Recognition of near-duplicate periodic patterns by continuous metrics with approximation guarantees

Olga D Anosova[a], Daniel E Widdowson[a], Vitaliy A Kurlin[a,*]

[a] *Department of Computer Science, University of Liverpool, Liverpool L69 3BX, UK*

Abstract

This paper rigorously solves the challenging problem of recognizing periodic patterns under rigid motion in Euclidean geometry. The 3-dimensional case is practically important for justifying the novelty of solid crystalline materials (periodic crystals) and for patenting medical drugs in a solid tablet form.

Past descriptors based on finite subsets fail when a unit cell of a periodic pattern discontinuously changes under almost any perturbation of atoms, which is inevitable due to noise and atomic vibrations. The major problem is not only to find complete invariants (descriptors with *no false negatives* and *no false positives* for all periodic patterns) but to design efficient algorithms for distance metrics on these invariants that should continuously behave under noise.

The proposed continuous metrics solve this problem in any Euclidean dimension and are algorithmically approximated with small error factors in times that are explicitly bounded in the size and complexity of a given pattern.

The proved Lipschitz continuity allows us to confirm all near-duplicates filtered by simpler invariants in major databases of experimental and simulated crystals. This practical detection of noisy duplicates will stop the artificial generation of 'new' materials from slight perturbations of known crystals. Several such duplicates are under investigation by five journals for data integrity.

Keywords: periodic pattern, isometry invariant, continuous distance metric

*The corresponding author was supported by the Royal Society APEX fellowship (APX/R1/231152), Royal Academy of Engineering Fellowship (IF2122/186) at the Cambridge Crystallographic Data Center, and the EPSRC New Horizons grant (EP/X018474/1).

Email address: vkurlin@liverpool.ac.uk, http://kurlin.org (Vitaliy A Kurlin)

2020 MSC: 52C25, 52C07, 51N20, 68W25

1. Continuous metric problem for periodic point sets and crystals

In Euclidean geometry, periodic sets of points model all periodic crystals since any atom has a physically meaningful nucleus represented by an atomic center [1]. This approach is more fundamental than using graphs with chemical bonds, which are not real sticks and only abstractly representing inter-atomic interactions depending on various thresholds for distances and angles [2].

A *lattice* $\Lambda \subset \mathbb{R}^n$ is the infinite set of all integer linear combinations $\sum_{i=1}^{n} c_i v_i$ of a basis $v_1, \dots, v_n$ of Euclidean space $\mathbb{R}^n$. Any basis defines a parallelepiped U called a primitive *unit cell* of Λ. The first picture in Fig. 1 shows different unit cells in red, green, and blue, which generate the same hexagonal lattice.

A *periodic point set* $S \subset \mathbb{R}^n$ is a finite union of lattice translates $\Lambda + p$ obtained from Λ by shifting the origin to a point p from a finite *motif* $M \subset U$.

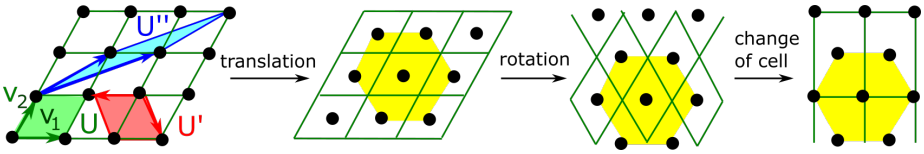


Figure 1: **Left**: three (of infinitely many) primitive cells U, U', U'' of the same minimal area for the hexagonal lattice Λ. **Other images** show periodic sets $\Lambda + M$ with different cells and motifs, which are all isometric to Λ whose hexagonal Voronoi domain is highlighted in yellow.

This paper is motivated by the growing crisis of artificial data in crystallography [3]because a slight modification of a known material can be claimed as 'new'. Indeed, the fundamental question "same or different" [4] was not rigorously answered for periodic crystals. In the mathematical language, what periodic crystals can we consider equivalent, i.e. in the same class under an equivalence relation? One classical equivalence between crystals is by symmetry, e.g. crystallographic space groups were classified into 230 types (if mirror images are distinguished) already in the 19th century by Fedorov and Schonflies.

In 2024, experimental databases, e.g. the Cambridge Structural Database (CSD) of 1.25+ million real materials [5], need much stronger classifications than by 230 space groups or by chemical compositions. In fact, many crystals such as diamond and graphite consist of the same elements but have vastly different properties due to essential differences in their geometric structures.

Structures of periodic crystals are experimentally determined in a rigid form. Hence their most practical equivalence is *rigid motion*, which is a composition of translations and rotations in $\mathbb{R}^n$. The slightly weaker equivalence is *isometry* (any distance-preserving transformation), including mirror reflections.

An *isometry class* (or pattern or or *orbit* under the action of all Euclidean isometries) consists of all sets that are isometric to each other. All isometry classes of periodic point sets form a continuously infinite space. Indeed, almost any perturbation of points such as atomic displacements caused by noise in data produces a non-isometric crystal, which might have an arbitrarily scaled-up primitive cell, an enlarged motif, and a different symmetry group as in Fig. 2.

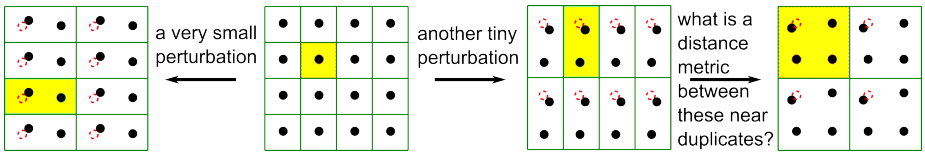


Figure 2: Past descriptors based on a primitive cell cannot continuously quantify a distance between near-duplicates. For example, under almost any perturbation, symmetry groups break down and a primitive cell volume discontinuously changes up to any integer factor.

Perturbations in Fig. 2 can be applied to any periodic crystal and mistakenly used for re-formatting old crystals as 'new' by additionally replacing atomic types with similar ones [6]. The first cases of geometric near-duplicates were exposed in the CSD, see [7, section 7], [8, section 6], and on a much larger scale in Google's GNoME database, which was reviewed in [9], [3, Tables 1-2]. To avoid machine learning on skewed data, the filtering of near-duplicates is necessary for any database of simulated or experimental objects, e.g. point clouds, that can have infinitely many representations in different coordinate systems.

A pseudo-symmetry approach, for a threshold $\varepsilon > 0$, calls periodic sets *equivalent* if their cell parameters and atomic coordinates differ by at most ε [10]. Then any sets can be joined by a long enough chain of ε-perturbations [11, Prop 2.1]. If we allow any threshold $\varepsilon > 0$, the transitivity axiom (if $A \sim B \sim C$, then $A \sim C$) implies that all periodic point sets in $\mathbb{R}^n$ become *equivalent*.

A mathematical approach to noisy data is to quantify perturbations by a distance metric satisfying all axioms in Definition 1.1 below and taking small positive values on pairs of sets in Fig.2, which is formalized in Problem 1.2(a) .

Definition 1.1 (metric)**.** A *metric* on isometry classes of periodic sets of unordered points in $\mathbb{R}^n$ is a real-valued function d satisfying these axioms:
(1.1a) $d(S,Q) = 0$ if and only if sets S, Q are isometric (denoted by $S \simeq Q$);
(1.1b) symmetry: $d(S,Q) = d(Q,S)$ for any periodic point sets S, Q in $\mathbb{R}^n$;
(1.1c) triangle inequality: $d(S,Q) + d(Q,T) \geq d(S,T)$ for any S, Q, T. ■

Without the first axiom in (1.1a), even the zero function $d(S,Q) = 0$ satisfies Definition 1.1a. The atomic vibrations [1, chapter 1] motivate a metric whose continuity is quantified via a maximum displacement of atoms in (1.2a) below.

