 2011, pp. 1–17.
[34] 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).
[35] M. Kac, *Can one hear the shape of a drum?*, Amer. Math. Monthly, 73 (1966), pp. 1–23.
[36] E. S. Keeping, *Introduction to statistical inference*, Courier Corporation, 1995.
[37] D. Kendall, D. Barden, T. Carne, and H. Le, *Shape and shape theory*, Wiley & Sons, 2009.
[38] J. B. Kruskal and M. Wish, *Multidimensional scaling*, no. 11, Sage, 1978.
[39] V. Kurlin, *A complete isometry classification of 3D lattices*, arxiv:2201.10543, (2022).
[40] V. Kurlin, *Polynomial-time algorithms for continuous metrics on atomic clouds of unordered points*, MATCH Comm. Math. Comp. Chemistry, 91 (2024), pp. 79–108.
[41] V. A. Kurlin, *Mathematics of 2-dimensional lattices*, Foundations of Computational Mathe-

matics, 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] S. Lim, F. Mémoli, and Z. Smith, *The gromov–hausdorff distance between spheres*, Geometry & Topology, 27 (2023), pp. 3733–3800.
[45] I. G. Macdonald, *Symmetric functions and Hall polynomials*, Oxford University Press, 1998.
[46] S. Majhi, J. Vitter, and C. Wenk, *Approximating gromov-hausdorff distance in euclidean space*, Computational Geometry, 116 (2024), p. 102034.
[47] 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.
[48] F. Mémoli, *Gromov–Wasserstein distances and the metric approach to object matching*, Foundations of Computational Mathematics, 11 (2011), pp. 417–487.
[49] F. Mémoli, Z. Smith, and Z. Wan, *The Gromov-Hausdorff distance between ultrametric spaces: its structure and computation*, arXiv:2110.03136, (2021).
[50] M. Mosca and V. Kurlin, *Voronoi-based similarity distances between arbitrary crystal lattices*, Crystal Research and Technology, 55 (2020), p. 1900197.
[51] 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).
[52] 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.
[53] 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.
[54] P. Sacchi, M. Lusi, A. J. Cruz-Cabeza, E. Nauha, and J. Bernstein, *Same or different – that is the question: identification of crystal forms from crystal structure data*, Cryst Eng Comm, 22 (2020), pp. 7170–7185.
[55] R. Sato, M. Cuturi, M. Yamada, and H. Kashima, *Fast and robust comparison of probability measures in heterogeneous spaces*, arXiv:2002.01615, (2020).
[56] 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.
[57] 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.
[58] M. Senechal, *Quasicrystals and geometry*, CUP Archive, 1996.
[59] 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.
[60] 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.
[61] 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.
[62] H. Weyl, *The classical groups: their invariants and representations*, Princeton Univ., 1946.
[63] D. Widdowson and V. Kurlin, *Resolving the data ambiguity for periodic crystals*, Advances in Neural Information Processing Systems, 35 (2022), pp. 24625–24638.
[64] D. Widdowson and V. Kurlin, *Navigation maps of the material space for automated self-driving labs of the future*, arxiv:2410.13796, (2024).
[65] 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.
[66] 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.
[67] 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.

SUPPLEMENTARY MATERIALS: POINTWISE DISTANCE DISTRIBUTIONS FOR DETECTING NEAR-DUPLICATES IN LARGE MATERIALS DATABASES*

DANIEL E. WIDDOWSON† AND VITALIY A. KURLIN‡

SM1. Details of experiments on the world's 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 $\mu^{(10)}[\mathrm{PDD}(S;100)]$ 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, 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 matrix $\mu^{(10)}[\mathrm{PDD}(S;100)]$ 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.

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 SM1 summarized below.

- The original 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. Table SM2 below shows how many matches were found, and how many also shared the same composition.

Table SM3 compares properties of past and new descriptors

*LaTeX2e Standard Macros were used from https://epubs.siam.org/journal-authors#macros

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).

CSD refcode	COD ID	Notes
LAVFAP	2001334	Mixed types in CIF
ZAYRUM	2003941	Mixed types in CIF
FONGAQ01	2005101	Mixed types in CIF
TIPYOG	2005914	Mixed types in CIF
HABTAF	2001740	Mixed types in CIF
AJIRAM01	2100097	Mixed types in CIF
LABSAI	2001822	Mixed types in CIF
DECTAI	4065524	Mixed types in CIF
WATMIO	4309447	Mixed types in CIF
NAJQUK	4323901	Mixed types in CIF
PIHJUL	4030494	Mixed types in 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 SM1

26 matches between the CSD and COD have identical geometry but different compositions.

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 SM2

*Number of exact matches (*EMD *within* 10^{-13}Å*) between four databases.*

Descriptor	Invariant	Continuity	Complete	Reconstruction	Time
primitive cell	×	×	×	×	✓
reduced cell	✓	×	×	×	✓
space group	✓	×	×	×	✓
PDF [SM8]	✓	✓	×	×	✓
SOAP [SM2]	✓	✓	×	×	✓
densities [SM4]	✓	✓	✓*	×	✓*
isosets [SM1]	✓	✓	✓	✓	✓*
AMD	✓	✓	×	×	✓
PDD	✓	✓	✓*	✓*	✓

TABLE SM3

Comparison of crystal descriptors with regards to the requirements of Problem 1.6. ✓ in the 'Computable' column indicates that only an approximate algorithm exists for distances, and ✓* in the 'Complete' and 'Reconstruction' columns means that the condition holds in* general position.

