MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **87** (2022) 529-559
ISSN: 0340-6253
doi: 10.46793/match.87-3.529W

Average Minimum Distances of Periodic Point Sets – Foundational Invariants for Mapping Periodic Crystals

Daniel Widdowson[1], **Marco M. Mosca**[1], **Angeles Pulido**[2], **Andrew I. Cooper**[3], **Vitaliy Kurlin**[*,1,3]

[1]*Computer Science, University of Liverpool, Liverpool, L69 3BX, UK*
d.e.widdowson@liverpool.ac.uk, m.m.mosca@liverpool.ac.uk

[2]*Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, UK*
apulido@ccdc.cam.ac.uk

[3]*Materials Innovation Factory, University of Liverpool, Liverpool, L69 3NY, UK*
aicooper@liverpool.ac.uk, vitaliy.kurlin@liverpool.ac.uk

(Received July 16, 2021)

Abstract

The fundamental model of any solid crystalline material (crystal) at the atomic scale is a periodic point set. The strongest natural equivalence of crystals is rigid motion or isometry that preserves all inter-atomic distances. Past comparisons of periodic structures often used manual thresholds, symmetry groups and reduced cells, which are discontinuous under perturbations or thermal vibrations of atoms. This work defines the infinite sequence of continuous isometry invariants (Average Minimum Distances) to progressively capture distances between neighbors. The asymptotic behaviour of the new invariants is theoretically proved in all dimensions for a wide class of sets including non-periodic. The proposed near linear time algorithm identified all different crystals in the world's largest Cambridge Structural Database within a few hours on a modest desktop. The ultra fast speed and proved continuity provide rigorous foundations to continuously parameterise the space of all periodic crystals as a high-dimensional extension of Mendeleev's table of elements.

*Corresponding author.

1 Motivations, problem statement and main results

A periodic *lattice* consists of all integer linear combinations of a basis in Euclidean space $\mathbb{R}^n$ whose vectors span a *unit cell* U. More generally, a *periodic point set* is the Minkowski sum $\Lambda + M = \{\vec{u} + \vec{v} : u \in \Lambda, v \in M\}$ of a *lattice* Λ and a *motif* M, which is a finite set of points in a unil cell U [42], see Fig. 1. Since atomic nuclei are well-defined physical objects, their geometric positions provide the most important information about a periodic crystal, while chemical bonds often depend on manually chosen thresholds for distances and angles. Though chemical elements can be added to points as labels, a pure point set representation allows us to study all periodic crystals together in a common space.

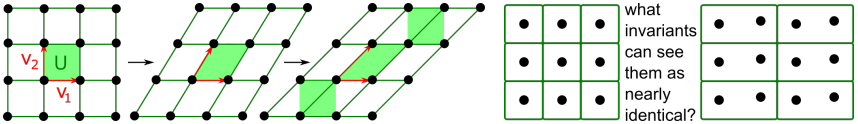

Figure 1. Left: a continuous deformation of lattices, all periodic sets form a continuous space. **Right**: many invariants such as symmetry groups are discontinuous under any perturbations that change a primitive cell.

Though the concept of a crystal pattern [11, section 8.1.4] covers a periodic point set, such patterns were studied up to coarse equivalences depending on compositions or threshold parameters. For example, [26, Table 1] calls the crystals of FeS_2 and PtP_2 equivalent by all past relations, though their cubic lattices have different sizes [36] and can be distinguished up to isometry.

An *isometry* of Euclidean $\mathbb{R}^n$ is any map that preserves inter-point distances. Any orientation-preserving isometry can be realised as a continuous rigid motion, for example a composition of translations and rotations in $\mathbb{R}^3$. The isometry equivalence is the strongest and most natural for rigid crystals. We consider orientation-reversing isometries including reflections for simplicity since a sign of orientation can be added to isometry invariants.

The recent work [17] initiated a classification of periodic point sets up to isometry. An isometry classification of periodic point sets is highly non-trivial both mathematically and computationally for the following four reasons.

Firstly, since any lattice can be generated by infinitely many different bases or *primitive* unit cells of a minimal volume, a representation of a periodic point set as a sum $S = \Lambda + M$ of a lattice and a motif is highly ambiguous. All attempts to find a canonical basis or a

reduced cell of a lattice are discontinuous under perturbations [17, section 1], which will be formalised by Theorem 15 in section 7. Fig. 1 shows how all lattices (similarly any periodic point sets) can be continuously deformed into each other. Hence all isometry classes of periodic point sets form a continuous space. The last two periodic sets in Fig. 1 are nearly identical but their similarity is hard to quantify without using thresholds. These sets substantially differ by symmetry groups and unit cell volumes but cannot be distinguished by density (the number of points per volume).

Secondly, even for a fixed lattice basis, shifting points within a unit cell changes their coordinates in the cell basis. Similarly, isometries preserve a rigid structure, but change a basis representation in a fixed coordinate system.

Thirdly, periodic structures were traditionally studied by symmetries and other group-theoretic invariants [19], which break down under almost any perturbations, see the last pair in Fig. 1 Such discrete invariants cut the continuous space of periodic point sets into separate strata. This discontinuous stratification is the main obstacle for understanding transitions between crystal phases and for detecting nearly identical structures, which are common due to inevitable noise in measurements or approximations in simulations.

Fourthly, the world's largest Cambridge Structural Database has more than 1.1M known structures kept as pairs (unit cell, motif), which should be converted into continuous invariants for reliable comparisons. Hence we need fast invariants that can be sequentially extended to distinguish all real structures.

Problem 1 (fast invariants)**.** *Find continuous isometry invariants of periodic point sets that can be computed in a near linear time in input sizes.* ■

Fig. 2 shows how infinitely many pairs (cell, motif) form a single class of lattices represented by a colored box in the middle picture illustrating the continuous space of all lattices. Consider the *crystal geometry map* {crystals} $\to$ {periodic point sets} representing any atom or ion by its atomic centre.

It is physically reasonable that the above map is injective meaning that any non-isometric crystals remain non-isometric after forgetting chemical types or charges, because different atoms or ions can be distinguished by their distances to neighbors in real crystals. We experimentally confirm this injectivity by computing new isometry invariants for all known molecular organic crystals.

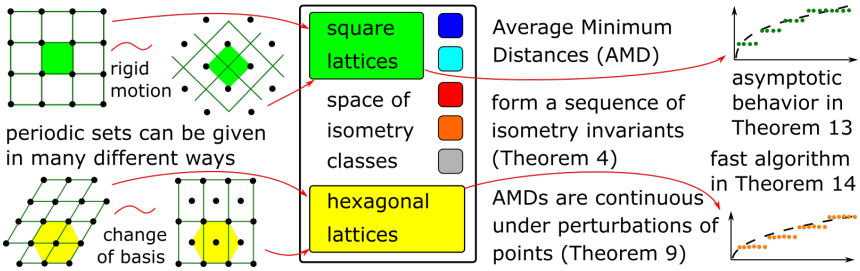

Figure 2. Ambiguous representations of periodic point sets by (unit cell, motif) are converted into continuous isometry invariants. Each colored box in the middle picture represents one isometry class of lattices.

Section 2 reviews key concepts and past results on isometry classifications. Section 3 introduces the Average Minimum Distances (AMD). Section 4 proves the Lipschitz continuity of AMD under perturbations of points. Section 5 describes the asymptotic behaviour of the infinite AMD sequence. A near linear time algorithm in section 6 computes AMD of real structures in milliseconds on a modest desktop. Section 7 discusses how AMD detected previously unknown duplicates and distinguished hundreds of thousands of other existing crystals in the Cambridge Structural Database (CSD).

2 Periodic point sets and a review of past work

Any point $p \in \mathbb{R}^n$ can be represented by the vector $\vec{p}$ from the origin of $\mathbb{R}^n$ to the point p, so p and $\vec{p}$ can be used interchangeably, though $\vec{p}$ can be drawn at any initial point. The *Euclidean* distance between points $p, q \in \mathbb{R}^n$ is $|p-q|$.

Definition 2 (a lattice Λ, a motif, a unit cell, a periodic point set S)**.** *Let vectors* $\vec{v}_1, \dots, \vec{v}_n$ *form a linear* basis *in* $\mathbb{R}^n$ *so that if* $\sum_{i=1}^{n} c_i \vec{v}_i = \vec{0}$ *for some real* c_i*, then all* $c_i = 0$*. Then a* lattice Λ *in* $\mathbb{R}^n$ *consists of all linear combinations* $\sum_{i=1}^{n} c_i \vec{v}_i$ *with integer coefficients* $c_i \in \mathbb{Z}$*. A* motif M *is a finite set of points* $p_1, \dots, p_m$ *in the* unit cell $U(\vec{v}_1, \dots, \vec{v}_n) = \left\{ \sum_{i=1}^{n} c_i \vec{v}_i \,:\, c_i \in [0,1) \right\}$*, which is the parallepiped spanned by* $\vec{v}_1, \dots, \vec{v}_n$*. A* periodic point set $S \subset \mathbb{R}^n$ *is the* Minkowski sum $S = \Lambda + M = \{\vec{u} + \vec{v} \,:\, u \in \Lambda, v \in M\}$*, so* S *is a finite union of translates of the lattice* Λ*. A unit cell* U *is* primitive *if* S *remains invariant under shifts by vectors only from* Λ *generated by* U *or* $\vec{v}_1, \dots, \vec{v}_n$*.* ■

Any lattice Λ can be considered as a periodic set with a 1-point motif $M = \{p\}$. This

point p can be arbitrarily chosen in a unit cell U. The lattice translate $\Lambda+\vec{p}$ is considered as a lattice, because p can be the origin of $\mathbb{R}^n$. The periodic sets in the top left part of Fig. 2 represent isometric square lattices, though the former has a point p at a corner of a unit cell U and the latter has p in the center of U. The periodic sets in the bottom left part of Fig. 2 represent isometric hexagonal lattices, because every black point has exactly six nearest neighbors that form a regular hexagon.

A lattice Λ of a periodic set $S = M + \Lambda \subset \mathbb{R}^n$ is not unique in the sense that S can be generated by a sublattice of Λ and a motif larger than M. If U is any unit cell of Λ, the sublattice 2Λ has the 2^n times larger unit cell 2^nU (twice larger along each of n basis vectors of U), hence contains 2^n times more points than M. Such an extended cell 2^nU is superfluous, because S is invariant under translations along not only integer linear combinations $\sum_{i=1}^{n} c_i\vec{v}_i$ with $c_i \in \mathbb{Z}$, but also along vectors with coefficients $c_i \in \frac{1}{2}\mathbb{Z}$.

Now we discuss the past work on comparing finite and periodic sets up to isometry. The wider area of Euclidean distance geometry is reviewed in [25]. The full distribution of all pairwise Euclidean distances $|a-b|$ between points a, b in a finite set $S \subset \mathbb{R}^m$ is a well-known isometry invariant. This invariant is complete or injective for finite sets in general position [8] in the sense that almost any finite set S can be uniquely reconstructed up to isometry from the set of all distances between points of S. The left hand side pictures in Fig. 3 show the counter-example pair T, K to the full completeness.