Problem 1.2 (continuous metric on periodic sets)**.** Find a metric d on periodic point sets in $\mathbb{R}^n$ such that all the metric axioms of Definition 1.1 hold and

(1.2a) d is Lipschitz continuous : there is a constant $\lambda > 0$ such that, for any sufficiently small $\varepsilon > 0$, if Q is obtained from any periodic set $S \subset \mathbb{R}^n$ by perturbing each point of S within its ε-neighborhod, then $d(S,Q) \leq \lambda\varepsilon$;
(1.2b) $d(S,Q)$ is computed or approximated (up to an explicit error factor) in a time that has a polynomially upper bound in the sizes of motifs of S, Q. ■

Problem 1.2 can be widened to any real data (instead of crystals) and equivalences (instead of isometry). Condition (1.2a) goes beyond a complete classification of periodic point sets modulo isometry. Indeed, any metric d satisfying (1.2a) detects all non-isometric sets $S \not\simeq Q$ by checking if $d(S,Q) \neq 0$. Con-

versely, detecting an isometry $S \simeq Q$ gives only a discontinuous metric d , e.g. $d(S,Q) = 1$ for any non-isometric $S \not\simeq Q$ and $d(S,Q) = 0$ for any $S \simeq Q$.

For finite sets under isometry, the persistent homology turned out to be weaker than anticipated [12]. In this case, Problem 1.2 was solved by easier and faster invariants [13, 14]. In the periodic case, Problem 1.2 was solved in dimension $n = 1$ [15] and for lattices in $\mathbb{R}^2$ [11, 16] but was open for $n > 2$.

Accuracies such as precision, recall, and F1-score make sense for finitely many classes with (usually manual) labels. However, experimental noise (or thermal vibrations of atoms) always produce slightly different objects whose deviations should be quantified by a continuous distance metric. Hence the real ground truth in many applications is not one of finitely many labels but an experimental structure within a continuous space of all potential objects. Using crystals as an example, Problem 1.2 states the necessary conditions towards continuous machine learning for any data including real (non-discrete) values.

This paper solves Problem 1.2 by defining a continuous metric on the complete invariant *isoset* from [17]. The first step introduces a boundary tolerant metric BT on local clusters around points of a periodic set S, which continuously changes when points cross a cluster boundary. This discontinuity at the boundary can be formally resolved by an extra factor, which smoothly goes down to 0 depending on an extra parameter. Without using extra parameters, the new metric BT will be exactly expressed in terms of simpler distances.

The second step uses the Earth Mover's Distance [18] to extend BT to complete invariants [17] that are weighted distributions of local clusters up to rotations. The resulting metric on periodic sets in $\mathbb{R}^n$ is approximated with a factor η, e.g. $\eta \approx 4$ in $\mathbb{R}^3$, in a time depending polynomially on the input size.

The third step proves the metric axioms and continuity $d(S,Q) \le 2\varepsilon$, which also has practical importance. Indeed, if $d(S,Q)$ is approximated by a value d with a factor η, we get the lower bound $\varepsilon \ge \frac{d}{2\eta}$ for the maximum displacement ε of points. Such a lower bound is impossible to guarantee by analyzing only finite subsets, which can be very different in identical periodic sets, see Fig. 3.

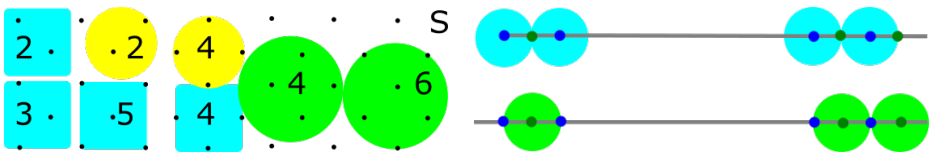


Figure 3: **Left**: for any lattice S and a fixed size of a box or a ball, one can choose many non-isometric finite subsets of different sizes. **Right**: the blue set S and green set Q in the line $\mathbb{R}$ have a small Hausdorff distance $d_H = \varepsilon$ but are not related by a small perturbation.

2. Past work on distances and invariants of periodic point sets

This section clarifies that all past descriptors of periodic crystals are either discontinuous under perturbations as in Fig. 2 or were not proved to be complete under rigid motion. Problem 1.2 was open for periodic sets for $n > 2$.

One can try comparing periodic point sets by finding an isometry of $\mathbb{R}^n$ that makes them as close as possible [19]. This approximate matching is much easier for finite sets. Hence it is very tempting to restrict any periodic point set to a large rectangular box or a cube with identified opposite sides (a fixed 3D torus). However, differently located boxes or balls of any fixed size can contain non-isometric finite sets as shown in Fig. 3 (left) for the square lattice. Then extra justifications are needed to show that a comparison of periodic sets by their finite subsets does not depend on the choices of these finite subsets.

Definition 2.1 (Hausdorff distance d_H, bottleneck distance d_B)**. (a)** For any sets S, Q in a metric space, $d_{\vec{H}}(S,Q) = \sup\limits_{p\in S} \inf\limits_{q\in Q} d(p,q)$ is the *directed Hausdorff* distance. The *Hausdorff distance* is $d_H(S,Q) = \max\{d_{\vec{H}}(S,Q), d_{\vec{H}}(Q,S)\}$.

(b) The *bottleneck distance* $d_B(S,Q) = \inf\limits_{g:S\to Q} \sup\limits_{p\in S} d(p,g(p))$ for sets S, Q of the same cardinality is minimized over bijections g and maximized over $p \in S$. ■

Fig. 3 (right) shows the sets S, Q consisting of blue and green points, respectively, where all green points of Q are covered by small closed blue balls centered at all points of S in the top right picture, and vice versa. Hence a small Hausdorff distance $d_H(S,Q)$ doesn't guarantee that the sets S, Q are related by

a small perturbation of points. A non-bijective matching of points is inappropriate for real atoms that cannot disappear and reappear from thin air. Hence the bottleneck distance d_B is more suitable for measuring atomic displacements than d_H. [8, Example 2.1] shows that the 1-dimensional lattices $\mathbb{Z}$ and $(1+\delta)\mathbb{Z}$ have $d_B = +\infty$ for any $\delta > 0$. If any lattices have equal density (or unit cell volume), they have a finite bottleneck distance d_B by [20, 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), the bottleneck distance $d_B(S, Q)$ becomes a well-defined wobbling distance [21], which is still discontinuous under perturbations by [8, Example 2.2], see the related results for non-periodic sets in [22, 23]

Another approach to comparing crystals is by Voronoi diagrams, which can be defined for periodic sets but remain combinatorially unstable as for finite sets. Under almost any perturbation of basis vectors in $\mathbb{R}^2$, a rectangular lattice becomes generic with a hexagonal Voronoi domain. Hence combinatorial descriptors of Voronoi domains discontinuously change under perturbations of non-generic sets as in Fig. 2. Geometric descriptors such as the area or volume can be continuously compared by the Hausdorff distance and helped define two continuous metrics between lattices in $\mathbb{R}^n$ [24], though their implementation sampled finitely many rotations without approximation guarantees.

Other comparisons of periodic sets use a manually chosen number of neighbors or a cut-off radius [19]. A reduction to a finite subset cannot provide a complete and continuous invariant of periodic sets because, under tiny perturbations, a *primitive* (minimal by volume) cell can become larger than any bounded subset of a fixed size, see Fig. 2. One can guarantee the continuity under perturbations by extra smoothing at a fixed cut-off radius so that non-matched points covertly cross a fixed boundary, e.g. [25] starts from a Gromov-Hausdorff distance between finite sets of any sizes and adds terms converging to 0 at the boundary. The continuity was shown for three motions [25, Fig. 3,4,5] but the triangle inequality needs a proof, else clustering may not be trustworthy [26].