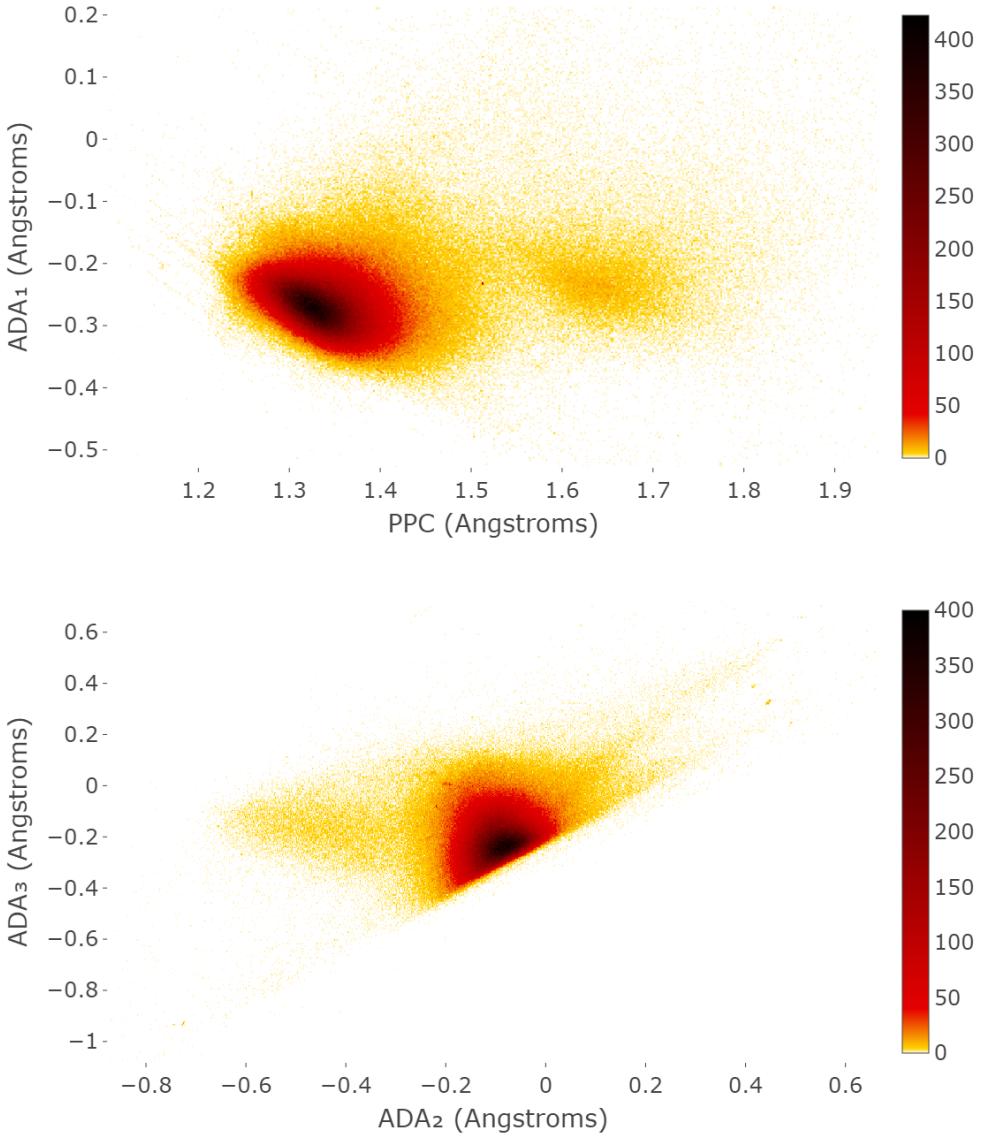

FIG. SM1. *The projections of the CSD in the invariants* $\mathrm{PPC}, \mathrm{ADA}_1, \mathrm{ADA}_2, \mathrm{ADA}_3$.

SM2. Examples and instructions for the PDD code and data. This appendix explains the code at https://pypi.org/project/average-minimum-distance.

SM2.1. Pseudocode for computing Pointwise Distance Distributions. 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