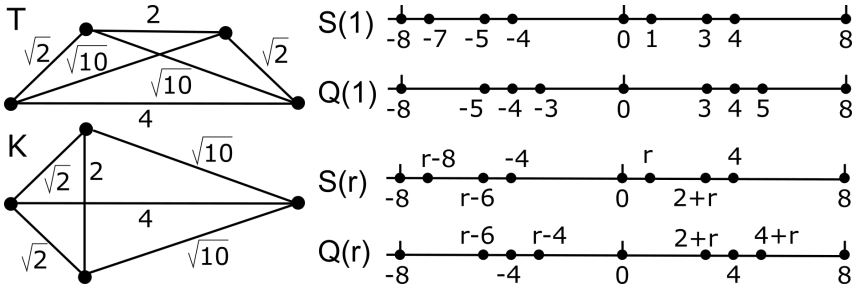

Figure 3. Non-isometric sets that cannot be distinguished by past invariants. **Left**: $T, K \subset \mathbb{R}^2$ have the same pairwise distances $\{2, \sqrt{2}, \sqrt{2}, \sqrt{10}, \sqrt{10}, 4\}$. **Right**: periodic sets $S(r) = \{0, r, r+2, 4\} + 8\mathbb{Z}$ and $Q(r) = \{0, r+2, 4, r+4\}+8\mathbb{Z}$ for $0 < r \le 1$ have the same Patterson function [32, p. 197, Fig. 2]. All these pairs are distinguished by AMD in section 3.

The isometry classification of finite point sets was algorithmically resolved by [1, Theorem 1] saying that an existence of an isometry between m-point sets in $\mathbb{R}^n$ can be checked in time $O(m^{n-2}\log m)$. For finite sets, Average Minimum Distances are similar to Mémoli's seminal work on *distributions of distances* [28], also known as *shape distributions* [5,22,27,31]. All algorithms for finite sets cannot be easily extended to periodic sets by fixing a unit cell, because any such reduced cell is discontinuous under perturbations by Theorem 15 in section 7.

An isometry classification of periodic point sets is already non-trivial in dimension 1. Complete invariants were found for 1D sets with only integer (or rational) coordinates [23]. Three-dimensional analogues of their invariants form the *Patterson function*, whose peaks correspond to inter-point vectors [21]. Periodic sets that cannot be distinguished by Patterson functions are called *homometric*, see Fig. 3 and more details in appendix B. The 4-point non-isometric sets T, K have periodic versions $S(1) = \{0,1,3,4\} + 8\mathbb{Z}$ and $Q(1) = \{0,3,4,5\}+8\mathbb{Z}$ in Fig. 3. Even more general homometric sets $S(r), Q(r)$ depending on $0 < r \le 1$ will be distinguished by the simplest new invariant AMD_1 in section 3.

More recently, for any periodic point set $S \subset \mathbb{R}^n$ with a motif M in a unit cell U, Edelsbrunner et al. [17] introduced the density functions $\psi_k(t)$ for any integer $k \ge 1$. The k-th *density function* $\psi_k(t)$ is the total volume of the regions within the unit cell U covered by exactly k balls $B(p;t)$ with a radius $t \ge 0$ and centres at points $p \in M$, divided by the unit cell volume $\mathrm{Vol}[U]$. The density function $\psi_k(t)$ was proved to be invariant under isometry, continuous under perturbations, complete for periodic sets in general position in $\mathbb{R}^3$, and computable in time $O(mk^3)$, where m is the motif size of S. Section 5 in [17] gives the counter-example to completeness: the 1-dimensional periodic sets $S_{15} = X + Y + 15\mathbb{Z}$ and $Q_{15} = X - Y + 15\mathbb{Z}$ for $X = \{0,4,9\}$ and $Y = \{0,1,3\}$, which appeared earlier in [23, section 4]. These non-isometric sets have the same density functions for all $k \ge 1$, see [3, Example 11], and are distinguished by AMD_3 in Example (5b).

The latest invariant isoset [4] reduces the isometry classification of all periodic point sets to a fintie collection of isometry classes of α-clusters around points in a motif at a certain radius α, which was motivated by the seminal work of Dolbilin with co-authors about Delone sets [7, 15, 16]. Checking if two isosets coincide needs a cubic algorithm, which is not yet implemented. Running times of algorithms are compared in section 7.

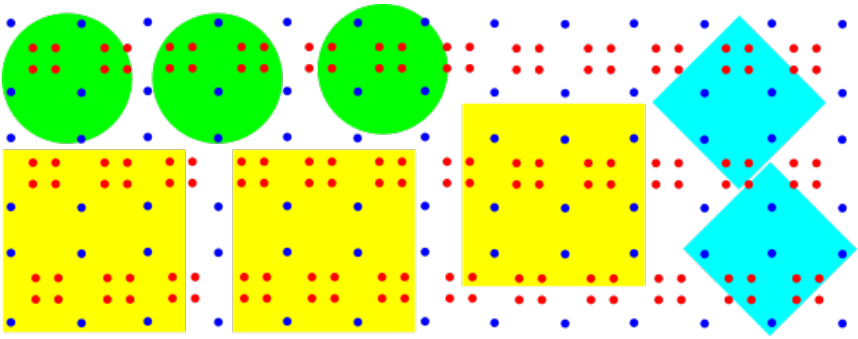

Figure 4. A periodic point set cannot be reliably represented by its finite subsets without a lattice. For example, disks and rectangular boxes of the same size can contain non-isometric finite subsets, which can discontinuously change under perturbations of points or cut-off parameters.

Other distance-based invariants widely used in applications are the radial distribution function [37], which additionally depends on a cut-off radius around atoms. Fig. 4 illustrates the key obstacle to represent a periodic point set by its finite subset of points. Even if such a subset is sufficiently large to cover a big extended cell, its content is highly variable. In the simpler cases of lattices, two metric functions on arbitrary lattices were defined by using Voronoi domains [30], though their computation was only approximate.

A similar attempt to tackle the discontinuity under perturbations uses a pseudo-symmetry approach [44] putting two structures in the same symmetry class if their points can be matched by shifts up to a small distance ε. Any tolerance $\varepsilon > 0$ leads to a logical contradiction in classifications, because any wildly different periodic point sets can be connected by a long chain of small perturbations within any small $\varepsilon > 0$. If we accept the transitivity axiom of an equivalence relation, which justifies a partition into well-defined classes, all above sets become equivalent to each other, so the classification is trivial.

A better way to deal with perturbations is to continuously quantify noise by a metric between isometry classes of periodic point sets. Such a metric d should satisfy three axioms, most importantly the first axiom saying that $d(S,Q) = 0$ *only* if S, Q are isometric. Most past attempts use a metric between descriptors or invariants of periodic structures [43], which can satisfy the above axiom only if the invariant is complete.

3 Average Minimum Distances and their invariance

This section introduces Average Minimum Distances in Definition 3 and proves their isometry invariance in Theorem 4. For a lattice $\Lambda \subset \mathbb{R}^n$, $\mathrm{AMD}_k(\Lambda)$ can be defined as the distance from a fixed point $p \in L$ to its k-th nearest neighbor in Λ, see Fig. 5.

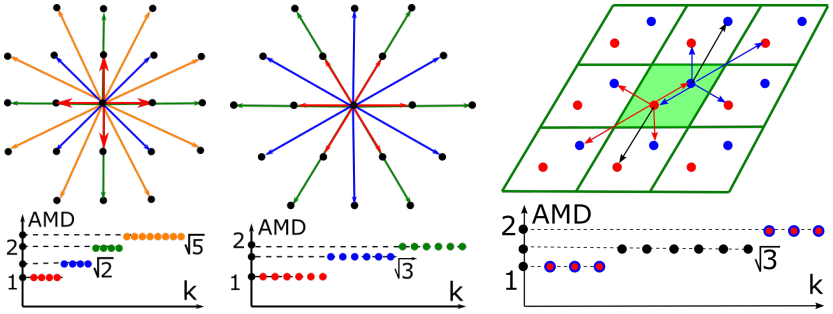

Figure 5. Left: in the square lattice, the k-th neighbors of the origin and corresponding AMD_k are shown in the same color, for example the shortest axis-aligned distances $\mathrm{AMD}_1 = \cdots = \mathrm{AMD}_4 = 1$ are in red, the longer diagonal distances $\mathrm{AMD}_5 = \cdots = \mathrm{AMD}_8 = \sqrt{2}$ are in blue. **Middle**: in the hexagonal lattice, the shortest distances are in red: $\mathrm{AMD}_1 = \cdots = \mathrm{AMD}_6 = 1$. **Right**: a honeycomb periodic point set has a 2-point motif and the first average distances $\mathrm{AMD}_1 = \mathrm{AMD}_2 = \mathrm{AMD}_3 = 1$.

Definition 3 (Average Minimum Distances AMD)**.** *Let a periodic point set $S = \Lambda + M \subset \mathbb{R}^n$ have points $p_1, \dots, p_m$ in a primitive unit cell. For a fixed integer $k \geq 1$ and $i = 1, \dots, m$, the i-th row of the $m \times k$ matrix $D(S;k)$ consists of the ordered Euclidean distances $d_{i1} \leq \cdots \leq d_{ik}$ measured from the point p_i to its first k nearest neighbors within the infinite set S, see Fig. 5. The* Average Minimum Distance $\mathrm{AMD}_k(S) = \dfrac{1}{m}\sum_{i=1}^{m} d_{ik}$ *equals the average of the k-th column in the matrix $D(S;k)$ of distances to neighbors.* ■

Definition 3 makes sense for any finite set $S = M$ of m points for $k \leq m-1$. Then the matrix $D(S;m-1)$ for the largest possible number $k = m-1$ of neighbors includes all pairwise distances, but differs from the usual symmetric distance matrix of S due to the ordered distances in each row. This pointwise information distinguishes the 4-point sets T, K in Fig. 3 as follows. The trapezium T and kite K in $\mathbb{R}^2$ can be represented by the points $(\pm 1, 1), (\pm 2, 0)$ and $(-2, 0), (-1, \pm 1), (2, 0)$, respectively. The matrices from

Definition 3 are $D(T;3) = \begin{pmatrix} \sqrt{2} & 2 & \sqrt{10} \\ \sqrt{2} & 2 & \sqrt{10} \\ \sqrt{2} & \sqrt{10} & 4 \\ \sqrt{2} & \sqrt{10} & 4 \end{pmatrix}$ and $D(K;3) = \begin{pmatrix} \sqrt{2} & \sqrt{2} & \sqrt{10} \\ \sqrt{2} & 2 & \sqrt{10} \\ \sqrt{2} & 2 & \sqrt{10} \\ \sqrt{10} & \sqrt{10} & 4 \end{pmatrix}$.
The first components of the vectors $\mathrm{AMD}^{(3)}(T) = (\sqrt{2}, 1 + \frac{\sqrt{10}}{2}, 2 + \frac{\sqrt{10}}{2})$, $\mathrm{AMD}^{(3)}(K) = (\frac{3\sqrt{2}+\sqrt{10}}{4}, 1 + \frac{\sqrt{2}+\sqrt{10}}{4}, 1 + \frac{3}{4}\sqrt{10})$ distinguish K, T in Fig. 3.