Crystallographers often compared periodic crystals by using reduced or conventional cells. In $\mathbb{R}^2$, a cell with basis vectors $\vec{v}_1, \vec{v}_2$ is *reduced* if $|\vec{v}_1| \le |\vec{v}_2|$ and $-\frac{1}{2}\vec{v}_1^2 \le \vec{v}_1 \cdot \vec{v}_2 \le 0$. The vectors $\vec{v}_1 = (2a, 0)$ and $\vec{v}_2^{\pm} = (-a, \pm b)$ for $b \ge a\sqrt{3}$ and both signs $\pm$ are reduced and define isometric lattices related by reflection. This ambiguity of bases can be resolved by an additional condition $\det(\vec{v}_1, \vec{v}_2) > 0$, which creates the inevitable discontinuity, see more details in [11, Fig. 4]. In $\mathbb{R}^3$, the most widely used reduced cell is Niggli's cell, which has a minimum volume and all angles as close to 90° as possible. Niggli's cell was known to be experimentally discontinuous since 1965 [27] or even earlier due to Fig. 2.

In $\mathbb{R}^3$, all generic periodic sets are distinguished by density functions [28], which can be computed at discrete values of a continuous radius $t \in \mathbb{R}$ [29]. A metric between density functions was defined in terms of suprema over infinitely many $t \in \mathbb{R}$, so the metric was approximated without guarantees. The density functions [30] coincide for the periodic sets $S_{15} = X + Y + 15\mathbb{Z}$, $Q_{15} = X - Y + 15\mathbb{Z}$, where $X = \{0, 4, 9\}$ and $Y = \{0, 1, 3\}$ [31, Example 11]. This pair and all generic periodic sets are distinguished by a faster Pointwise Distance Distribution (PDD) due to [8, Theorem 4.4] whose averages AMD are incomplete by [8, Example 3.3] but the PDD also cannot distinguish any mirror images.

A distance between invariant values can be a metric on isometry classes only if the underlying invariant is complete under isometry. Otherwise, non-isometric sets can have identical invariant values with a distance of 0. Hence a complete classification should take into account a potential high complexity of periodic sets. Inspired by [32, 33], the isometry classification of periodic sets was reduced [17] to only rotations of local clusters whose radius can be determined from S.

Section 3 reminds us of a complete invariant isoset from [17]. Section 4 introduces a Lipschitz continuous metric (Definition 4.4 and Theorem 4.9), whose polynomial time bounds (Corollaries 5.4, 5.10) are proved in section 5. Section 6 proves a lower bound (Theorem 6.5) for the new metric via faster PDDs. Section 7 discusses the significance of the continuous metric for detecting near-duplicates in major crystal databases and for upholding scientific integrity.

3. Isometry classification of periodic point sets by complete invariants

This section reviews the complete invariant [17] based on local clusters and their symmetry groups, which were previously studied in [32, 33].

Definition 3.1 (*global* clusters and *m-regular* periodic sets)**.** For any point p in a periodic set $S \subset \mathbb{R}^n$, the *global cluster* is $C(S,p) = \{\vec{q}-\vec{p} \,:\, q \in S\}$. For any $p,q \in \mathbb{R}^n$, let the set $\mathrm{O}(\mathbb{R}^n;p,q)$ consist of all isometries of $\mathbb{R}^n$ that map p to q. Global clusters $C(S,p)$ and $C(S,q)$ are called *isometric* if there is $f \in \mathrm{O}(\mathbb{R}^n;p,q)$ such that $f(S)=S$. A periodic point set $S \subset \mathbb{R}^n$ is called *m-regular* if all global clusters of S form exactly $m \geq 1$ isometry classes. ■

For any point $p \in S$, its global cluster is a view of S from the position of a point p. We view all astronomical stars in the universe S from our planet Earth at p. Any lattice is 1-regular since all its global clusters are related by translations. Though global clusters $C(S,p), C(S,q)$ at any different points $p,q \in S$ contain the same set S, they may not match under the translation shifting p to q. The global clusters are infinite, hence distinguishing them up to isometry is not easier than original periodic sets. However, the m-regularity of a periodic set can be checked in terms of finite local α-clusters below.

Definition 3.2 (*local α-clusters* $C(S,p;\alpha)$ and *symmetry groups* $\mathrm{Sym}(S,p;\alpha)$)**.** For a point p in a periodic point set $S \subset \mathbb{R}^n$ and any $\alpha \geq 0$, the local *α-cluster* $C(S,p;\alpha)$ is the set of all vectors $\vec{q}-\vec{p}$ such that $q \in S$ and $|\vec{q}-\vec{p}| \leq \alpha$. Let the group $\mathrm{O}(\mathbb{R}^n;p)$ consist of all isometries that fix p. If $p=0$ is the origin, $\mathrm{O}(\mathbb{R}^n;0)$ is the usual orthogonal group. The *symmetry* group $\mathrm{Sym}(S,p;\alpha)$ consists of all isometries $f \in \mathrm{O}(\mathbb{R}^n;p)$ that map $C(S,p;\alpha)$ to itself so that $f(p)=p$. ■

For any periodic set S, if α is smaller than the minimum distance between all points of S, then any α-cluster $C(S,p;\alpha)$ is one point $\{p\}$. Its symmetry group consists of all isometries fixing the center p, so $\mathrm{Sym}(S,p;\alpha) = \mathrm{O}(\mathbb{R}^n;p)$. When α is increasing, the α-clusters $C(S,p;\alpha)$ become larger and there can be fewer (not more) isometries $f \in \mathrm{O}(\mathbb{R}^n;p)$ that bijectively map $C(S,p;\alpha)$ to itself. So

the group $\mathrm{Sym}(S, p; \alpha)$ can become smaller (not larger) and eventually stabilizes (stops changing), which will be formalized later in Definition 3.5.

Definition 3.3 (*bridge length* $\beta(S)$)**.** For a periodic point set $S \subset \mathbb{R}^n$, the *bridge length* is a minimum distance $\beta(S) > 0$ such that any $p, q \in S$ can be connected by a sequence of points $p_0 = p, p_1, \dots, p_k = q$ such that any two successive points p_{i-1}, p_i are close so that $|\vec{p}_{i-1} - \vec{p}_i| \le \beta(S)$ for $i = 1, \dots, k$. ■

The theorem from [32, p. 20] proves that any 1-regular periodic point set is uniquely determined (up to isometry) by one sufficiently large α-cluster. [33, Theorem 1.3] describes how a family of clusters uniquely determines a periodic point set up to isometry. These results motivated the concepts of the *isotree*, *stable* radius, and *isoset* in Definitions 3.4, 3.5, 3.8, respectively, leading to the isometry classification of periodic point sets via isosets in Theorem 3.10. The *isotree* in Definition 3.4 is inspired by a clustering dendrogram but points of S split into isometry classes of α-clusters at different radii α, not at a fixed α.

Definition 3.4 (*isotree* $\mathrm{IT}(S)$ of α*-partitions*)**.** Fix a periodic point set $S \subset \mathbb{R}^n$. Points $p, q \in S$ are α*-equivalent* if their α-clusters $C(S, p; \alpha)$ and $C(S, q; \alpha)$ can be related by an isometry that matches their centers. The *isometry class* $[C(S, p; \alpha)]$ consists of all α-clusters isometric to $C(S, p; \alpha)$. The α*-partition* $P(S; \alpha)$ is the splitting of S into α-equivalence classes of points. Call a value α *singular* if $P(S; \alpha) \neq P(S; \alpha - \varepsilon)$ for any small enough $\varepsilon > 0$. Represent each α-equivalence class by a vertex of the *isotree* $\mathrm{IT}(S)$. The top vertex of $\mathrm{IT}(S)$ represents the 0-equivalence class coinciding with S. For any successive singular values $\alpha < \alpha'$, connect the vertices representing any classes $A \in P(S; \alpha)$ and $A' \in P(S; \alpha')$ such that $A' \subset A$ by an edge of the length $\alpha' - \alpha$ in $\mathrm{IT}(S)$. ■