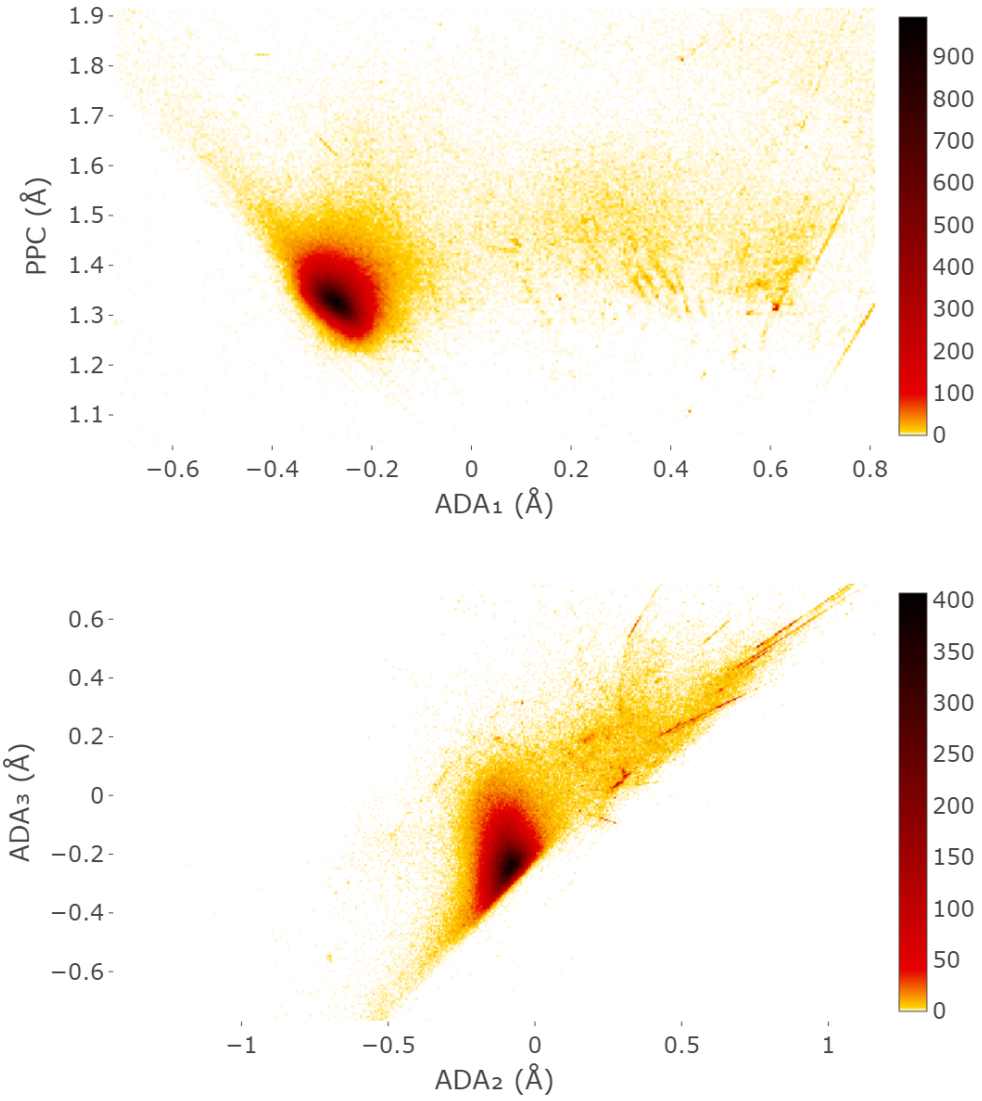

FIG. SM2. *The projections of the COD in the invariants* $\mathrm{PPC}, \mathrm{ADA}_1, \mathrm{ADA}_2, \mathrm{ADA}_3$.

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 the key objects and functions in use:

- A generator `g`, which creates points from the set S to find distances to,
- KDTrees (canonically k is the dimension here, in our case it's denoted n),

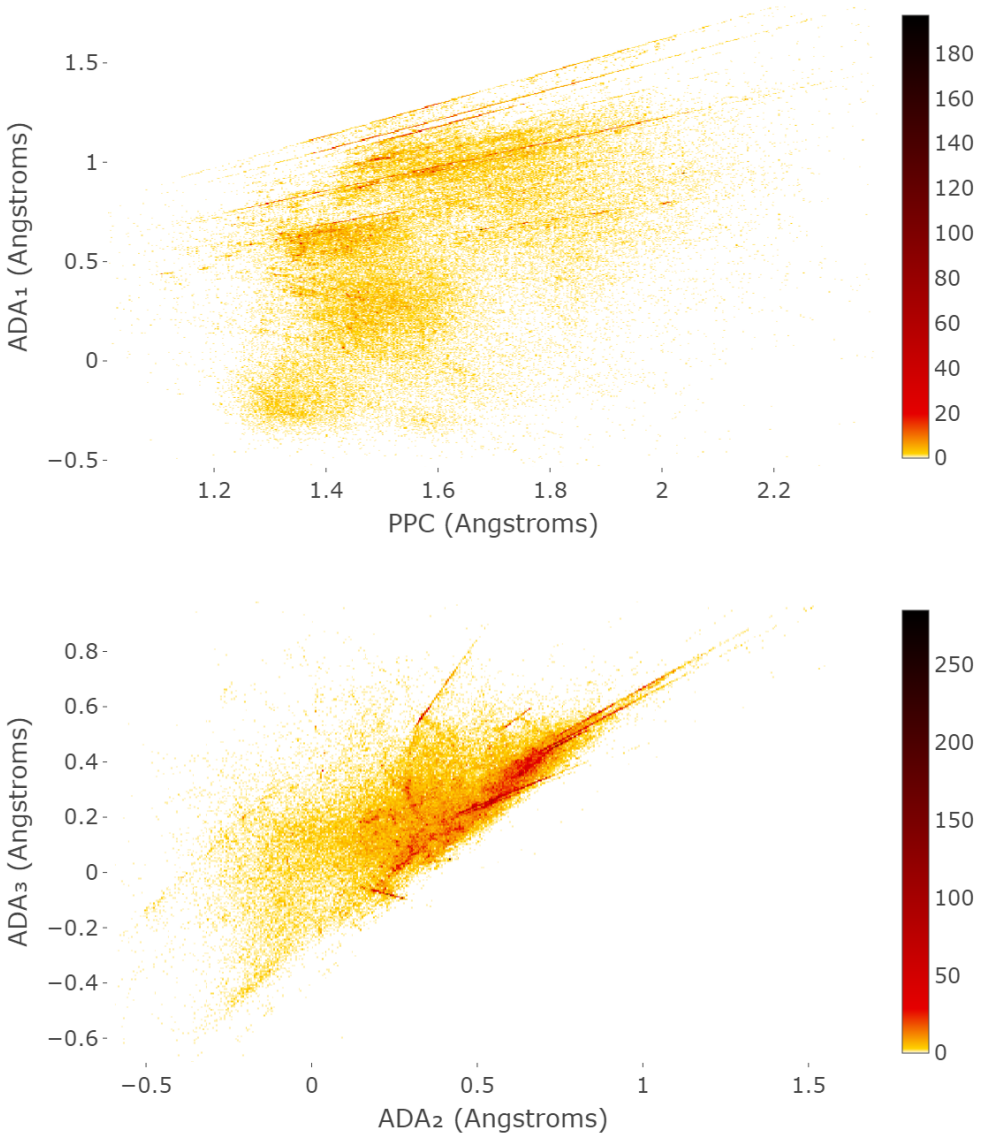

FIG. SM3. *The projections of the ICSD in the invariants* $\mathrm{PPC}, \mathrm{ADA}_1, \mathrm{ADA}_2, \mathrm{ADA}_3$.

data structures designed for fast nearest-neighbor lookup in $\mathbb{R}^n$.