If S is periodic, all AMD values form the infinite sequence $\{\mathrm{AMD}_k\}_{k=1}^{+\infty}$. In practice, we compute the vector $\mathrm{AMD}^{(k)} = (\mathrm{AMD}_1, \dots, \mathrm{AMD}_k)$ up to a certain number k of neighbors. However, k is not a parameter that changes the output. If we increase k, we get more values without changing the previous ones, so k is similar to a desired degree of approximation. Since the asymptotic behaviour of AMD_k will be explicitly described in Theorem 13, the infinite AMD sequence can be informally compared with the sequence of coefficients in a degree k Taylor polynomial approximating an analytic function.

Theorem 4 (isometry invariance of AMD_k)**.** *For any finite or periodic point set $S \subset \mathbb{R}^n$, the Average Minimum Distance $\mathrm{AMD}_k(S)$ from Definition 3 is an isometry invariant of S for any $k \geq 1$.* ■

Proof. If S is periodic, first we show that the unordered collections of rows of the matrix $D(S;k)$, and hence $\mathrm{AMD}_k(S)$, is independent of a primitive unit cell. Let U, U' be different primitive cells of the periodic point set $S \subset \mathbb{R}^n$ with a lattice Λ. Any point $q \in S \cap U'$ can be translated along $\vec{v} \in \Lambda$ to a point $p \in S \cap U$ and vice versa. These translations establish a bijection between the motifs $S \cap U \leftrightarrow S \cap U'$ and preserve all distances. Hence the matrix $D(S;k)$ is the same for both cells U, U' up to a permutation of rows.

Now we prove that $D(S;k)$, and hence $\mathrm{AMD}_k(S)$, is preserved by any isometry $f : S \to Q$. Any primitive cell U of S is bijectively mapped by f to the unit cell $f(U)$ of Q, which should be also primitive. Indeed, if Q is preserved by a translation along a vector $\vec{v}$ that doesn't have all integer coefficients in the basis of $f(U)$, then $S = f^{-1}(Q)$ is preserved by the translation along $f^{-1}(\vec{v})$, which also doesn't have all integer coefficients in the basis of U, i.e. U was non-primitive. Since both primitive cells U and $f(U)$ contain the same number of points from S and $Q = f(S)$, the isometry f gives a bijection between all motif points of S, Q. For any sets S, Q, since f preserves distances, every list of ordered distances from any point $p_i \in S \cap U$ 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. The matrices $D(S;k), D(Q;k)$ are identical up to permutations of rows, hence $\mathrm{AMD}_k(S) = \mathrm{AMD}_k(Q)$. □

Example 5. ***(5a)*** *Table 1 implies by Theorem 4 that $S(r), Q(r)$ in Fig. 6 are not isometric for $0 < r \le 1$. The mirror image of $S(r) = \{0, r, r+2, 4\} + 8\mathbb{Z}$ under the reflection $t \mapsto 4 - t$ coincides with $S(2-r) = \{0, 2-r, 4-r, 4\} + 8\mathbb{Z}$, so they are equivalent up to isometries including reflections. Similarly, $Q(r)$ and $Q(2-r)$ are isometric by $t \mapsto -t$. Though $\mathrm{AMD}_k(S(r))$ seem to be independent of r, the first column of $D(S(r);3)$ has the minimum distance r, which distinguishes $S(r)$ between each other for all $0 < r \le 1$.*

(5b) *The 1-dimensional periodic sets $S_{15} = \{0, 1, 3, 4, 5, 7, 9, 10, 12\} + 15\mathbb{Z}$ and $Q_{15} = \{0, 1, 3, 4, 6, 8, 9, 12, 14\} + 15\mathbb{Z}$ in Fig. 6 are not isometric due to $\mathrm{AMD}_k(S_{15}) \ne \mathrm{AMD}_k(Q_{15})$ for $k = 3, 4$ in Table 2. All their density functions $\psi_k(t)$ are identical [17, section 5].* ■

Table 1. Matrices $D(S;k)$ and AMD from Definition 3 distinguish the sets $S(r) = \{0, r, r+2, 4\} + 8\mathbb{Z}$ and $Q(r) = \{0, r+2, 4, r+4\} + 8\mathbb{Z}$ for any $0 < r \le 1$.

$D(S(r);3)$	distance to neighbor 1	distance to neighbor 2	distance to neighbor 3
$p_1 = 0$	$\lvert 0 - r\rvert = r$	$\lvert 0 - (2+r)\rvert = 2 + r$	$\lvert 0 - 4\rvert = 4$
$p_2 = r$	$\lvert r - 0\rvert = r$	$\lvert r - (2+r)\rvert = 2$	$\lvert r - 4\rvert = 4 - r$
$p_3 = 2 + r$	$\lvert (2+r) - 4\rvert = 2 - r$	$\lvert (2+r) - r\rvert = 2$	$\lvert (2+r) - 0\rvert = 2 + r$
$p_4 = 4$	$\lvert 4 - (2+r)\rvert = 2 - r$	$\lvert 4 - r\rvert = 4 - r$	$\lvert 4 - 0\rvert = 4$
$\mathrm{AMD}_k(S(r))$	$\mathrm{AMD}_1 = 1$	$\mathrm{AMD}_2 = 2.5$	$\mathrm{AMD}_3 = 3.5$

$D(Q(r);3)$	distance to neighbor 1	distance to neighbor 2	distance to neighbor 3
$p_1 = 0$	$\lvert 0 - (2+r)\rvert = 2 + r$	$\lvert 0 - (r+4-8)\rvert = 4 - r$	$\lvert 0 - 4\rvert = 4$
$p_2 = 2 + r$	$\lvert (2+r) - 4\rvert = 2 - r$	$\lvert (2+r) - (4+r)\rvert = 2$	$\lvert (2+r) - 0\rvert = 2 + r$
$p_3 = 4$	$\lvert 4 - (4+r)\rvert = r$	$\lvert 4 - (2+r)\rvert = 2 - r$	$\lvert 4 - 0\rvert = 4$
$p_4 = 4 + r$	$\lvert (4+r) - 4\rvert = r$	$\lvert (4+r) - (2+r)\rvert = 2$	$\lvert (4+r) - 8\rvert = 4 - r$
$\mathrm{AMD}_k(Q(r))$	$\mathrm{AMD}_1 = 1 + 0.5r$	$\mathrm{AMD}_2 = 2.5 - 0.5r$	$\mathrm{AMD}_3 = 3.5$

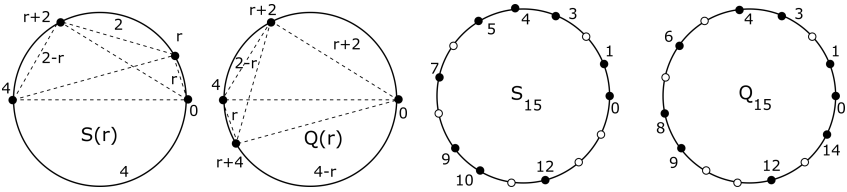

Figure 6. **Left**: circular versions of the periodic point sets $S(r) = \{0, r, r+2, 4\} + 8\mathbb{Z}$ and $Q(r) = \{0, r+2, 4, r+4\} + 8\mathbb{Z}$ for $0 < r \le 1$. The distances between points (shown outside the disk) are arc lengths (shown inside the disk). **Right**: S_{15}, Q_{15} from Example 5b are distinguished by AMD_3, not by the density functions $\psi_k(t)$ for all $k \ge 1$ [3, Example 11].

Table 2. First row: points from the motif M of S_{15} and Q_{15} in Fig. 6. **Further rows**: distances from each $p \in M$ to its k-th neighbor in S_{15} and Q_{15}.

S_{15}	0	1	3	4	5	7	9	10	12	AMD_k
$k=1$	1	1	1	1	1	2	1	1	2	11/9
$k=2$	3	2	2	1	2	2	2	2	3	19/9
$k=3$	3	3	2	3	2	3	3	3	3	25/9
$k=4$	4	4	3	3	4	3	4	5	4	34/9
Q_{15}	0	1	3	4	6	8	9	12	14	AMD_k
$k=1$	1	1	1	1	2	1	1	2	1	11/9
$k=2$	1	2	2	2	2	2	3	3	2	19/9
$k=3$	3	2	3	3	3	4	3	3	2	26/9
$k=4$	3	3	3	4	3	4	5	4	4	33/9

4 Continuity of Average Minimum Distances

For the isometry invariance of $\mathrm{AMD}(S;k)$ in Theorem 4, a unit cell U in Definition 3 should be primitive. If U contains m points and we make one edge of U twice longer, the resulting non-primitive unit cell contains $2m$ points and the matrix $D(S;k)$ will be twice larger. A translated copy of any point $p_i \in U$ will have exactly the same ordered distances to its neighbors as p_i due to periodicity. After doubling U, every row is repeated twice in $D(S;k)$. The requirement of a primitive cell U makes $D(S;k)$ discontinuous similarly to the cell volume $\mathrm{Vol}[U]$ in Fig. 1. One way to resolve this discontinuity is to average each column of $D(S;k)$ to get $\mathrm{AMD}_k(S)$ in Definition 3.

Continuity of $\mathrm{AMD}_k(S)$ is most naturally measured relative to a maximum perturbation of points needed to get one set from another, as formalised below.

Definition 6 (bottleneck distance between sets)**.** *For a bijection $g : S \to Q$ between finite or periodic point sets $S, Q \subset \mathbb{R}^n$, the* maximum deviation *is the supremum $\sup_{p\in S} |p - g(p)|$ over $p \in S$. The* bottleneck distance *is defined as $d_B(S,Q) = \inf_{g:S\to Q} \sup_{p\in S} |p - g(p)|$ is the infimum over bijections $g : S \to Q$.* ■

The bottleneck distance is impractical to compute because of a minimisation over infinitely many bijections. Theorem 15 in Section 7 will justify that there is no continuous way to select a unit cell of a lattice. However, continuity of isometry invariants can be checked for all small perturbations in the bottleneck distance. Continuity Theorem 9 requires Lemmas 7 and 8.