For any periodic point set $S \subset \mathbb{R}^n$, the root vertex of $\mathrm{IT}(S)$ at $\alpha = 0$ is the single class S, because any 0-cluster $C(S, p; 0)$ of a point $p \in S$ consists only of its center p. When the radius α is increasing, α-clusters $C(S, p; \alpha)$ include more points and hence may not be isometric. In other words, any α-equivalence class from $P(S; \alpha)$ may split into two or more classes, which cannot merge at

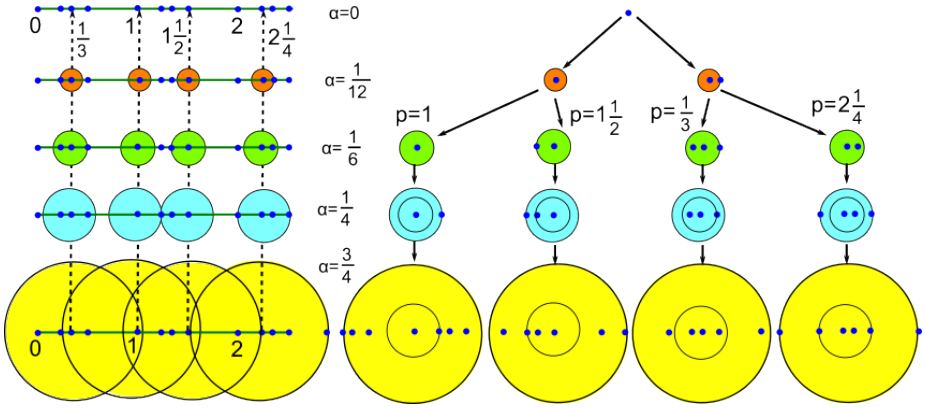


Figure 4: **Left**: the 1-dimensional set $S_4 = \{0, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}\} + \mathbb{Z}$ has four points in the unit cell $[0, 1)$ and is 4-regular by Definition 3.1. **Right**: the colored disks show α-clusters in the line $\mathbb{R}$ with radii $\alpha = 0, \frac{1}{12}, \frac{1}{6}, \frac{1}{4}, \frac{3}{4}$ and represent points in the isotree $\mathrm{IT}(S_4)$ from Definition 3.4.

any larger α'. Branched vertices of $\mathrm{IT}(S)$ correspond to the values of α when an α-equivalence class is split into subclasses for α' slightly larger than α. So the number $|P(S;\alpha)|$ of α-equivalence is non-decreasing in α, see Fig. 4.

The α-clusters of the 1-dimensional periodic point set $S_4 \subset \mathbb{R}$ in Fig. 4 are intervals in $\mathbb{R}$, shown as disks for better visibility. In Fig. 4, this class persists until $\alpha = \frac{1}{12}$, when all points $p \in S_4$ are split into two classes: one represented by 1-point cluster $\{p\}$ for $p \in \{0, \frac{1}{2}\} + \mathbb{Z}$, and another represented by 2-point clusters $\{p, p + \frac{1}{12}\}$, $p \in \{\frac{1}{4}, \frac{1}{3}\} + \mathbb{Z}$. The periodic set S_4 has four α-equivalence classes for any radius $\alpha \geq \frac{1}{6}$. For any point $p \in \mathbb{Z} \subset S_4$, the symmetry group $\mathrm{Sym}(S_4, p; \alpha) = \mathbb{Z}_2$ is generated by the reflection in p for $\alpha \in [0, \frac{1}{4})$. For all $p \in S_4$, the symmetry group $\mathrm{Sym}(S_4, p; \alpha)$ is trivial for any $\alpha \geq \frac{1}{4}$. For any periodic point set $S \subset \mathbb{R}^n$, the α-partitions of S stabilize in the sense below.

Definition 3.5 (the *minimum stable radius* $\alpha(S)$)**.** Let $S \subset \mathbb{R}^n$ be a periodic point, $\beta \geq \beta(S)$ be an upper bound of the bridge length $\beta(S)$ from Definition 3.3. A radius $\alpha \geq \beta$ is called *stable* if the following conditions hold:

(3.5a) the α-partition $P(S;\alpha)$ equals the $(\alpha - \beta)$-partition $P(S;\alpha - \beta)$;

(3.5b) the groups stabilize so that $\mathrm{Sym}(S, p; \alpha) = \mathrm{Sym}(S, p; \alpha - \beta)$ for any $p \in S$,

i.e. any isometry $f \in \mathrm{Sym}(S, p; \alpha - \beta)$ preserves the larger cluster $C(S, p; \alpha)$.

A minimum value of a stable radius α satisfying (3.5ab) for $\beta = \beta(S)$ from Definition 3.3 is called *the minimum stable radius* and denoted by $\alpha(S)$. ■

Due to the upper bounds in Lemma 3.6(b,c), the minimum stable radius $\alpha(S) \geq 0$ exists and is achieved because $P(S; \alpha)$ and $\mathrm{Sym}(S, p; \alpha)$ are continuous on the right (unchanged when α increases by a sufficiently small value).

Any m-regular periodic point set $S \subset \mathbb{R}^n$ has at most m α-equivalence classes, so the isotree $\mathrm{IT}(S)$ stabilizes with maximum m branches. Though (3.5b) is stated for all points $p \in S$ for simplicity, it suffices to check condition (3.5b) for points only from a finite motif M of S due to periodicity.

All stable radii of S form the interval $[\alpha(S), +\infty)$ by Lemma A.4 in the appendix. The periodic set S_4 in Fig. 4 has $\beta(S_4) = \frac{1}{2}$ and $\alpha(S) = \frac{3}{4}$ since the α-partition and symmetry groups $\mathrm{Sym}(S_4, p; \alpha)$ are stable for $\frac{1}{4} \leq \alpha \leq \frac{3}{4}$.

Condition (3.5b) doesn't follow from condition (3.5a) due to the following example. Let Λ be the 2D lattice with the basis $(1, 0)$ and $(0, \beta)$ for $\beta > 1$. Then β is the bridge length of Λ. Condition (3.5a) is satisfied for any $\alpha \geq 0$, because all points of any lattice are equivalent up to translations. However, condition (3.5b) fails for any $\alpha < \beta + 1$. Indeed, the α-cluster of the origin $(0, 0)$ contains five points $(0, 0), (\pm 1, 0), (0, \pm\beta)$, whose symmetries are generated by the two reflections in the axes x, y, but the $(\alpha - \beta)$-cluster of the origin $(0, 0)$ consists of its center and has the symmetry group $\mathrm{O}(\mathbb{R}^2)$. It is possible that condition (3.5b) might imply condition (3.5a), but in practice it makes sense to verify (3.5b) only after checking much simpler condition (3.5a). Both conditions are essentially used in the proof of Isometry Classification Theorem 3.10.

Conditions (3.5ab) appeared in [33] with different notations $\rho, \rho + t$. Since many applied papers use ρ for the physical density and have many types of bond distances, we replaced t and $\rho + t$ with the bridge length β and radius α, respectively, as for growing α-shapes in Topological Data Analysis [12].

Recall that the *covering radius* $R(S)$ of a periodic point set $S \subset \mathbb{R}^n$ is the minimum radius $R > 0$ such that $\bigcup\limits_{p \in S} \bar{B}(S; R) = \mathbb{R}^n$, or the largest radius of an open ball in the complement $\mathbb{R}^n \setminus S$. For m-regular point sets in $\mathbb{R}^n$, an upper bound of $\alpha(S)$ can be extracted from [33, Theorem 1.3] whose proof motivated a stronger bound in Lemma 3.6(c), see comparisons in Example 3.7(c).

A periodic point set S is *locally antipodal* if the local cluster $C(S, p; 2R(S))$ is centrally symmetric for any point $p \in S$, i.e. bijectively maps to itself under $\vec{q} \mapsto 2\vec{p} - \vec{q}$, $q \in \mathbb{R}^n$. [34, Theorem 1] says that all locally antipodal Delone sets, hence all periodic sets S, are globally antipodal, i.e. S is preserved under the isometry $\vec{q} \mapsto 2\vec{p} - \vec{q}$ for any fixed $p \in S$, e.g. any lattice is antipodal.