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, respectively, are assumed

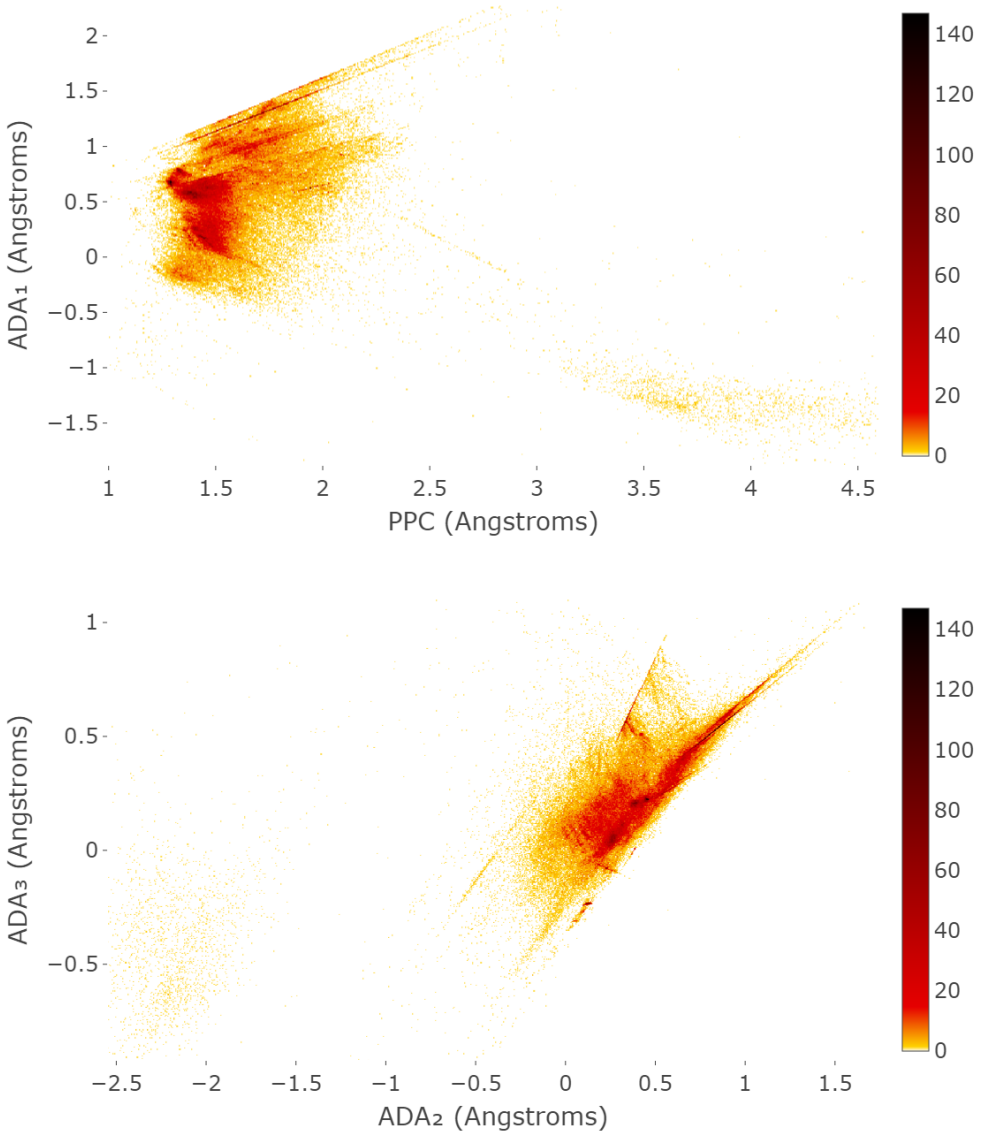

FIG. SM4. *The projections of the MP in the invariants* $\mathrm{PPC}, \mathrm{ADA}_1, \mathrm{ADA}_2, \mathrm{ADA}_3$.

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
```

SM2.2. 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`. 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
```

Here is the list of included example scripts and their parameters:

- `kite_trapezium_example.py` prints the PDDs of the 4-point sets K (kite) and T (trapezium) in Fig. SM5 (left), along with their EMD.
- `1D_sets_example.py` shows that the 1D periodic sets in Fig. SM5 (right) are distinguished by their PDDs for any $0 < r \leq 1$. This script requires r to be passed after the file name, e.g. '`python 1D_sets_example.py 0.5`'.

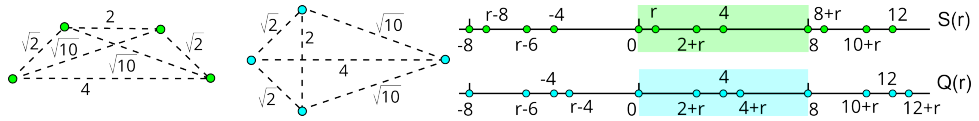

FIG. SM5. ***Left:*** *the 4-point sets* $K = \{(\pm 2, 0),\ (\pm 1, 1)\}$ *and* $T = \{(\pm 2, 0), (-1, \pm 1)\}$ *have the same pairwise distances* $\sqrt{2}, \sqrt{2}, 2, \sqrt{10}, \sqrt{10}, 4$. ***Right:*** *the sequences* $S(r) = \{0, r, 2+r, 4\} + 8\mathbb{Z}$ *and* $Q(r) = \{0, 2+r, 4, 4+r\} + 8\mathbb{Z}$ *for* $0 < r \leq 1$ *have the same Patterson function [SM6, p. 197, Fig. 2].*

- `T2_14_15_example.py` compares the crystals shown in Fig. SM6, whose original CIFs are included. This optionally accepts the number k 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.