Lemma 7 (Lemma 2 in [17]). *Let periodic point sets $S, Q \subset \mathbb{R}^n$ have a bottleneck distance $d_B(S,Q) < r(Q)$, where the* packing radius *$r(Q)$ is the minimum half-distance between points of Q. Then S, Q have a common lattice Λ with a unit cell U such that $S = \Lambda + (U \cap S)$ and $Q = \Lambda + (U \cap Q)$.* ■

Lemma 8 (perturbed distances). *For some $\varepsilon > 0$, let $g : S \to Q$ be a bijection between finite or periodic sets such that $|a - g(a)| \le \varepsilon$ for all $a \in S$. For any $i \ge 1$, let $a_i \in S$ and $b_i \in Q$ be the i-nearest neighbors of points $a \in S$ and $b = g(a) \in Q$, respectively. Then the Euclidean distances from a, b to their i-th neighbors a_i, b_i are 2ε-close, i.e. $||a - a_i| - |b - b_i|| \le 2\varepsilon$.* ■

Proof. Translate the full set Q by the vector $a - g(a)$. So we assume that $a = g(a)$ and $|b - g(b)| < 2\varepsilon$ for all $b \in S$. Assume by contradiction that the distance from a to its i-th neighbor b_i is less than $|a - a_i| - 2\varepsilon$.

Then all first i neighbors $b_1, \dots, b_i$ of a within Q belong to the open ball with the center a and the radius $|a - a_i| - 2\varepsilon$. Since the bijection g shifted every point $b_1, \dots, b_i$ by at most 2ε, their preimages $g^{-1}(b_1), \dots, g^{-1}(b_i)$ belong to the open ball with the center $a = g(a)$ and the radius $|a - a_i|$. Then the i-th neighbor of a within S is among these i preimages. Hence the distance from a to its i-th nearest neighbor is strictly less than the required distance $|a - a_i|$. A similar contradiction is obtained from the assumption that the distance from a to its new i-th neighbor b_i is more than $|a - a_i| + 2\varepsilon$. □

Theorem 9 (continuity of AMD under any small perturbations). *Let finite or periodic sets $S, Q \subset \mathbb{R}^n$ satisfy $d_B(S,Q) < r(Q)$, where $r(Q)$ is the packing radius of Q. Then $|\mathrm{AMD}_k(S) - \mathrm{AMD}_k(Q)| \le 2d_B(S,Q)$ for $k \ge 1$.* ■

Proof. By Lemma 7 the sets S, Q have a common lattice Λ. Any primitive cell U of Λ is a unit cell of S, Q, i.e. $S = \Lambda + (S \cap U)$ and $Q = \Lambda + (Q \cap U)$. Since the bottleneck distance $\varepsilon = d_B(S,Q) < r(Q)$, we can define a bijection g from every point $a \in S$ to its unique ε-closest neighbor $g(a) \in Q$.

If U is a non-primitive unit cell of S, the matrix $D(S;k)$ can be constructed as in Definition 3, but each row will be repeated $n(S) > 1$ times, where $n(S)$ is $\mathrm{Vol}[U]$ divided by the volume of a primitive unit cell of S. The average $\mathrm{AMD}_k(S)$ in the k-th column is independent of the factor $n(S) > 1$.

Since the above conclusions hold for Q instead of S, we now compare the matrices $D(S;k)$ and $D(Q;k)$ built on the same unit cell U and have equal sizes. By Lemma 8 the corresponding elements of the matrices $D(S;k)$ and $D(Q;k)$ differ by at most 2ε, i.e. $|D_{ij}(S;k) - D_{ij}(Q;k)| \le 2\varepsilon$. The average of the k-th column changes by at most 2ε, i.e. $|\mathrm{AMD}_k(S) - \mathrm{AMD}_k(Q)| \le 2\varepsilon$. □

5 The explicit asympotic behaviour of AMD

The main result of this section is Theorem 13 explicitly describing the asymptotic growth of AMD_k as $k \to +\infty$ for a wide class of sets including non-periodic sets. The average minimum distance $\mathrm{AMD}_k(S)$ approaches $c(S)\sqrt[n]{k}$, where the point packing coefficient $c(S)$ is introduced below. The volume of the unit ball in $\mathbb{R}^n$ is $V_n = \dfrac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)}$, where $\Gamma(m) = (m-1)!$ and $\Gamma(\frac{m}{2}+1) = \sqrt{\pi}(m-\frac{1}{2})(m-\frac{3}{2})\cdots\frac{1}{2}$ for integer $m \ge 1$.

Definition 10 ((U,m)-sets S, point packing coefficient $c(S)$)**.** *Let U be a unit cell of a lattice $\Lambda \subset \mathbb{R}^n$. For any fixed $m \ge 1$, a set $S \subset \mathbb{R}^n$ is called a* (U,m)-set *if $S \cap (U + \vec{v})$ consists of m points for any vector $\vec{v} \in \Lambda$. For any point $p \in S \cap U$, let $d_k(S;p)$ be the distance from p to its k-th neighbor in S. The* Average Minimum Distance *is* $\mathrm{AMD}_k(S;U) = \dfrac{1}{m}\sum\limits_{p\in S\cap U} d_k(S;p)$. *The* Point Packing Coefficient *is* $c(S) = \sqrt[n]{\dfrac{\mathrm{Vol}[U]}{mV_n}}$, *where S has m points in U.* ■

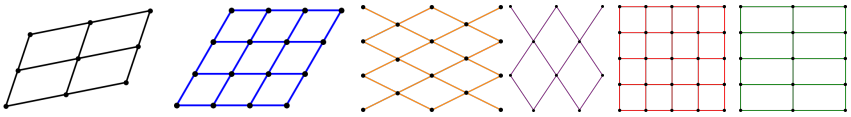

Figure 7. The six 2-dimensional lattices Λ_i have the point packing coefficients $c(\Lambda_i)$ below and the curves of AMD_k values for $k = 1,\dots,300$ in Fig. 8. **1st**: a generic black lattice Λ_1 with the basis $(1.25, 0.25), (0.25, 0.75)$ and $c(\Lambda_1) = \sqrt{\dfrac{7}{8\pi}} \approx 0.525$. **2nd**: the blue hexagonal lattice Λ_2 with the basis $(1,0), (1/2, \sqrt{3}/2)$ and $c(\Lambda_2) = \sqrt{\dfrac{\sqrt{3}}{2\pi}} \approx 0.528$. **3rd**: the orange rhombic lattice Λ_3 with the basis $(1, 0.5), (1, -0.5)$ and $c(\Lambda_3) = \sqrt{\dfrac{1}{\pi}} \approx 0.564$. **4th**: the purple rhombic lattice Λ_4 with the basis $(1, 1.5), (1, -1.5)$ and $c(\Lambda_4) = \sqrt{\dfrac{3}{\pi}} \approx 0.977$. **5th**: the red square lattice Λ_5 with the basis $(1,0), (0,1)$ and $c(\Lambda_5) = \sqrt{\dfrac{1}{\pi}} \approx 0.564$. **6th**: the green rectangular lattice Λ_6 with the basis $(2,0), (0,1)$ and $c(\Lambda_6) = \sqrt{\dfrac{2}{\pi}} \approx 0.798$.

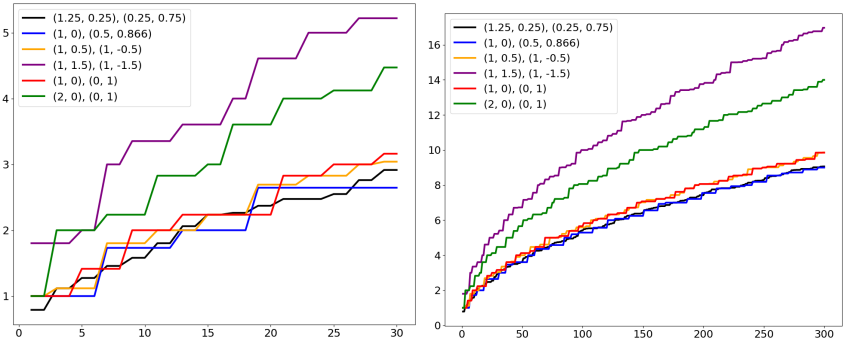

Figure 8. Left: AMD_k up to $k = 30$ for the 2D lattices in Fig. 7. **Right**: AMD_k extended to $k = 300$. The orange and red lattices have close point packing coefficients $c(\Lambda_i)$, so their AMD curves approach each other by Theorem 13 but distinguish these lattices.

An example of a (U, m)-set is a non-periodic perturbation of a periodic set $S = \Lambda + M$, where a lattice Λ is generated by a unit cell U, a motif M has m points. Since S may not be periodic, $\mathrm{AMD}_k(S; U + \vec{v})$ can depend on a shift vector $\vec{v} \in \Lambda$. Even if S is periodic, a unit cell U in Definition 10 can be non-primitive. However, $\mathrm{Vol}[U]/m$ is independent of a choice of U. Hence $c(S)$ is an isometry invariant, also for any (U, m)-set S, because any shifted cell $U + \vec{v}$ contains the same number m of points from S by Definition 10.

If all points have the weight V_n, then $(c(S))^n$ is inversely proportional to the density $\rho = \dfrac{mV_n}{\mathrm{Vol}[U]}$ of S. The *diameter* of a unit cell U is $d = \sup\limits_{a,b \in U} |a - b|$.

Lemma 11 (bounds on points within a ball)**.** *Let $S \subset \mathbb{R}^n$ be any (U, m)-set with a unit cell U, which generates a lattice Λ and has a diameter d. For any point $p \in S \cap U$ and a radius r, consider the* lower *union $U'(p;r) = \bigcup\{(U + \vec{v})$ such that $\vec{v} \in \Lambda, (U + \vec{v}) \subset \bar{B}(p;r)\}$ and the* upper *union $U''(p;r) = \bigcup\{(U + \vec{v})$ such that $\vec{v} \in \Lambda, (U + \vec{v}) \cap \bar{B}(p;r) \neq \emptyset\}$. Then the number of points from S in the closed ball $\bar{B}(p;r)$ with center p and radius r has the bounds* $\left(\dfrac{r-d}{c(S)}\right)^n \le m\dfrac{\mathrm{Vol}[U'(p;r)]}{\mathrm{Vol}[U]} \le |S \cap \bar{B}(p;r)| \le m\dfrac{\mathrm{Vol}[U''(p;r)]}{\mathrm{Vol}[U]} \le \left(\dfrac{r+d}{c(S)}\right)^n$. ■

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

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

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

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

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

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

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

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

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

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

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

Lemma 12 (distance bounds)**.** *Let $S \subset \mathbb{R}^n$ be any (U,m)-set with a unit cell U of diameter d. For any point $p \in S \cap U$, let $d_k(S;p)$ be the distance from p to its k-th nearest neighbor in S. Then $c(S)\sqrt[n]{k} - d < d_k(S;p) \le c(S)\sqrt[n]{k} + d$ for any $k \ge 1$.* ■

Proof. The closed ball $\bar{B}(p;r)$ of the radius $r = d_k(S;p)$ has more than k points (including p) from S. The upper bound of Lemma 11 for $r = d_k(S;p)$ implies that $k < |S \cap \bar{B}(p;r)| \le \dfrac{(r+d)^n}{(c(S))^n}$. Taking the n-th roots, we get $\sqrt[n]{k} < \dfrac{r+d}{c(S)}$, so $r = d_k(S;p) > c(S)\sqrt[n]{k} - d$.