Lemma 3.6 (upper bounds for a stable radius $\alpha(S)$ and bridge length $\beta(S)$)**.**
(a) Let $S \subset \mathbb{R}^n$ be a periodic point set with a unit cell U, which has the longest edge b and longest diagonal d. Set $r(U) = \max\{b, \frac{d}{2}\}$. Then the bridge length $\beta(S)$ from Definition 3.3 has the upper bound $\min\{2R(S), r(U)\} \geq \beta(S)$.

(b) For any antipodal periodic set $S \subset \mathbb{R}^n$ whose covering radius is $R(S)$, the minimum stable radius has the upper bound $2R(S) + \beta(S) > \alpha(S)$.

(c) Let $S \subset \mathbb{R}^n$ be any periodic point set with the bridge length β. For any point $p \in S$ and a radius $\alpha_0 \geq 2R(S)$, the order $|\mathrm{Sym}(S, p; \alpha_0)|$ of the group $\mathrm{Sym}(S, p; \alpha_0)$ should be finite. Let $p_1, \dots, p_m \in S$ be all points of an asymmetric unit of S. Set $L = \left\lceil \sum\limits_{i=1}^{m} \left(\log_2 |\mathrm{Sym}(S, p_i; \alpha_0)| - \log_2 |\mathrm{Sym}(S, p_i)| \right) \right\rceil$. Then the minimum stable radius $\alpha(S)$ from Definition 3.5 has the upper bound $\alpha_0 + (L + m)\beta \geq \alpha(S)$. If $\alpha_0 = 2R(S)$, then $(L + m + 1)2R(S) \geq \alpha(S)$. ■

Proof. **(a)** The lemma in [32, section 2] proved that, in any Delone set S with the covering radius $R(S)$, any two points $p, q \in S$ can be connected by a finite sequence of points $p_0 = p, p_1, \dots, p_k = q$ such that $|\vec{p}_{i-1} - \vec{p}_i| \leq 2R(S)$ for $i = 1, \dots, k$. In particular, any periodic point set S has the upper bound $2R(S) \geq \beta(S)$. It remains to prove the second upper bound $r(U) \geq \beta(S)$.

For a point $p \in S$, shift the unit cell U so that p becomes the origin of $\mathbb{R}^n$ and a vertex of U, so the lattice Λ can be considered a subset of the periodic

point set S. Any points of Λ can be connected by a sequence of lattice points such that any successive points have a distance not greater than the longest edge-length b of U. Any point of a motif $M \subset U$ of S is at most $r(U)$ away from a vertex of U, where d is the length of the longest diagonal of U. Any points of S can be connected by a sequence whose successive points are at most $r(U) = \max\{b, \frac{d}{2}\}$ away from each other, so $\beta(S) \le r(U)$ by Definition 3.3.

(b) We will prove that the conditions of Definition 3.5 hold for $\alpha = 2R(S) + \beta(S)$ and $\beta = \beta(S)$. To prove condition 3.5(a), we check below that any $2R(S)$-equivalent points $p, q \in S$ are α-equivalent for any $\alpha > 2R(S)$. The $2R(S)$-equivalence means that there is an isometry $f \in \mathrm{O}(\mathbb{R}^n; p, q)$ such that $f(C(S, p; 2R(S))) = C(S, q; 2R(S))$. Set $Q = f(S)$. Then $f(C(S, p; 2R(S))) = C(f(S), f(p); 2R(S))$ means that $C(S, q; 2R(S)) = C(Q, q; 2R(S))$. [34, Theorem 3] implies that if antipodal periodic point sets $S, Q \subset \mathbb{R}^n$ have a common point q with $C(S, q; 2R(S)) = C(Q, q; 2R(S))$, then $S = Q$. In our case, $f(S) = S$ implies that f makes the points p and $q = f(p)$ α-equivalent for any $\alpha > 2R(S)$. Condition 3.5(b) says that any isometry $f \in \mathrm{Sym}(S, p; 2R(S))$ should belong to $\mathrm{Sym}(S, p; \alpha)$ for any point $p \in S$ and radius $\alpha > 2R(S)$. Indeed, [34, Theorem 3] implies that $Q = f(S)$ and S should coincide, so f isometrically maps any cluster $C(S, p; \alpha)$ to itself, hence $f \in \mathrm{Sym}(S, p; \alpha)$.

(c) Lemma A.7, which was briefly proved in [32, p. 20], says that the symmetry group $\mathrm{Sym}(S, p; 2R(S))$ is finite. For any initial radius $\alpha_0 \ge 2R(S)$, we aim to find a radius $\alpha = \alpha_0 + k\beta$ such that both conditions 3.5(a,b) hold for a suitable index $k = 1, 2, 3, \dots$ whose upper bound we will determine below.

If condition 3.5(a) fails for some $\alpha = \alpha_0 + k\beta$, the number $|P(S; \alpha_0 + (k-1)\beta)|$ of α-equivalence classes increases at least by one when $\alpha_0 + (k-1)\beta$ increases to $\alpha_0 + k\beta$. Since an asymmetric unit of S consists of $m \ge 1$ points, there are at most $m - 1$ incremental values $0 = k_0 \le k_1 \le \ldots \le k_{m-1}$ when $1 \le |P(S; \alpha_0 + (k_i - 1)\beta)| < |P(S; \alpha_0 + k_i\beta)| \le m$ for $i = 1, \dots, m-1$.

In a degenerate case, if all points of S are $(\alpha_0 + (k-1)\beta)$-equivalent, this single class can split into the maximum $m > 1$ classes of $(\alpha_0 + k\beta)$-equivalence,

then $k_1 = \ldots = k_{m-1} = k \geq 1$. For any successive incremental values $k_{i-1} < k_i$, the number $|P(S; \alpha_0 + k\beta)|$ of $(\alpha_0 + k\beta)$-equivalence classes is constant for $k = k_{i-1} + 1, \ldots, k_i$, so condition 3.5(a) holds for every radius $\alpha = \alpha_0 + k\beta$.

By reordering the points $p_1, \ldots, p_m$ from an asymmetric unit of S, we can assume that $p_1, \ldots, p_i$ represent i classes of $(\alpha + k_{i-1}\beta)$-equivalence for any fixed $i = 1, \ldots, m$. Set $L(k) = \sum_{i=1}^{m} \log_2 |\mathrm{Sym}(S, p_i; \alpha_0 + k\beta)|$. When k increases, any group $\mathrm{Sym}(S, p_i; \alpha_0 + k\beta)$ can become only smaller, not larger, so $L(k)$ is non-increasing. If $L(k-1) = L(k)$ for any $0 < k \neq k_1, \ldots, k_{m-1}$, both conditions 3.5(a,b) hold, so $\alpha_0 + k\beta$ is a stable radius. We will find an upper bound for a minimum value of such k. If condition 3.5(b) fails for all radii $\alpha = \alpha_0 + k\beta$ with $k = k_{i-1}+1, \ldots, k_i$, then at least one of the groups $\mathrm{Sym}(S, p; \alpha_k + j\beta)$ for $p \in \{p_1, \ldots, p_i\}$ is a proper subgroup of $\mathrm{Sym}(S, p; \alpha_0 + (k-1)\beta)$. The order of a proper subgroup is at most a half of the order of the group, so

$$\log_2 |\mathrm{Sym}(S, p; \alpha_0 + k\beta)| \leq \log_2 |\mathrm{Sym}(S, p; \alpha_0 + (k-1)\beta)| - 1, k = k_{i-1}+1, \ldots, k_i.$$

Hence the sum $L(k)$ decreases at least by 1 for any failure of condition 3.5(b) from $L(0) = \sum_{i=1}^{m} \log_2 |\mathrm{Sym}(S, p_i; \alpha_0)|$ to $L(+\infty) = \sum_{i=1}^{m} \log_2 |\mathrm{Sym}(S, p_i)|$, where $\mathrm{Sym}(S, p_i)$ is the symmetry group of the global cluster $C(S; p_i)$. Adding $m-1$ potential failures of condition 3.5(a) for $\alpha_0 + k_i\beta$ with $i = 1, \ldots, m-1$, the radius $\alpha_0 + k\beta$ cannot be stable for a maximum $L + m - 1$ values of k, where

$$L = [L(0) - L(+\infty)] = \left[\sum_{i=1}^{m} \left(\log_2 |\mathrm{Sym}(S, p_i; \alpha_0)| - \log_2 |\mathrm{Sym}(S, p_i)| \right) \right].$$

Then any $\alpha = \alpha_0 + k\beta$ with $k \geq L + m$ is stable, so $\alpha(S) \leq \alpha_0 + (L+m)\beta$. To get $\alpha(S) \leq (L+m+1)2R(S)$, set $\alpha_0 = 2R(S)$ and use $\beta \leq 2R(S)$ from (a). □

The upper bound in Lemma 3.6(a) holds for any unit cell of S. If a cell is non-reduced and too long, its reduced form can have smaller bounds for $\beta(S)$.