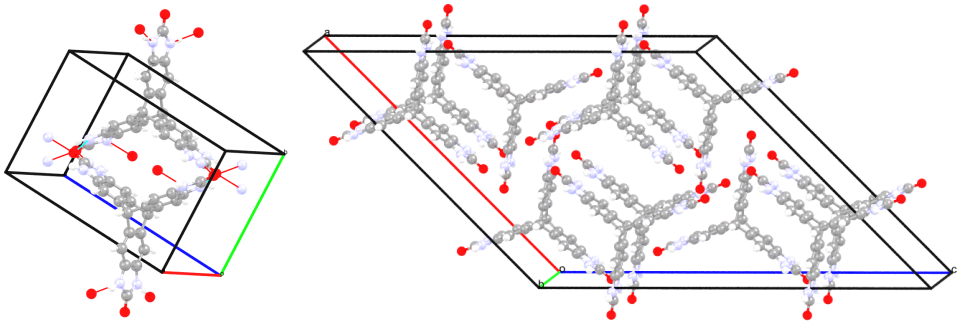

FIG. SM6. *Crystals 14, 15 from the database of 5679 simulated crystals reported in [SM7] consist of identical T2 molecules and have very different Crystallographic Information Files (with different motifs in unit cells of distinct shapes) but are nearly identical under isometry.*

- `CSD_duplicates_example.py` computes and compares the PDDs of isometric crystals from the CSD discussed in section SM1, 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`.

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.

SM3. Detailed proofs of auxiliary lemmas and Theorem 4.2. This appendix proves Lemmas 3.4-3.5, which were used in Theorem 3.6, and Theorem 4.2.

Proof of Lemma 3.4. Intersect the three regions $U^-(p;r) \subset C(p;r) \subset U^+(p;r)$ with S in $\mathbb{R}^n$ and count all points: $|S \cap U^-(p;r)| \leq |S \cap C(p;r)| \leq |S \cap U^+(p;r)|$.

The union $U^-(p;r)$ consists of $m^-(p;r) = \dfrac{\mathrm{vol}[U^-(p;r) \cap R^l]}{\mathrm{vol}[U]}$ shifted cells, which all have the same volume $\mathrm{vol}[U \cap R^l]$. Since $|S \cap U| = m$, we get $|S \cap U^-(p;r)| = \dfrac{\mathrm{vol}[U^-(p;r) \cap R^l]}{\mathrm{vol}[U]} m$. Similarly, we count all points of S in the upper union as follows:

$|S \cap U^+(p;r)| = \dfrac{\mathrm{vol}[U^+(p;r) \cap R^l]}{\mathrm{vol}[U]} m$. The bounds for $|S \cap C(p;r)|$ become

$$\frac{\mathrm{vol}[U^-(p;r) \cap R^l]}{\mathrm{vol}[U]} m \le |S \cap C(p;r)| \le \frac{\mathrm{vol}[U^+(p;r) \cap R^l]}{\mathrm{vol}[U]} m,$$

which proves the internal inequalities $m^-(p;r)m \le |S \cap C(p;r)| \le m^+(p;r)m$. Then

$$\mathrm{vol}[U^-(p;r) \cap R^l] \le \frac{\mathrm{vol}[U \cap R^l]}{m} |S \cap C(p;r)| \le \mathrm{vol}[U^+(p;r) \cap R^l].$$

For the width w of the unit cell U, the smaller cylinder $C(p; r-w)$ is completely contained within the lower union $U^-(p;r)$. Indeed, if $|\vec{q} - \vec{p}| \le r - w$, then $q \in U + \vec{v}$ for some $\vec{v} \in \Lambda$. Then $(U + \vec{v})$ is covered by the cylinder $C(q;w)$, hence by $C(p;r)$ due to the triangle inequality. The inclusion $C(p; r-w) \subset U^-(p;r)$ implies the lower bound for the volumes: $(r-w)^l V_l = \mathrm{vol}[C(p;r-w) \cap R^l] \le \mathrm{vol}[U^-(p;r) \cap R^l]$, where V_l is the unit ball volume in $\mathbb{R}^l$. Then $\dfrac{(r-w)^l V_l}{\mathrm{vol}[U \cap R^l]} \le \dfrac{\mathrm{vol}[U^-(p;r) \cap R^l]}{\mathrm{vol}[U \cap R^l]} = m^-(p;r)$, which implies the first required inequality in the lemma:

$$\left(\frac{r-w}{\mathrm{PPC}(S)} \right)^l = \frac{(r-w)^l m V_l}{\mathrm{vol}[U \cap R^l]} \le \frac{\mathrm{vol}[U^-(p;r) \cap R^l]}{\mathrm{vol}[U \cap R^l]} m = m^-(p;r) m.$$

The last required inequality is proved similarly by using $U^+(p;r) \subset C(p; r+w)$. ∎

Proof of Lemma 3.5. Let $q \in S$ be a k-th neighbor of p in S. There can be several points $q \in S$ at the distance $|q-p| = d_k(S;p)$ but the argument below works for any q. The closed cylinder $C(p;r)$ with $r = d_k(S;p)$ contains the k-th neighbor q of p and hence has more than k points (including p) from S. The upper bound of Lemma 3.4 for $r = d_k(S;p)$ implies that $k < |S \cap C(p;r)| \le \dfrac{(r+w)^l}{(\mathrm{PPC}(S))^l}$. Taking the l-th roots gives $\sqrt[l]{k} < \dfrac{r+w}{\mathrm{PPC}(S)}$, so $r = d_k(S;p) > \mathrm{PPC}(S)\sqrt[l]{k} - w$.