For any smaller radius $r < d_k(S;p)$, the closed ball $\bar{B}(p;r)$ contains at most k points (including p) from S. The lower bound of Lemma 11 for any $r < d_k(S;p)$ implies that $\dfrac{(r-d)^n}{c(S)^n} \le |S \cap \bar{B}(p;r)| \le k$. Since $\dfrac{(r-d)^n}{c(S)^n} \le k$ holds for the constant upper bound k and any radius $r < d_k(S;p)$, the same inequality holds for the radius $r = d_k(S;p)$. Similarly to the upper bound above, we get $\dfrac{r-d}{c(S)} \le \sqrt[n]{k}$, $r = d_k(S;p) \le c(S)\sqrt[n]{k} + d$. Combine the two bounds above as follows: $c(S)\sqrt[n]{k} - d < d_k(S;p) \le c(S)\sqrt[n]{k} + d$. □

Theorem 13 (asymptotic behaviour of AMD)**.** *For any (U,m)-set $S \subset \mathbb{R}^n$ from Definition 10, we have $|\mathrm{AMD}_k(S;U) - c(S)\sqrt[n]{k}| \le d$ for any $k \ge 1$ and $\lim\limits_{k\to+\infty} \dfrac{\mathrm{AMD}_k(S;U)}{\sqrt[n]{k}}$ equals the Point Packing Coefficient $c(S)$.* ■

Proof. Averaging the bounds of Lemma 12 over all points $p \in S \cap U$, we get $c(S)\sqrt[n]{k} - d < \mathrm{AMD}_k(S;U) = \dfrac{1}{m}\sum\limits_{p\in S\cap U} d_k(S;p) \le c(S)\sqrt[n]{k} + d$, which imply that $|\mathrm{AMD}_k(S;U) - c(S)\sqrt[n]{k}| \le d$, $k \ge 1$. So $\lim\limits_{k\to+\infty} \dfrac{\mathrm{AMD}_k(S;U)}{\sqrt[n]{k}} = c(S)$. □

6 A near linear time algorithm and experiments

This section describes the AMD algorithm in Theorem 14 and experiments on big datasets. The input for computing AMD is a periodic point set S given by the basis vectors of its unit cell U and the Cartesian coordinates of m motif points. The length k of the vector $\mathrm{AMD}^{(k)}(S) = (\mathrm{AMD}_1, \dots, \mathrm{AMD}_k)$ is independent of a periodic point set S. Increasing k adds more components to the vector $\mathrm{AMD}^{(k)}$ without changing any previous values.

The size of an input for any real periodic structure is proportional to the number m of points in a motif. Theorem 14 solves Problem 1 requiring a near linear time in both k, m. For a unit cell U with a diameter d, define the *skewness* $\nu = \dfrac{d}{\sqrt[n]{\mathrm{Vol}[U]}}$. Reduced cells of most real structures have a small skewness ν. For any fixed ν, in Theorem 14 below $(5\nu)^n V_n \to 0$ as $n \to +\infty$ by Stirling's approximation of the factorial $n!$ hidden in the unit ball volume V_n.

Theorem 14 (a near linear time algorithm for AMD)**.** *Let a periodic set $S \subset \mathbb{R}^n$ have m points in a unit cell U. Then $\mathrm{AMD}_i(S)$ can be computed for $i = 1, \dots, k$ in time $O((5\nu)^n V_n k m \log^2(km))$, where V_n is the unit ball volume in $\mathbb{R}^n$, d and $\nu = \dfrac{d}{\sqrt[n]{\mathrm{Vol}[U]}}$ are the diameter and* skewness *of the cell U.* ■

Proof. Let the origin $0 \in \mathbb{R}^n$ be in the center of the unit cell U. If d is the diameter of U, any point $p \in M = S \cap U$ is covered by the closed ball $\bar{B}(0, 0.5d)$. By Lemma 12 all k neighbors of p are covered by the ball $\bar{B}(0;r)$ of radius $r = c(S)\sqrt[n]{k} + 1.5d$. To generate all Λ-translates of M within $\bar{B}(0;r)$, we gradually extend U in spherical layers by adding more shifted cells until we get the upper union $U''(0;r) \supset \bar{B}(0;r)$. By Lemma 11 the union $U''(0;r)$ includes k neighbors of motif points and has at most $\mu \le m\dfrac{\mathrm{Vol}[U''(0;r)]}{\mathrm{Vol}[U]} \le$

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

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

A cover tree on μ points can be built in time $O(\mu \log \mu)$ by [6, Theorems 4, 6]. Then all k neighbors of $m \le \mu$ motif points $p \in M$ can be found in time $O(mk \log k)$ by [18, chapter 3], which extends [14, Theorem 2] for $k = 1$, where hidden constants depend on a type of a point set, but not on its size. By Definition 3 we lexicographically sort m lists of ordered distances in time $O(km \log m)$, because a comparison of any two ordered lists of length k takes $O(k)$ time. The ordered lists of distances are the rows of the matrix $D(S; k)$. Then $\mathrm{AMD}_i(S)$ are found as column averages in time $O(km)$. Using $\log \mu = n \log 2 + O(\log(k + m(2.5\nu)^n V_n)) = nO(\log(km))$, the total time is

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

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

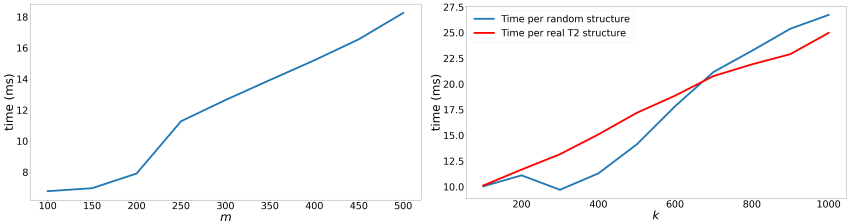

Figure 9. The blue curves show the AMD time averaged over 3000 sets whose points are uniformly generated in unit cells with random edge-lengths in $[1, 2]$ and angles in $[\frac{\pi}{3}, \frac{2\pi}{3}]$. **Left**: $k = 200$. **Right**: $m = 250$ for the blue curve, 5679 structures for the red curve.

Our implementations of AMD invariants in Python [39] and C++ [29] use the n-d tree [12] on μ points in $\mathbb{R}^n$, which performs faster in small dimensions than a cover tree but has no proved complexity for a nearest neighbor search. Fig. 9 illustrates a near linear running time justified by Theorem 14. The red curve on the right was obtained on 5679 predicted structures, which have the same chemical composition and contain about $m = 250$ points on average in a unit cell, which has led to four new crystals [34].

The only other implemented sequence of continuous isometry invariants are density functions $\psi_k(t)$ requiring a cubic time in k, see [17, section 6.3]. Because of this cubic increase, we ran the C++ code [38] only up to $k = 8$ for four representative points out of 46 atoms per molecule. These simplified sets contain about 25 points on average in a unit cell. Each $\psi_k(t)$ depends on a continuous radius t and can be evaluated only at discretely

sampled values of t. We computed the first eight $\psi_k(t)$ sampled at t equal to multiples of $0.2\mathring{A}$ up to about $12\mathring{A}$, where $1\mathring{A} = 10^{-10}$m. Smallest distances between atoms in crystals are about $0.8\mathring{A}$ at this atomic scale. With the above parameters, the experiments on similar machines (Dell XPS 15 6-core, 2.20GHz, 16GB and MacBook Pro, 2.3GHz, 8GB) took about 1 min per structure, more than 4 days for 5679 simplified sets.

In comparison with the above times of density functions, vectors $\mathrm{AMD}^{(k)}$ were computed for full periodic structures and much larger k on the modest desktop AMD Ryzen 5 6-core 4.60Ghz, 32GB DDR4. The red curve in Fig. 9 for the 5679 structures implies the average time of 10ms for $\mathrm{AMD}^{(100)}$ and 27ms for $\mathrm{AMD}^{(1000)}$, so the total time was about 17 min. Despite the world's largest Cambridge Structural Database (CSD) has much more diverse compositions in comparison with T2 structures based on the same molecule, $\mathrm{AMD}^{(100)}$ required the similar time of 13.4ms on average, less than 52 min in total for all 228,994 molecular organic real crystal structures in the CSD.

7 A discussion of contributions and further steps

In conclusion, the AMD sequence is a novel continuous invariant of periodic point sets up to isometry in $\mathbb{R}^n$ by Theorems 4 and 9. Theorem 13 is especially strong due to the asymptotic formula working even for non-periodic sets S with the new point packing coefficient $c(S)$. Theorem 14 justified a near linear running time in both input parameters k, m, which enabled visualisations of huge data with modest resources, see appendix B.

The key motivation for continuous isometry invariants is the discontinuity of reduced cells, which was experimentally known [2] since 1980. Theorem 15 below will disprove any possibility of a continuous reduction. Let B be the space of linear bases $b = \{\vec{v}_1, \dots, \vec{v}_n\}$. If we concatenate n vectors into one vector with n^2 coordinates, the space B becomes a subset of $\mathbb{R}^{n^2}$ with the Euclidean topology. Since any basis generates a lattice, we have the projection $g : B \to L$, where L is the space of all lattices in $\mathbb{R}^n$. This space L can have the minimal topology that makes g continuous so that the preimage $g^{-1}(N(\Lambda))$ of any open neighborhood $N(\Lambda)$ is a union of open neighborhoods of bases from $g^{-1}(\Lambda) \subset B$. Continuity of $g : B \to L$ means that any small perturbation of a basis gives rise to a small perturbation of the lattice generated by this basis. A desired reduction would be a continuous map $h : L \to B$ such that $g \circ h(\Lambda) = \Lambda$ is the identity. If we continuously change a lattice $\Lambda \in L$, its reduced basis $h(\Lambda) \in B$ should also change continuously.

Theorem 15 (discontinuity of reduced cells)**.** *Let $g : B \to L$ map any basis b of $\mathbb{R}^n$ to its lattice Λ. There is no continuous map $h : L \to B$ such that $g \circ h$ is the identity.* ■

Proof. Let $h(\mathbb{Z}^n) = \{\vec{v}_1, \dots, \vec{v}_n\} \in B$ be a reduced basis of integer lattice $\mathbb{Z}^n \subset \mathbb{R}^n$. Consider the continuous path $\gamma : [0,1] \to B$, where $\gamma(t)$ is the basis $\vec{v}_1 + t\vec{v}_2, \vec{v}_2, \dots, \vec{v}_n$. Since the bases $\gamma(0) \neq \gamma(1)$ define the same lattice $\mathbb{Z}^n \subset \mathbb{R}^n$, the composition $g \circ \gamma$ is a continuous loop $g \circ \gamma : [0,1] \to L$ in the space of lattices with $g \circ \gamma(0) = \mathbb{Z}^n = g \circ \gamma(1)$.