Example 3.7 (upper bounds for $\alpha(S)$ and $\beta(S)$)**.** Let $\Lambda(b) \subset \mathbb{R}^n$ be a lattice whose unit cell is a rectangular box with the longest edge $b \geq 1$.

(a) In Lemma 3.6(a), the upper bound $b \geq \beta(S)$ is tight because $\beta(\Lambda(b)) = b$.

(b) In Lemma 3.6(b), the ratio $(2R(S)+\beta(S))/\alpha(S) \geq 1$ tends to 1 as $b \to +\infty$ for any fixed n. Indeed, a cluster $C(\Lambda(b), 0; \alpha)$ is n-dimensional only for $\alpha \geq b$, so the group $\mathrm{Sym}(\Lambda(b), 0; \alpha)$ stabilizes at $\alpha = b$, hence $\alpha(S) = b + \beta(\Lambda(b)) = 2b$ is the minimum stable radius. The covering radius $R(\Lambda(b))$ is half of the longest diagonal of the rectangular cell U. If $b \to +\infty$ and all other sizes of U remain fixed, the ratio $(2R(\Lambda(b)) + \beta(\Lambda(b)))/\alpha(S)$ tends to 1 for any fixed n.

(c) Lemma 3.6(c) was motivated by [33, Theorem 1.3], which implies the upper bound $\beta(S) + 2m(n^2+1)\log_2(2 + R(S)/r(S)) > \alpha(S)$ for m-regular point sets. Let $\Lambda \subset \mathbb{R}^2$ be a lattice whose unit cell is a rhombus with sides 1. Then $m = 1$, $n = 2$, $r(\Lambda) = 0.5$, $\beta(\Lambda) = 1$, and $\alpha(\Lambda) = 2$. If Λ deforms from a square lattice to a hexagonal lattice, the covering radius $R(\Lambda)$ varies in the range $[\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{2}}]$. The past bound above gives the estimate $1+2(2^2+1)\log_2(2+\frac{2}{\sqrt{3}}) \approx 17.6 > \alpha(\Lambda) = 2$. For any lattice Λ in this family, the symmetry group $\mathrm{Sym}(\Lambda, 0) = \mathrm{Sym}(\Lambda, 0; 1)$ stabilizes at $\alpha_0 = 1$. Lemma 3.6(c) for $\alpha_0 = 1$ gives $L = \log_2(2) - \log_2(2) = 0$, so the upper bound $\alpha_0 + (L+m)\beta(S) \geq \alpha(S)$ is tight: $2 \geq \alpha(\Lambda)$. In practice, if L is large because some local clusters $C(S; p; \alpha_0)$ have too many symmetries, one can increase the radius α_0 to reduce L for a better bound of $\alpha(S)$. ■

Definition 3.8 reminds of the *isoset*, which was initially introduced in [17, Definition 9]. We also cover the case of rigid motion and prove Completeness Theorem 3.10 in the appendix in more detail than in [17, Theorem 9].

Definition 3.8 (isoset $I(S;\alpha)$ at a radius $\alpha \geq 0$)**.** Let a periodic point set $S \subset \mathbb{R}^n$ have a motif M of m points. Split all points $p \in M$ into α-equivalence classes. Each α-equivalence class of (say) k points in M can be associated with the *isometry class* $\sigma = [C(S, p; \alpha)]$ of an α-cluster centered at some $p \in M$. The *weight* of σ is $w = k/m$. The *isoset* $I(S;\alpha)$ is the unordered set of all isometry classes $(\sigma; w)$ with weights w for all points p in the motif M. If we replace isometry with rigid motion, we get the *oriented* isoset $I^o(S;\alpha)$. ■

All points p of a lattice $\Lambda \subset \mathbb{R}^n$ from one α-equivalence class for any radius $\alpha \geq 0$ because all α-clusters $C(\Lambda, p; \alpha)$ are isometrically equivalent to each other

by translations. Hence the isoset $I(\Lambda;\alpha)$ is one isometry class of weight 1 for $\alpha \geq 0$, see examples in Fig. 6. All isometry classes σ in $I(S;\alpha)$ are in a 1-1 correspondence with all α-equivalence classes in the α-partition $P(S;\alpha)$ from Definition 3.4. So $I(S;\alpha)$ without weights can be viewed as a set of points in the isotree $\mathrm{IT}(S)$ at the radius α. The size of the isoset $I(S;\alpha)$ equals the number $|P(S;\alpha)|$ of α-equivalence classes in the α-partition. Formally, $I(S;\alpha)$ depends on α because α-clusters grow in α. To distinguish any $S, Q \subset \mathbb{R}^n$ up to isometry, we will compare their isosets at a maximum stable radius of S, Q.

Example 3.9 (isosets of simple lattices)**. (a)** Any lattice $\Lambda \subset \mathbb{R}^n$ is 1-regular by Definition 3.1 and can be assumed to contain the origin 0 of $\mathbb{R}^n$. Then the isoset $I(\Lambda;\alpha)$ consists of a single isometry class of a cluster $C(\Lambda, 0;\alpha)$. So the isotree $\mathrm{IT}(\Lambda)$ is a linear path, which is horizontally drawn for the hexagonal and square lattices Λ_6, Λ_4 in Fig. 5. If both Λ_6, Λ_4 have a minimum inter-point distance 1, then the bridge length from Definition 3.3 is $\beta = 1$.

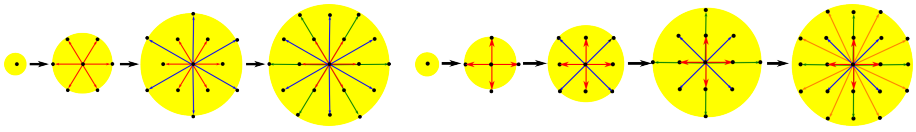


Figure 5: The isotree of any lattice Λ is $[0, +\infty)$ is a line $\mathbb{R}$ parametrized by the radius α. **Left**: the isotree of the hexagonal lattice Λ_6. **Right**: the isotree of the square lattice Λ_4.

(b) For the hexagonal lattice $\Lambda_6 \subset \mathbb{R}^2$, $C(\Lambda_6, (0,0);\alpha)$ includes points $p \neq (0,0)$ only for $\alpha \geq 1$. The cluster $C(\Lambda_6, (0,0); 1) = \{(0,0), (\pm 1, 0), (\pm\frac{1}{2}, \pm\frac{\sqrt{3}}{2})\}$ appears in the 2nd step of Fig. 5 (left). The symmetry group $\mathrm{Sym}(\Lambda_6, (0,0);\alpha)$ becomes the dihedral group D_6 (all symmetries of a regular hexagon) for $\alpha \geq 1$. Hence any $\alpha \geq \beta + 1 = 2$ is stable. The isoset $I(\Lambda_6; 1)$ is the isometry class of the cluster $C(\Lambda_6, (0,0); 1)$ of six vertices of the regular hexagon and its center.