For any radius r such that $\sqrt{r^2 + h^2} < d_k(S;p)$, the closed cylinder $C(p;r)$ contains only points at a maximum distance $\sqrt{r^2+h^2}$ from p. Then $C(p;r)$ does not include the k-th neighbor q of p and hence contains at most k points (including p) from S. The lower bound of Lemma 3.4 for $r < \sqrt{(d_k(S;p))^2 - h^2}$ implies that $\dfrac{(r-w)^l}{(\mathrm{PPC}(S))^l} \le |S \cap C(p;r)| \le k$. Since the inequality $\dfrac{(r-w)^l}{(\mathrm{PPC}(S))^l} \le k$ holds for the constant upper bound k and any radius $r < \sqrt{(d_k(S;p))^2 - h^2}$, the same inequality holds for the radius $r = \sqrt{(d_k(S;p))^2 - h^2}$. Then $\dfrac{r-w}{\mathrm{PPC}(S)} \le \sqrt[l]{k}$,

$$r = \sqrt{(d_k(S;p))^2 - h^2} \le \mathrm{PPC}(S)\sqrt[l]{k} + w, \quad d_k(S;p) \le \sqrt{(\mathrm{PPC}(S)\sqrt[l]{k} + w)^2 + h^2}.$$

EXAMPLE SM3.1 (stronger asymptotic $\mathrm{ADA}_k(S) \to 0$ as $k \to +\infty$ for $\mathbb{Z}^n$). *The survey [SM5] describes progress on the generalized Gauss circle problem expressing the number of points from the cubic lattice $\mathbb{Z}^n$ within a ball of a radius r as $k = V_n r^n - O(r^{\alpha_n + \varepsilon})$ for any $\varepsilon > 0$, where $\alpha_n < n-1$ for $n \ge 2$. The cubic lattice has* $\mathrm{PPC}(\mathbb{Z}^n) =$

$1/\sqrt[n]{V_n}$. *Let d_k denote the distance from the origin 0 to its k-th neighbor in $\mathbb{Z}^n$. Then* $k = V_n d_k^n - O(d_k^{\alpha_n+\varepsilon})$, *so* $d_k = \sqrt[n]{\dfrac{k+O(d_k^{\alpha_n+\varepsilon})}{V_n}} = \mathrm{PPC}(\mathbb{Z}^n)\sqrt[n]{k+O(d_k^{\alpha_n+\varepsilon})}$. *Then*

$$\frac{\mathrm{ADA}_k(\mathbb{Z}^n)}{\mathrm{PPC}(\mathbb{Z}^n)} = \frac{d_k}{\mathrm{PPC}(\mathbb{Z}^n)} - \sqrt[n]{k} = \sqrt[n]{k+O(d_k^{\alpha_n+\varepsilon})} - \sqrt[n]{k} = \frac{O(d_k^{\alpha_n+\varepsilon})}{P_n(\sqrt[n]{k+O(d_k^{\alpha_n+\varepsilon})},\ \sqrt[n]{k})},$$

where P_n is a homogeneous polynomial of degree $n-1$, e.g. $P_2(x,y) = x+y$, $P_3(x,y) = x^2+xy+y^2$. Because the numerator has the power $\alpha_n < n-1$ of $d_k = O(\sqrt[n]{k})$ for $n \geq 2$, the final expression and hence $\mathrm{ADA}_k(\mathbb{Z}^n)$ have limit 0 as $k \to +\infty$.

Theorem 4.1 will be proved similar to [SM9, Theorem 13] by Lemmas SM3.2, SM3.3, SM3.4. Partial cases of Lemmas SM3.2 and SM3.3 appeared for $l = n$ in [SM4, Lemma 2] and for $\mathbb{R}^n$ in [SM9, Lemma 8], respectively.

LEMMA SM3.2 (common lattice). *Let l-periodic point sets $S, Q \subset \mathbb{R}^n$ have a bottleneck distance $d_B(S,Q) < \min\{r(S), r(Q)\}$. Then S, Q have a common lattice Λ with a unit cell U such that $S = \Lambda + (U \cap S)$ and $Q = \Lambda + (U \cap Q)$.*

Proof of Lemma SM3.2. 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 l-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 l-dimensional 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 $p \in \Lambda(S) \subset S$ whose all integer multiples $k\vec{p} \in \Lambda(S)$ do not belong to $\Lambda(Q)$ for $k \in \mathbb{Z} - \{0\}$. Any such multiple $k\vec{p} \in \Lambda(S) \subset S$ can be translated by a vector of $\Lambda(Q)$ to a point $t(k)$ in the unit cell $U(Q)$ so that $k\vec{p} \equiv t(k) \pmod{\Lambda(Q)}$. Since the cell $U(Q)$ contains infinitely many points $t(k)$ for $k \neq 0$, one can find a pair $t(i) \neq t(j)$ 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 $\Lambda(Q)$.

$$t(i+m(j-i)) \equiv (i+m(j-i))\vec{p} = i\vec{p} + m(j\vec{p} - i\vec{p}) \equiv t(i) + m(t(j) - t(i)).$$

These points for $m \in \mathbb{Z}$ lie in a straight line with gaps $|t(j) - t(i)| < \delta$. 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 $|t(j) - t(i)| < \delta = r(Q) - d_B(S,Q)$, there is $m \in \mathbb{Z}$ such that $t(i) + m(t(j) - t(i))$ is outside the union $Q + \bar{B}(0; d_B(S,Q))$ of all these smaller balls. Then $t(i) + m(t(j) - t(i))$ 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 $(i + m(j-i))\vec{p} \in \Lambda(S) \subset S$, which is equivalent to $t(i) + m(t(j) - t(i))$ modulo $\Lambda(Q)$, has a distance more than $d_B(S,Q)$ from any point of Q. This conclusion contradicts the definition of $d_B(S,Q)$. ∎