It remains to show that the continuous map h lifts the loop $g \circ \gamma$ to the path $\gamma : [0,1] \to B$ with disjoint endpoints $\gamma(0) \neq \gamma(1)$, which is a contradiction with the existence of such h. Since h is continuous, for all sufficiently small $t > 0$, the basis $h \circ g \circ \gamma(t)$ should be close to $h(\mathbb{Z}^n)$, hence should coincide with $\gamma(t)$, because all other bases of the lattice $h \circ g \circ \gamma(t)$ close to $\mathbb{Z}^n$ are sufficiently away from $\gamma(t)$ in Euclidean metric on B.

This local extension argument works around any $t \in [0,1]$ where we already know that $h \circ g \circ \gamma(t) = \gamma(t)$. Since any infinite cover of the compact line segment $[0,1]$ by open neighborhoods contains a finite subcover, we need only finitely many steps to get $\gamma(1) = h \circ g \circ \gamma(1) = h(\mathbb{Z}^n) = h \circ g \circ \gamma(0) = \gamma(0)$, which contradicts the fact that the initial bases $\gamma(1) \neq \gamma(0)$. □

A limitation of AMD is its potential incompleteness, though we do not know any non-isometric periodic sets that have equal AMD_k for all $k \geq 1$. Theorem 16 hints at completeness of AMD for periodic point sets in general position.

Theorem 16 (completeness for generic finite sets)**.** *Let a finite set $S \subset \mathbb{R}^n$ consist of m points such that all pairwise distances between points of S are distinct. Then S can be uniquely reconstructed from the matrix $D(S; m-1)$ in Definition 3 up to isometry.* ■

Proof. Since the distances between all m points of $S \subset \mathbb{R}^n$ are distinct, every distance appears in the matrix $D(S; m-1)$ exactly twice, once as the distance from a point p_i to its neighbor p_j, and once more as the distance from p_j to p_i, though these equal entries are not symmetric. We will convert $D(S; m-1)$ into the distance matrix $D(S)$. Let $d_1 < d_2 < \dots < d_{m-1}$ be all distances from the first point $p_1 \in S$ to all $m-1$ others. Each distance d_i from the first row of $D(S; m-1)$ appears exactly once more in another (say, i'-th) row of $D(S; m-1)$. Then d_i is the distance between the points p_1 and $p_{i'}$ numbered as the i'-th row. The map of indices $i \mapsto i'$ is a permutation of $\{2, \dots, m\}$. We

set $D_{11}=0$ and $D_{1,i'}=d_i$ for each $i=2,\dots,m$. Then we similarly permute indices in the 2nd row of $D(S;m-1)$, starting from the 3rd index due to the symmetry of $D(S)$, and so on. The full distance matrix $D(S)$ uniquely determines a set with ordered points $S\subset\mathbb{R}^n$ modulo isometries by the classical multi-dimensional scaling in [25, Section 8.5.1]. $\square$

Encouraged by Theorem 16 and fast speed of the AMD algorithm, we have computed the invariant $\mathrm{AMD}^{(100)}$ vector for all 229K molecular organic crystals from the world's largest Cambridge Structural Database (CSD). We found 405 pairs with identical $\mathrm{AMD}^{(100)}$ and checked them by the traditional packing similarity [13] measuring the Root Mean Square Devitation (RMSD) between atomic positions in 15 (by default) matched molecules from two crystals. It turned out that these 405 pairs have RMSD$\approx 10^{-15}\mathring{A}$ in the range of typical floating point errors. So all these pairs of crystals are duplicates but remained undetected in the CSD, because RMSD is slow to compute pairwisely.

The AMD invariants were enough to predict the lattice energy of molecular crystals within 5kJ/mole [35] without using any chemical data. In the partial case of lattices, their space of isometry classes was continuously parameterised by root forms [9,10] in dimension two and three. AMD were recently extended to Pointwise Distance Distributions (PDD) whose continuity was proved [40] under Earth Mover's Distance, which was used for comparing chemical compositions [24]. The above root forms of lattices combined with PDD are enough to explicitly reconstruct a periodic point set in general position, which justifies a geometric inverse design for any periodic crystals.

The fact that $\mathrm{AMD}^{(100)}$ distinguished 229K real crystals supports the Crystal Isometry Principle (CRISP) saying that the map {periodic crystals} $\to$ {periodic point sets} is injective, where both crystals and point sets are considered up to isometry. Indeed, replacing one chemical element by another inevitably changes its distances to neighbors, which is easily captured by the new distance-based invariants AMD or PDD. The pairwise computations for over 660K periodic crystals (with full 3D geometry and without disorder) from the Cambridge Structural Database (CSD) have identified five pairs of crystals that have identical $\mathrm{AMD}^{(100)}$ invariants and differ only by the chemical type of a single atom:

HIFCAB and JEPLIA (Cd $\leftrightarrow$ Mn),

COLYEI and POCLOK (Eu $\leftrightarrow$ Sm),

DTBIPT and DTHBPD10 (Pt $\leftrightarrow$ Pd),

LALNET and POCPAA (Cd ↔ Ni),

AFIBOH and NENCUF (Cd ↔ Zn).

After the above five pairs of 'needles in a haystack' were automatically detected by invariant computations, it was easy enough to manually check that in each pair both crystals have identical unit cell parameters, atomic coordinates and structure factors. These coincidences seem unrealistic to all crystallographers who looked the raw data.

Our colleagues in the Cambridge Crystallographic Data Centre are now contacting the journals that published the underlying papers. The above examples were not discovered by any of the past tools such as RMSD or PXRD, because they are slow even for comparing a single newly deposited crystal with more than 1.1M entries in the CSD. Moreover, checking the cell and motif data for coincidence cannot reliably detect duplicates, because cell parameters and atomic coordinates can be easily changed for any crystal.

Acknowledgments: We thank the co-authors of [17] and Nikolai Dolbilin for fruitful discussions and all reviewers for their valuable time and suggestions. This research was supported by the £3.5M EPSRC grant 'Application-driven Topological Data Analysis' (2018-2023, EP/R018472/1), the £10M Leverhulme Research Centre for Functional Materials Design (2016-2026) and the last author's Royal Academy of Engineering Fellowship 'Data Science for Next Generation Engineering of Solid Crystalline Materials' (2021-2023, IF2122/186).

References

[1] H. Alt, K. Mehlhorn, H. Wagener, E. Welzl, Congruence, similarity, and symmetries of geometric objects, *Discr. Comput. Geom.* **3** (1988) 237–256.

[2] L. Andrews, H. Bernstein, G. Pelletier, A perturbation stable cell comparison technique, *Acta Crystall. A* **36** (1980) 248–252.

[3] O. Anosova, V. Kurlin, Introduction to periodic geometry and topology, arXiv:2103.02749 (2021).

[4] O. Anosova, V. Kurlin, An isometry classification of periodic point sets, in: J. Lindblad, F. Malmberg, N. Sladoje (Eds.), *Discrete Geometry and Mathematical Morphology*, Springer, Cham, 2021, pp. 229–241.

[5] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, *Trans. PAMI* **24** (2002) 509–522.

[6] A. Beygelzimer, S. Kakade, J. Langford, Cover trees for nearest neighbor, in: W. Cohen, A. Moore (Eds.), *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 97–104.

[7] M. Bouniaev, N. Dolbilin, Regular and multi-regular t-bonded systems, *J. Inf. Proc.* **25** (2017) 735–740.

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

[9] M. Bright, A. I. Cooper, V. Kurlin, A complete and continuous map of the lattice isometry space for all 3-dimensional lattices, arXiv:2109.11538v1 (2021).

[10] M. Bright, A. I. Cooper, V. Kurlin, Easily computable continuous metrics on the space of isometry classes of 2-dimensional lattices, arXiv:2109.10885v1 (2021).

[11] M. I. Aroyo (Ed.), *International Tables for Crystallography. Volume A: Space–Group Symmetry*, Wiley, New York, 2016.

[12] R. A. Brown, Building a balanced k-d tree in $o(kn \log n)$ time, *J. Comput. Graph. Tech.* **4** (2015) 50–68.

[13] J. Chisholm, S. Motherwell, Compack: a program for identifying crystal structure similarity using distances, *J. Appl. Cryst.* **38** (2005) 228–231.

[14] R. R. Curtin, D. Lee, W. B. March, P. Ram, Plug-and-play dual-tree algorithm runtime analysis, *J. Mach. Learn. Res.* **16** (2015) 3269–3297.

[15] N. Dolbilin, M. Bouniaev, Regular t-bonded systems in R^3, *Eur. J. Comb.* **80** (2019) 89–101.

[16] N. Dolbilin, J. Lagarias, M. Senechal, Multiregular point systems, *Discr. Comput. Geom.* **20** (1998) 477–498.

[17] H. Edelsbrunner, T. Heiss, V. Kurlin, P. Smith, M. Wintraecken, The density fingerprint of a periodic point set, in: K. Buchin, É. C. de Verdière (Eds), *37th International Symposium on Computational Geometry (SoCG 2021)*, Dagstuhl Pub., Wadern, 2021.

[18] Y. Elkin, New geometric methods for a data-led discovery of topological shapes, PhD thesis, Univ. Liverpool, 2021.

[19] G. Fadda, G. Zanzotto, On the arithmetic classification of crystal structures, *Acta Cryst. A* **57** (2001) 492–506.

[20] J. Franklin, Ambiguities in the x-ray analysis of structures, *Acta Cryst. A* **30** (1974) 698–702.

[21] J. P. Glusker, B. K. Patterson, M. Rossi, *Patterson and Pattersons: Fifty Years of the Patterson Function*, Clarendon Press, Oxford, 1987.

[22] C. Grigorescu, N. Petkov, Distance sets for shape filters and shape recognition, *IEEE Trans. Image Proc.* **12** (2003) 1274–1286.

[23] F. Grünbaum, C. Moore, The use of higher-order invariants in the determination of generalized Patterson cyclotomic sets, *Acta Cryst. A* **51** (1995) 310–323.

[24] C. J. Hargreaves, M. S. Dyer, M. W. Gaultois, V. A. Kurlin, M. J. Rosseinsky, The earth mover's distance as a metric for the space of inorganic compositions, *Chem. Mat.* **32** (2020) 10610–10620.

[25] L. Liberti, C. Lavor, *Euclidean Distance Geometry – An Introduction*, Springer, Cham, 2017.

[26] J. Lima-de Faria, E. Hellner, F. Liebau, E. Makovicky, and E. Parthé, Nomenclature of inorganic structure types, *Acta Cryst. A* **46** (1990) 1–11.

[27] S. Manay, D. Cremers, B. W. Hong, A. J. Yezzi, S. Soatto, Integral invariants for shape matching, *Trans. PAMI* **28** (2006) 1602–1618.

[28] F. Mémoli, Gromov–Wasserstein distances and the metric approach to object matching, *Found. Comput. Math.* **11** (2011) 417–487.

[29] M. Mosca, AMD implemented in C++, `https://github.com/mmmosca/AMD`.

[30] M. Mosca, V. Kurlin, Voronoi-based similarity distances between arbitrary crystal lattices, *Crystal Res. Tech.* **55** (2020) #1900197.