(c) For the square lattice $\Lambda_4 \subset \mathbb{R}^2$, $C(\Lambda_4, (0,0);\alpha)$ has points $p \neq (0,0)$ only for $\alpha \geq 1$. $C(\Lambda_4, (0,0); 2) = \{(0,0), (\pm 1, 0), (0, \pm 1), (\pm\sqrt{2}, \pm\sqrt{2}), (\pm 2, 0), (0, \pm 2)\}$ includes the origin $(0,0)$ with its 12 neighbors in the 4th step of Fig. 5 (right). The group $\mathrm{Sym}(\Lambda_4, (0,0);\alpha)$ becomes the dihedral group D_4 (all symmetries of

a square) for $\alpha \geq 1$. So any $\alpha \geq \beta + 1 = 2$ is stable. The isoset $I(\Lambda_4; 1)$ is the isometry class of $C(\Lambda_4, (0,0); 1)$ of four vertices of the square and its center. ■

An equality $\sigma = \xi$ between isometry classes of clusters means that some (hence any) clusters $C(S, p; \alpha)$ and $C(Q, q; \alpha)$ representing σ, ξ, respectively, are related by $f \in \mathrm{O}(\mathbb{R}^n; p, q)$, which will be algorithmically tested in Corollary 5.4.

Theorem 3.10 (isometry classification of periodic point sets)**.** For any periodic point sets $S, Q \subset \mathbb{R}^n$, let α be a common stable radius satisfying Definition 3.5 for an upper bound $\beta \geq \beta(S), \beta(Q)$. Then S, Q are isometric (related by rigid motion, respectively) if and only if there is a bijection $\varphi : I(S; \alpha) \to I(Q; \alpha)$ (between oriented isosets, respectively) that preserves all their weights. ■

Theorem 3.10 was inspired by [33, Theorem 1.3] saying that, for a multi-regular point set X, "the only Delone sets Y all of whose ρ-stars are isometric to ρ-stars of X are sets globally isometric to X". After renaming ρ-stars as α-clusters, we collected their isometry classes (with weights) into the *isoset* to rephrase [33, Theorem 1.3] as a classification of all periodic point sets by isosets. The α-equivalence and isoset in Definition 3.8 can be refined by labels such as chemical elements, which keeps Theorem 3.10 valid for labeled points.

When comparing sets from a finite database, it suffices to build their isosets only up to a common upper bound of a stable radius α in Lemma 3.6(c).

4. Continuous metrics on isometry classes of periodic sets in $\mathbb{R}^n$

This section proves the continuity of the isoset $I(S; \alpha)$ in Theorem 4.9 by using the Earth Mover's Distance (EMD) from Definition 4.4. For a point $p \in \mathbb{R}^n$ and a radius ε, the *closed* ball $\bar{B}(p; \varepsilon) = \{q \in \mathbb{R}^n : |\vec{q} - \vec{p}| \leq \varepsilon\}$ has as its the boundary $(n-1)$-dimensional sphere $\partial\bar{B}(p; \varepsilon) \subset \mathbb{R}^n$. The *$\varepsilon$-offset* of any set $C \subset \mathbb{R}^n$ is the Minkowski sum $C + \bar{B}(0; \varepsilon) = \{\vec{p} + \vec{q} : p \in C, q \in \bar{B}(0; \varepsilon)\}$.

Then the directed *Hausdorff* distance from Definition 2.1(a) $d_{\vec{H}}(C, D)$ is the minimum radius $\varepsilon \geq 0$ such that $C \subseteq D + \bar{B}(0; \varepsilon)$. Definition 4.1 introduces the crucial new metric, which will be explicitly computed in Lemma 5.6.

Definition 4.1 (boundary tolerant metric BT on isometry classes of clusters)**.** For a radius α and periodic point sets $S, Q \subset \mathbb{R}^n$, let clusters $C(S,p;\alpha), C(Q,q;\alpha)$ represent isometry classes $\sigma \in I(S;\alpha), \xi \in I(Q;\alpha)$, respectively. The *boundary tolerant* metric $\mathrm{BT}(\sigma,\xi)$ is defined as the minimum $\varepsilon \geq 0$ such that

(4.1a) $C(Q,q;\alpha-\varepsilon) \subseteq f(C(S,p;\alpha)) + \bar{B}(0;\varepsilon)$ for some $f \in \mathrm{O}(\mathbb{R}^n;p,q)$, and

(4.1b) $C(S,p;\alpha-\varepsilon) \subseteq g(C(Q,q;\alpha)) + \bar{B}(0;\varepsilon)$ for some $g \in \mathrm{O}(\mathbb{R}^n;q,p)$. ■

In Definition 4.1, if one cluster consists of only its centre, e.g. $C(S,p;\alpha) = \{p\}$, then the boundary tolerant metric is $\mathrm{BT} = \max\{|\vec{s}-\vec{q}| \mid s \in C(Q,q;\alpha)\}$.

Lemma 4.2 (correctness of BT)**.** The metric $\mathrm{BT}(\sigma,\xi)$ in Definition 4.1 is independent of cluster representatives and satisfies the metric axioms below:

(4.2a) $\mathrm{BT}(\sigma,\xi) = 0$ if and only if $\sigma = \xi$ as isometry classes of α-clusters;

(4.2b) symmetry : $\mathrm{BT}(\sigma,\xi) = \mathrm{BT}(\xi,\sigma)$ for any isometry classes of α-clusters;

(4.2c) triangle inequality : $\mathrm{BT}(\sigma,\zeta) \leq \mathrm{BT}(\sigma,\xi) + \mathrm{BT}(\xi,\zeta)$ for any σ,ξ,ζ. ■

Example 4.3 (square lattice vs hexagonal)**.** The isoset $I(\Lambda;\alpha)$ of any lattice $\Lambda \subset \mathbb{R}^n$ containing the origin 0 consists of a single isometry class $[C(\Lambda,0;\alpha)]$, see Example 3.9. For the square (hexagonal) lattice with minimum inter-point distance 1 in Fig. 6, the cluster $C(\Lambda,0;\alpha)$ consists of only 0 for $\alpha < 1$ and includes four (six) nearest neighbors of 0 for $\alpha \geq 1$. Hence $\mathrm{Sym}(\Lambda,0;\alpha)$ stabilizes as the symmetry group of the square (regular hexagon) for $\alpha \geq 1$. The lattices have the minimum stable radius $\alpha(\Lambda) = 2$ and $\beta(\Lambda) = 1$ by Example 3.7(c). Fig. 6 illustrates the computations whose extra details are in Example A.5. ■

Non-isometric periodic sets S, Q such as perturbations in Fig. 2 can have isosets of different numbers of isometry classes. A distance between these weighted distributions of different sizes can be measured by EMD below.

Definition 4.4 (Earth Mover's Distance on isosets)**.** Let periodic point sets $S, Q \subset \mathbb{R}^n$ have a common stable radius α and isosets $I(S;\alpha) = \{(\sigma_i, w_i)\}$ and $I(Q;\alpha) = \{(\xi_j, v_j)\}$, where $i = 1,\ldots,m(S)$ and $j = 1,\ldots,m(Q)$. The

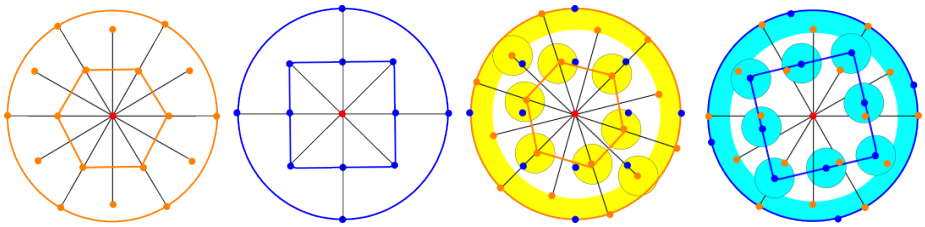


Figure 6: Example 4.3 computes the metric BT from Definition 4.1 for the isometry classes of the 2-clusters in the square and hexagonal lattices Λ_4, Λ_6. **1st**: the 2-cluster $C(\Lambda_6, 0; 2)$ with its boundary circle $\partial\bar{B}(0; 2)$; **2nd**: the 2-cluster $C(\Lambda_4, 0; 2)$ with its boundary circle $\partial\bar{B}(0; 2)$; **3rd**: for $\varepsilon = \sqrt{2} - 1 \approx 0.41$, the cluster $C(\Lambda_4, 0; 2)$ is covered by the yellow ε-offset of $C(\Lambda_6, 0; 2) \cup \partial\bar{B}(0; 2)$ rotated through 15° clockwise. **4th**: $C(\Lambda_6, 0; 2)$ is covered by the blue ε-offset of $C(\Lambda_4, 0; 2) \cup \partial\bar{B}(0; 2)$ rotated through 15° anticlockwise, so $\mathrm{BT} = \sqrt{2} - 1$.