LEMMA SM3.3 (perturbed distances). *For some $\varepsilon > 0$, let $g : S \to Q$ be a bijection between any discrete sets in a space X with a metric d_X such that $d_X(g(p), p) \leq \varepsilon$ for all $p \in S$. Then, for any $i \geq 1$, let $p_i \in S$, $t_i \in Q$ be i-th nearest neighbors of points $p \in S$, $t = g(p) \in Q$, respectively. Then the distances from the points p, t to their i-th neighbors p_i, t_i in X are 2ε-close to each other, i.e. $|d_X(p, p_i) - d_X(t, t_i)| \leq 2\varepsilon$.*

Proof of Lemma SM3.3. Shifting the point $g(p)$ back to p, assume that $p = g(p)$ is fixed and all other points change their positions by at most 2ε. Assume by contradiction that the distance from p to its new i-th neighbor t_i is less than $d_X(p, p_i) - 2\varepsilon$. Then all first new i neighbors $t_1, \dots, t_i \in Q$ of p belong to the open ball with the center p and the radius $d_X(p, p_i) - 2\varepsilon$. Because the bijection g shifted every point $t_1, \dots, t_i$ by at most 2ε, their preimages $g^{-1}(t_1), \dots, g^{-1}(t_i)$ belong to the open ball with the center p and the radius $d_X(p, p_i)$. Then the i-th neighbor of p within S is among these i preimages, i.e. the distance from p to its i-th nearest neighbor should be strictly less than the assumed value $d_X(p, p_i)$. We similarly get a contradiction by assuming that the distance from p to its new i-th neighbor t_i is more than $d_X(p, p_i) + 2\varepsilon$. ∎

LEMMA SM3.4 (perturbed distance vectors). *For $\varepsilon > 0$, let $g : S \to Q$ be a bijection between any discrete sets in a space X with a metric d_X so that $d_X(g(p), p) \le \varepsilon$ for all $p \in S$. Then g changes the vector $\vec{R}(S, p) = (d_X(p, p_1), \dots, d_X(p, p_k))$ of the first k minimum distances from any point $p \in S$ to its k nearest neighbors $p_1, \dots, p_k \in S$ by at most $2\varepsilon\sqrt[q]{k}$ in the distance L_q. So if $t_1, \dots, t_k \in Q$ are k nearest neighbors of $t = g(p)$ within Q and $\vec{R}(Q, t) = (d_X(t, t_1), \dots, d_X(t, t_k))$ is the vector of the first k minimum distances from $t = g(p)$ in Q, then the L_∞-distance $|\vec{R}(S, p) - \vec{R}(Q, t)|_\infty \le 2\varepsilon\sqrt[q]{k}$.*

Proof of Lemma SM3.4. By Lemma SM3.3 every coordinate of $\vec{R}(S, p)$ changes by at most 2ε. Hence the distance $L_q(\vec{R}(S, p), \vec{R}(Q, t)) \le \big(\sum\limits_{i=1}^{k} (2\varepsilon)^q\big)^{1/q} = 2\varepsilon\sqrt[q]{k}$. ∎

Proof of Theorem 4.2. The bottleneck distance between the given sets $S, Q \subset X$ is $d_B(S, Q) = \inf\limits_{g: S \to Q} \sup\limits_{p \in S} d_X(g(p), p)$. Then for any $\delta > 0$ there is a bijection $g : S \to Q$ such that $\sup\limits_{p \in S} d_X(g(p), p) \le d_B(S, Q) + \delta$. If the given sets S, Q are finite, one can set $\delta = 0$. Indeed, there are only finitely many bijections $g : S \to Q$, hence the infimum in the definition above is achieved for one of these bijection g.

(a) For any discrete sets $S, Q \subset X$ be with finite subsets M, T of the same number m of points, respectively, we use the notations of Definition 3.1. The given 1-1 perturbation $g : S \to Q$ defines the simplest 1-1 flow from the row of any $p \in M$ in the matrix $D(S, M; k)$ to the row of $g(p) \in T$ in $D(Q, T; k)$ by setting $f_{ii} = \frac{1}{m}$ and $f_{ij} = 0$ for $i \ne j$, where $i, j = 1, \dots, m$. All rows of $D(S, M; k)$ that are identical to each other are collapsed to a single row, similarly for $D(Q, T; k)$. By summing up weights of all collapsed rows, the above flow induces a flow from all distance vectors in $\mathrm{PDD}(S, M; k)$ to all distance vectors in $\mathrm{PDD}(Q, T; k)$.

Then $\mathrm{EMD}_q(\mathrm{PDD}(S, M; k), \mathrm{PDD}(Q, T; k)) \le \frac{1}{m} \sum\limits_{i=1}^{m} L_q(\vec{R}_i(S), \vec{R}_i(Q))$, because EMD_q minimizes the cost over all flows in Definition 4.2. The upper bound $L_q(\vec{R}_i(S), \vec{R}_i(Q)) \le 2(\varepsilon + \delta)\sqrt[q]{k}$ from Lemma SM3.4 implies that

$$\mathrm{EMD}_q(\mathrm{PDD}(S, M; k), \mathrm{PDD}(Q, T; k)) \le \frac{1}{m} \sum_{i=1}^{m} 2(\varepsilon + \delta)\sqrt[q]{k} = 2(\varepsilon + \delta)\sqrt[q]{k},$$

which holds for any small $\delta > 0$. By taking the limit for $\delta \to 0$, we get the required upper bound $\mathrm{EMD}_q(\mathrm{PDD}(S, M; k), \mathrm{PDD}(Q, T; k)) \le 2\varepsilon\sqrt[q]{k}$.