[31] R. Osada, T. Funkhouser, B. Chazelle, D. Dobkin, Shape distributions, *ACM Trans. Graph.* **21** (2002) 807–832.

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

[33] D. Probst, J. L. Reymond, Visualization of very large high-dimensional data sets as minimum spanning trees, *J. Cheminf.* **12** (2020) 1–13.

[34] A. Pulido, L. Chen, T. Kaczorowski, D. Holden, M. Little, S. Chong, B. Slater, D. McMahon, B. Bonillo, C. Stackhouse, A. Stephenson, C. Kane, R. Clowes, T. Hasell, A. Cooper, G. Day, Functional materials discovery using energy–structure maps, *Nature* **543** (2017) 657–664.

[35] J. Ropers, M. M. Mosca, O. Anosova, V. Kurlin, A. I. Cooper, Fast predictions of lattice energies by continuous isometry invariants of crystal structures, arXiv:2108.07233 (2021).

[36] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, The open quantum materials database, *JOM* **65** (2013) 1501–1509.

[37] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, E. K. Gross, How to represent crystal structures for machine learning, *Phys. Rev. B* **89** (2014) #205118.

[38] P. Smith, Density functions implemented in C++, `https://github.com/Phil-Smith1/Density_Functions_Analysis`.

[39] D. Widdowson, AMD in Python, `https://github.com/dwiddo/AMD`.

[40] D. Widdowson, V. Kurlin, Pointwise distance distributions of periodic sets, arXiv:2108.04798 (2021).

[41] D. Widdowson, M. Mosca, A. Pulido, V. Kurlin, A. I. Cooper, Average minimum distances of periodic point sets, arXiv:2009.02488 (2020).

[42] B. Zhilinskii, *Introduction to Lattice Geometry Through Group Action*, EDP Sci., Paris, 2016.

[43] L. Zhu, M. Amsler, T. Fuhrer, B. Schaefer, S. Faraji, S. Rostami, S. A. Ghasemi, A. Sadeghi, M. Grauzinyte, C. Wolverton, S. Goedecker, A fingerprint based metric for measuring similarities of crystalline structures, *J. Chem. Phys.* **144** (2016) #034203.

[44] P. H. Zwart, R. W. Grosse-Kunstleve, A. A. Lebedev, G. N. Murshudov, P. D. Adams, Surprises and pitfalls arising from (pseudo) symmetry, *Acta Cryst. D* **64** (2008) 99–107.

A Appendix A: definitions and proof of Lemma 7

This section covers key facts about isometries in $\mathbb{R}^n$ and their invariants.

Definition 17 (isometry)**.** *An* isometry *of $\mathbb{R}^n$ is a map $f : \mathbb{R}^n \to \mathbb{R}^n$ that preserves the Euclidean distance, so $|p-q| = |f(p)-f(q)|$ for any points $p, q \in \mathbb{R}^n$. The map f also preserves the* orientation *if the matrix whose columns are images under f of the standard basis vectors $\vec{e}_1, \ldots, \vec{e}_n$ has a positive determinant. In this case f can be called a* rigid motion*, because f is included into a continuous family of isometries $f_t : \mathbb{R}^n \to \mathbb{R}^n$, $t \in [0,1]$, where $f_1 = f$ and f_0 is the identity map $f_0(p) = p$ for any $p \in \mathbb{R}^n$.* ■

Any isometry of $\mathbb{R}^n$ can be decomposed into at most $n+1$ reflections over hyperspaces, for example over planes in $\mathbb{R}^3$, hence is bijective and can be inverted. A composition of isometries is also an isometry and defines the operation in the group $\mathrm{Iso}(\mathbb{R}^n)$ of all isometries in $\mathbb{R}^n$. Rigid motions are orientation-preserving isometries and form the smaller subgroup $\mathrm{Iso}^+(\mathbb{R}^n) \subset \mathrm{Iso}(\mathbb{R}^n)$.

Example rigid motions in $\mathbb{R}^3$ are translations by vectors and rotations around straight lines. It suffices to classify sets up to general isometries, because if we know that two point sets S, Q are isometric, one can easily check if a possible isometry $S \to Q$ preserves an orientation defined as follows.

Let a set $S \subset \mathbb{R}^n$ have $n+1$ points $p_0, \ldots, p_n$ that are not in any $(n-1)$-dimensional subspace. The determinant of the $n \times n$ matrix with columns $p_i - p_0$, $i = 1, \ldots, n$, is the signed volume of the parallelepiped spanned by these n vectors. The sign of this determinant can be considered as an *orientation* of S. An isometry f preserves an orientation if the determinant obtained from $f(p_0), \ldots, f(p_n) \in Q$ has the same sign as S.

For any $n \times n$ matrix A, recall that A^T denotes the *transpose* matrix with elements $A^T_{ij} = A_{ji}$, $i, j = 1, \ldots, n$. A matrix A is *orthogonal* if the inverse matrix A^{-1} equals the transpose A^T. Orthogonality of a matrix A means that $\vec{v} \mapsto A\vec{v}$ maps any orthonormal basis to another orthonormal basis. All orthogonal matrices A have the determinant $\det A = \pm 1$. If $\det A = 1$, then the map $\vec{v} \mapsto A\vec{v}$ preserves an orientation of $\mathbb{R}^n$.

All orthogonal matrices A with $\det A = 1$ form the special orthogonal group $\mathrm{SO}(\mathbb{R}^n)$, where the operation is the matrix multiplication. The group $\mathrm{SO}(\mathbb{R}^2)$ consists of rotations about the origin in the plane. The group $\mathrm{SO}(\mathbb{R}^3)$ consists of rotations about axes passing

through the origin in $\mathbb{R}^3$. All orientation-preserving isometries in $\mathbb{R}^n$ can be decomposed into translations and high-dimensional rotations $R \in \mathrm{SO}(\mathbb{R}^n)$ around the origin in $\mathbb{R}^n$.

For a given equivalence relation, any objects can be distinguished (claimed to non-equivalent) only by an invariant is preserved the given equivalence hence is independent of a representation of an object. Surprisingly many descriptors of crystals include non-invariants, for example parameters of an ambiguous unit cell or atomic coordinates in an arbitrary cell basis.

Definition 18 (isometry invariants)**.** *An* isometry class *of sets is a collection of all sets that are isometric to each other, i.e. any sets S, Q from the same class are related by an isometry $S \to Q$. An* isometry invariant *is a function I that maps all sets of a given type, for example all periodic point sets, to a simpler space (numbers, matrices) so that $I(S) = I(Q)$ for any isometric sets S, Q. An invariant I is called* complete *if the converse is also true: if $I(S) = I(Q)$, then S, Q are isometric.* ■

Proof of Lemma 7. Let $S = \Lambda(S) + (U(S) \cap S)$ and $Q = \Lambda(Q) + (U(Q) \cap Q)$, where $U(S), U(Q)$ are initial unit cells of S, Q and the lattices $\Lambda(S), \Lambda(Q)$ contain the origin. By shifting all points of S, Q (but not their lattices), we guarantee that S contains the origin 0 of $\mathbb{R}^n$. Assume by contradiction that the given periodic point sets S, Q have no common lattice. Then there is a vector $\vec{p} \in \Lambda(S)$ whose all integer multiples $k\vec{p} \notin \Lambda(Q)$ for $k \in \mathbb{Z} - 0$. Any such multiple $k\vec{p}$ can be translated by a vector $\vec{v}(k) \in \Lambda(Q)$ to the initial unit cell $U(Q)$ so that $\vec{q}(k) = k\vec{p} - \vec{v}(k) \in U(Q)$.

Since $U(Q)$ contains infinitely many points $\vec{q}(k)$, one can find a pair $\vec{q}(i), \vec{q}(j)$ at a distance less than $\delta = r(Q) - d_B(S, Q) > 0$. The formula $\vec{q}(k) \equiv k\vec{p} \pmod{\Lambda(Q)}$ implies that $\vec{q}(i + k(j - i)) \equiv (i + k(j - i))\vec{p} \pmod{\Lambda(Q)} \equiv \vec{q}(i) + k(\vec{q}(j) - \vec{q}(i)) \pmod{\Lambda(Q)}$. If the point $\vec{q}(i) + k(\vec{q}(j) - \vec{q}(i))$ belongs to $U(Q)$, we get the equality $\vec{q}(i + k(j - i)) = \vec{q}(i) + k(\vec{q}(j) - \vec{q}(i))$. All these points over $k \in \mathbb{Z}$ lie on a straight line within $U(Q)$ and have the distance $|\vec{q}(j) - \vec{q}(i)| < \delta$ between successive points.

The closed balls with radius $d_B(S, Q)$ and centers at points in Q are at least 2δ away from each other. Then one of the points $\vec{q}(i + k(j - i))$ is more than $d_B(S, Q)$ away from Q. Hence the point $(i + k(j - i))\vec{p} \in S$ also has a distance more than $d_B(S, Q)$ from any point of Q, which contradicts Definition 6. □

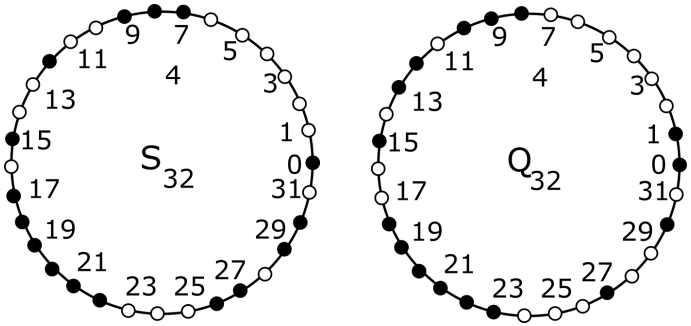

Figure 10. Circular versions of the periodic point sets S_{32}, Q_{32}, which are distinguished by AMD_2. Distances between any points are measured along round arcs in each circle.

The density functions narrowly distinguish the following periodic sets [17, Example 1]

$$S_{32} = \{0, 7, 8, 9, 12, 15, 17, 18, 19, 20, 21, 22, 26, 27, 29, 30\} + 32\mathbb{Z},$$

$$Q_{32} = \{0, 1, 8, 9, 10, 12, 13, 15, 18, 19, 20, 21, 22, 23, 27, 30\} + 32\mathbb{Z}$$

in Fig. 10. Tables 3, 4 distinguish these similar sets by AMD_k for $k = 2, 3$.

S_{32}	0	7	8	9	12	15	17	18	19	20	21	22	26	27	29	30	AMD_k
$k=1$	2	1	1	1	3	2	1	1	1	1	1	1	1	1	1	1	20/16
$k=2$	3	2	1	2	3	3	2	1	1	1	1	2	3	2	2	2	31/16
$k=3$	5	5	4	3	4	3	2	2	2	2	2	3	4	3	3	3	50/16

Table 3. First row: 16 points from the motif $M \subset S_{32}$ in Fig. 10. **Further rows**: distance from each point $p \in M$ to its k-th nearest neighbor in S_{32}.