(b) In the l-periodic case by Lemma SM3.2, the given sets S, Q should have a common l-dimensional lattice Λ. Any primitive cell U of Λ is a common unit cell

of S, Q, i.e. $S = \Lambda + (S \cap U)$ and $Q = \Lambda + (Q \cap U)$, so $\mathrm{PPC}(S) = \mathrm{PPC}(Q)$. Then all L_∞ distances between rows in $\mathrm{PDA}(S;k), \mathrm{PDA}(Q;k)$ are the same as between the corresponding rows in $\mathrm{PDD}(S;k), \mathrm{PDD}(Q;k)$, see Definition 3.7. Hence $\mathrm{EMD}_q(\mathrm{PDA}(S;k), \mathrm{PDA}(Q;k)) = \mathrm{EMD}_q(\mathrm{PDD}(S;k), \mathrm{PDD}(Q;k)) \le 2\varepsilon \sqrt[q]{k}$ by (a).

The remaining inequality follows from the PDA case. Indeed, each element of $\mathrm{PND}(S;k)$ in a row i and a column $j = 1, \dots, k$ is obtained from the corresponding element of $\mathrm{PDA}(S;k)$ by dividing by $\mathrm{PPC}(S)\sqrt[q]{j} \ge \mathrm{PPC}(S)$. Then each distance L_q between corresponding rows in $\mathrm{PND}(S;k)$, $\mathrm{PND}(Q;k)$ is at least $\mathrm{PPC}(S)$ times smaller than between the same rows in $\mathrm{PDA}(S;k)$, $\mathrm{PDA}(Q;k)$. Then

$$\mathrm{EMD}_q(\mathrm{PND}(S;k), \mathrm{PND}(Q;k)) \le \frac{\mathrm{EMD}_q(\mathrm{PDA}(S;k), \mathrm{PDA}(Q;k))}{\mathrm{PPC}(S)} \le \frac{2\varepsilon \sqrt[q]{k}}{\mathrm{PPC}(S)}. \quad \square$$

Proof of Theorem 4.4. Considering $\mathrm{PDD}(S;k)$ as a weighted distribution of rows, $\mathrm{AMD}(S;k)$ is its centroid from [SM3, section 3]. The argument below follows the proof of [SM3, Theorem 1] for $q = +\infty$ and similarly works for other invariants in parts (b,c). In the notations of Definition 4.1, we use the inequality $||\vec{u}||_q + ||\vec{v}||_q|| \ge ||\vec{u}+\vec{v}||_q$ for the q-norm $||\vec{v}||_q = \big(\sum\limits_{i=1} |v_i|^q\big)^{1/q}$ of the Minkowski metric L_q as follows:

$$\mathrm{EMD}_q(\mathrm{PDD}(S;k), \mathrm{PDD}(Q;k)) = \sum_{i=1}^{m(S)} \sum_{j=1}^{m(Q)} f_{ij} L_q(\vec{R}_i(S), \vec{R}_j(Q)) =$$

$$\sum_{i=1}^{m(S)} \sum_{j=1}^{m(Q)} ||f_{ij}\big(\vec{R}_i(S) - \vec{R}_j(Q)\big)||_q \ge || \sum_{i=1}^{m(S)} \sum_{j=1}^{m(Q)} f_{ij}(\vec{R}_i(S) - \vec{R}_j(Q))||_q =$$

$$|| \sum_{i=1}^{m(S)} \big(\sum_{j=1}^{m(Q)} f_{ij}\vec{R}_i(S)\big) - \sum_{j=1}^{m(Q)} \big(\sum_{i=1}^{m(S)} f_{ij}\vec{R}_j(Q)\big)||_q =$$

$$|| \sum_{i=1}^{m(S)} w_i(S)\vec{R}_i(S) - \sum_{j=1}^{m(Q)} w_j(Q)\vec{R}_j(Q)||_q = L_q(\mathrm{AMD}(S;k), \mathrm{AMD}(Q;k)). \quad \square$$

REFERENCES

[SM1] O. Anosova and V. Kurlin, *An isometry classification of periodic point sets*, in Proceedings of Discrete Geometry and Mathematical Morphology, 2021, pp. 229–241.

[SM2] A. P. Bartók, R. Kondor, and G. Csányi, *On representing chemical environments*, Physical Review B—Condensed Matter and Materials Physics, 87 (2013), p. 184115.

[SM3] S. Cohen and L. Guibas, *The Earth Mover's Distance: Lower bounds and invariance under translation*, tech. report, Stanford University, 1997.

[SM4] H. Edelsbrunner, T. Heiss, V. Kurlin, P. Smith, and M. Wintraecken, *The density fingerprint of a periodic point set*, in SoCG, 2021, pp. 32:1–32:16.

[SM5] A. Ivic, E. Krätzel, M. Kühleitner, and W. Nowak, *Lattice points in large regions and related arithmetic functions: recent developments in a very classic topic*, Publications of the Scientific Society at the Johann Wolfgang Goethe University, (2006), pp. 89–128.

[SM6] A. Patterson, *Ambiguities in the X-ray analysis*, Phys. Rev., 65 (1944), pp. 195–201.

[SM7] A. Pulido et al., *Functional materials discovery using energy–structure maps*, Nature, 543 (2017), pp. 657–664.

[SM8] M. W. Terban and S. J. Billinge, *Structural analysis of molecular materials using the pair distribution function*, Chemical Reviews, 122 (2021), pp. 1208–1272.

[SM9] 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.