Q_{32}	0	1	8	9	10	12	13	15	18	19	20	21	22	23	27	30	AMD_k
$k=1$	1	1	1	1	1	1	1	2	1	1	1	1	1	1	3	2	20/16
$k=2$	2	3	2	1	2	2	2	3	2	1	1	1	1	2	4	3	32/16
$k=3$	5	6	4	3	2	3	3	3	3	2	2	2	2	3	5	3	51/16

Table 4. First row: 16 points from the motif $M \subset Q_{32}$. **Further rows**: distance from each point $p \in M$ to its k-th nearest neighbor in Q_{32}, see Fig. 10.

B Appendix B: homometric sets and visualisations

This section defines homometric sets, which had different interpretations in past papers [20,32]. These sets are hard to distinguish, because they have identical diffraction patterns depending only on the difference set below.

Definition 19 (homometric sets)**.** *For any finite set $S \subset \mathbb{R}^n$ of m points, the* difference *multi-set* $\mathrm{Dif}(S)$ *consists of the m^2 vector differences $\vec{a}-\vec{b}$ for all points $a, b \in S$, counted with multiplicities. Periodic point sets $S, Q \subset \mathbb{R}^n$ with a common lattice Λ and a primitive unit cell U are* homometric *if* $\mathrm{Dif}(S \cap U) \equiv \mathrm{Dif}(Q \cap U) \pmod{\Lambda}$ *with multiplicities respected, so all pairs of vectors $u \in \mathrm{Dif}(S \cap U)$ and $v \in \mathrm{Dif}(Q \cap U)$ that are equal up to lattice translations have the same multiplicity in both difference sets. The* homometry *is the equivalence relation of periodic point sets being homometric.* ■

If a set S consists of m points, $\mathrm{Dif}(S)$ includes the zero vector with multiplicity m. The above definition clarifies the past attempts below.

Patterson [32, p.197] called periodic point sets $S, Q \subset \mathbb{R}^n$ *homometric* if $\mathrm{Dif}(S \cap U) \equiv \mathrm{Dif}(Q \cap U) \pmod{\Lambda}$ without mentioning weights or multiplicities. Franklin [20, equations (17)-(18)] renamed them as *homometric modulo a lattice* Λ and called S, Q *homometric* if $\mathrm{Dif}(S \cap U) = \mathrm{Dif}(Q \cap U)$, not modulo the lattice Λ. Additionally both definitions required that S, Q are not isometric. However, after removing this restriction, we expect to get an equivalence relation so that any periodic point set should be homometric to itself even if another unit cell of a lattice Λ is chosen.

The following example shows that the equation $\mathrm{Dif}(S \cap U) = \mathrm{Dif}(Q \cap U)$ without translations by the lattice Λ fails this reflexivity condition.

Franklin [20, p. 699] considered the sets $S_3 = \{0, 1\} + 3\mathbb{Z}$ and $Q_3 = \{0, 2\} + 3\mathbb{Z}$, which are isometric by $x \mapsto x + 1 \pmod 3$. However, $\mathrm{Dif}(\{0, 1\}) = \{0, 0, -1, +1\} \neq \mathrm{Dif}(\{0, 2\}) = \{0, 0, -2, +2\}$. These sets are equal modulo 3, hence lattice translations are needed. If we consider the set S_3 with a twice larger unit cell and period 6 as $\{0, 1, 3, 4\} + 6\mathbb{Z}$, then

$$\mathrm{Dif}(\{0, 1, 3, 4\}) = \{0, 0, 0, 0, \pm 1, \pm 1, \pm 2, \pm 3, \pm 3, \pm 4\}.$$

This difference set can be considered equal to $\mathrm{Dif}(\{0, 1\})$ modulo 3 only if the multiplicities in both sets are normalised so that their sums are equal. Hence Definition 19

requires a primitive unit cell U. Most importantly, Proposition 20 below justifies that homometry in Definition 19 is independent of a primitive unit cell and is an *equivalence* relation satisfying

(1) *reflexivity* : a periodic set S is homometric to S, i.e. $S \sim S$;

(2) *symmetry* : if S is homometric to Q, i.e. $S \sim Q$, then $Q \sim S$;

(3) *transitivity* : if $S \sim Q$ and $Q \sim T$, then $S \sim T$.

Proposition 20 makes the experimental concept of homometric crystals verifiable in an algorithmic way. It might be a folklore result, but we couldn't find a proof in the literature.

Proposition 20 (algorithm for homometric sets)**.** *(a) For any periodic point set $S \subset \mathbb{R}^n$ with a lattice Λ, the difference set* $\mathrm{Dif}(S \cap U) \pmod{\Lambda}$ *does not depend on a primitive unit cell U of S. So the homometry in Definition 19 is an equivalence relation.*

(b) Given a common primitive unit cell U containing m points of periodic point sets $S, Q \subset \mathbb{R}^n$, there is an algorithm of complexity $O(m^2 \log m)$ to determine if S, Q are homometric. ■

Proof. Let U, U' be primitive cells of the periodic set $S \subset \mathbb{R}^n$. Any point $q \in S \cap U'$ can be translated along a vector $\vec{v} \in \Lambda$ to a point $p \in S \cap U$ and vice versa. These translations establish a bijection $S \cap U \leftrightarrow S \cap U'$, which can change any point only by a vector of Λ. So $\mathrm{Dif}(S \cap U) \equiv \mathrm{Dif}(S \cap U') \pmod{\Lambda}$ with multiplicities respected by the bijection above.

To determine if periodic sets $S, Q \subset \mathbb{R}^n$ are homometric by Definition 19, first we compute all $O(m^2)$ pairwise vector differences between points from the motifs $S \cap U$ and $Q \cap U$. To check if these vector sets coincide, we could lexicographically order them in time $O(m^2 \log m)$, e.g. by using coordinates in the basis of the cell U. Then a single pass over $O(m^2)$ vector differences is enough to decide if $\mathrm{Dif}(S) \equiv \mathrm{Dif}(Q) \pmod{\Lambda}$. □

We illustrate Proposition 20 for $S(1) = \{0, 1, 3, 4\} + 8\mathbb{Z}$ and $Q(1) = \{0, 3, 4, 5\} + 8\mathbb{Z}$ in Fig. 3. Their 4-point motifs have distinct difference sets:

S_8	0	1	3	4
0	0	−1	−3	−4
1	1	0	−2	−3
3	3	2	0	−1
4	4	1	3	0

and

Q_8	0	3	4	5
0	0	−3	−4	−5
3	3	0	−1	−2
4	4	1	0	−1
5	5	2	1	0

The difference sets coincide modulo 8 with multiplicities shown as subscripts:

$$\mathrm{Dif}(S(1)) \equiv \{0_4, 1_2, 2_1, 3_2, 4_2, 5_2, 6_1, 7_2\} \equiv \mathrm{Dif}(Q(1)).$$

Then $S(1), Q(1)$ are homometric by Definition 19. The equivalence of differences modulo 8 gives rise to a bijection between all 16 elements of the matrices above, hence to a bijection between the sets of vector differences $\mathrm{Dif}(S(1)) \to \mathrm{Dif}(Q(1))$. For example, the difference $(8i+1)-(8j+4) = 8(i-j)-3 \equiv 5 \pmod 8$ in $S(1)$ can be bijectively mapped to the vector difference $(8i+5)-8j = 8(i-j)+5$ in $Q(1)$.

Previously it was impossible to visualise a large dataset of diverse crystals, because traditional comparison tools such as the COMPACK algorithm in the Mercury software [13] are slow and designed for pairwise comparisons of crystals with the same chemical composition or the same symmetry group. The Crystal Isometry Principle (CRISP) about injectivity of the map {periodic crystals} $\to$ {periodic point sets} allows us to study crystal similarities without restricting their symmetry group or composition. Mendeleev's table similarly parameterises all chemical elements by their period and group number.

Any crystal dataset can be considered as a discrete sample from the common continuous space of all periodic point sets. Though the ambient continuous space is high-dimensional, one can easily visualise [33] any finite subset in this space as a Minimum Spanning Tree (MST). This MST spans all given crystals represented by vertices and minimises the total length of edges (distances between crystal invariants). Hence any crystal and its nearest neighbor are always connected in any MST. Fig. 11 shows such MST for 12576 crystalline drugs, which was computed within one hour on a modest desktop.

The earlier version [41] of this paper has more Minimum Spanning Trees of larger datasets in appendices, which are being extended into another applications paper.

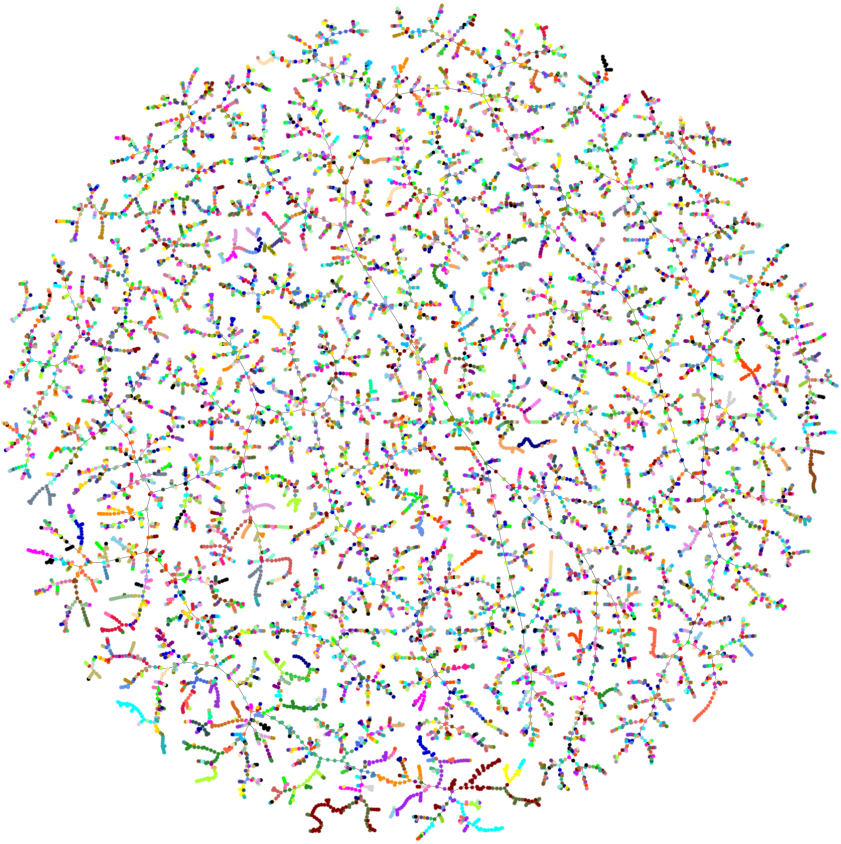

Figure 11. TMap of all 12576 crystalline drugs in the Cambridge Structural Database (CSD), based on L_∞-distances between $\mathrm{AMD}^{(200)}$ vectors. All drugs in the same CSD family have the same (random) color.