Earth Mover's Distance [18] is $\mathrm{EMD}(I(S;\alpha), I(Q;\alpha)) = \sum_{i=1}^{m(S)} \sum_{j=1}^{m(Q)} f_{ij} \mathrm{BT}(\sigma_i, \xi_j)$ minimized over *flows* $f_{ij} \in [0, 1]$ subject to $\sum_{j=1}^{m(Q)} f_{ij} \le w_i$ for $i = 1, \dots, m(S)$, $\sum_{i=1}^{m(S)} f_{ij} \le v_j$ for $j = 1, \dots, m(Q)$, and $\sum_{i=1}^{m(S)} \sum_{j=1}^{m(Q)} f_{ij} = 1$. ■

Lemma 4.5 (EMD is a metric on isosets)**.** The Earth Mover's Distance from Definition 4.4 satisfies the metric axioms for all α and periodic sets S, Q, T.

(4.5a) $\mathrm{EMD}(I(S;\alpha), I(Q;\alpha)) = 0$ if and only if $I(S;\alpha) = I(Q;\alpha)$;

(4.5b) $\mathrm{EMD}(I(S;\alpha), I(Q;\alpha)) = \mathrm{EMD}(I(Q;\alpha), I(S;\alpha))$;

(4.5b) $\mathrm{EMD}(I(S;\alpha), I(Q;\alpha)) + \mathrm{EMD}(I(Q;\alpha), I(T;\alpha)) \ge \mathrm{EMD}(I(S;\alpha), I(T;\alpha))$.

Example 4.6 (EMD for lattices with $d_B = +\infty$)**.** [8, Example 2.1] showed that the lattices $S = \mathbb{Z}$ and $Q = (1 + \delta)\mathbb{Z}$ have the bottleneck distance $d_B(S, Q) = +\infty$ for any $\delta > 0$. We show that S, Q have Earth Mover's Distance $\mathrm{EMD} = 2\delta$ at their common stable radius $\alpha = 2 + 2\delta$. The bridge lengths are $\beta(S) = 1$ and $\beta(Q) = 1 + \delta$. The α-cluster $C(S, 0; \alpha)$ contains non-zero points for $\alpha \ge 1$, e.g. $C(S, 0; 1) = \{0, \pm 1\}$. The symmetry group $\mathrm{Sym}(S, 0; \alpha) = \mathbb{Z}_2$ includes a non-trivial reflection with respect to 0 for all $\alpha \ge 1$, so the stable radius

of S is any $\alpha \geq \beta + 1 = 2$. Similarly, Q has $\beta(Q) = 1 + \delta$ and stable radii $\alpha \geq 2(1 + \delta)$. The Earth Mover's Distance between $I(S; \alpha)$ and $I(Q; \alpha)$ at the common stable radius $\alpha = 2 + 2\delta$ equals the metric BT between the only α-clusters $C(S, 0; \alpha) = \{0, \pm 1, \pm 2\}$ and $C(Q, 0; \alpha) = \{0, \pm(1 + \delta), \pm 2(1 + \delta)\}$.

By Definition 4.1 we look for a minimum $\varepsilon > 0$ such that the cluster $C(S, 0; \alpha - \varepsilon)$ is covered by ε-offsets of $\pm(1 + \delta), \pm 2(1 + \delta)$ and vice versa. If $\varepsilon < 2\delta$, the points $\pm 2 \in C(S, 0; \alpha - \varepsilon)$ cannot be ε-close to $\pm(1 + \delta), \pm 1(+\delta)$, but $\varepsilon = 2\delta$ is large enough. The cluster $C(Q, 0; \alpha - 2\delta) = \{0, \pm(1+\delta)\}$ is covered by the 2δ-offset of $C(S, 0; \alpha) = \{0, \pm 1, \pm 2\}$, so $\mathrm{EMD}(I(S; \alpha), I(Q; \alpha)) = 2\delta$. ■

Definition 4.7 (*packing radius*)**.** For a discrete set $Q \subset \mathbb{R}^n$, the *packing radius* $r(Q)$ is the minimum half-distance between any points of Q. Also, $r(Q)$ is the maximum radius r such that the open balls $B(p; r)$ are disjoint for all $p \in Q$. ■

Lemma 4.8 is proved in the appendix and is needed for Theorem 4.9.

Lemma 4.8. Let periodic point sets $S, Q \subset \mathbb{R}^n$ have bottleneck distance $d_B(S, Q) < r(Q)$, where $r(Q)$ is the packing radius. Then S, Q have a common lattice Λ with a unit cell U such that $S = \Lambda + (U \cap S)$ and $Q = \Lambda + (U \cap Q)$. ■

For rigid motion instead of general isometry, Definition 4.1 of a boundary tolerant metric BT is updated to BT^o by considering only orientation-preserving isometries from $\mathrm{SO}(\mathbb{R}^n; p, q)$, which also makes the continuity below valid for oriented isosets $I^o(S; \alpha)$ under EMD using BT^o instead of BT in Definition 4.4.

Theorem 4.9 (continuity of isosets under perturbations)**.** Let periodic point sets $S, Q \subset \mathbb{R}^n$ have a bottleneck distance $d_B(S, Q) < r(Q)$, where $r(Q)$ is the packing radius in Definition 4.7. Then the isosets $I(S; \alpha), I(Q; \alpha)$ are close in the Earth Mover's Distance: $\mathrm{EMD}(I(S; \alpha), I(Q; \alpha)) \leq 2d_B(S, Q)$ for $\alpha \geq 0$. ■

Proof. By Lemma 4.8 the given periodic point sets S, Q have a common unit cell U. Let $g : S \to Q$ be a bijection such that $|\vec{p} - g(\vec{p})| \leq \varepsilon = d_B(S, Q) = \inf_{g: S \to Q} \sup_{p \in S} |\vec{p} - g(\vec{p})|$ for all points $p \in S$. Since the bottleneck distance $\varepsilon < r(Q)$ is small, the bijective image $g(p)$ of any point $p \in S$ is a unique ε-close point of Q

and vice versa. Hence we can assume that the common unit cell $U \subset \mathbb{R}^n$ contains the same number (say, m) points from S and Q. Expand the initial $m(S)$ isometry classes $(\sigma_i, w_i) \in I(S;\alpha)$ to m isometry classes (with equal weights $\frac{1}{m}$) represented by clusters $C(S,p;\alpha)$ for m points $p \in S \cap U$. If the i-th initial isometry class had a weight $w_i = \frac{k_i}{m}$, $i = 1,\dots,m(S)$, the expanded isoset contains k_i equal isometry classes of weight $\frac{1}{m}$. For example, the 1-regular set S_1 in Fig. 10 has the isoset consisting of a single class $[C(S_1,p;\alpha)]$, which is expanded to four identical classes of weight $\frac{1}{4}$ for the four points in the motif. The isoset $I(Q;\alpha)$ is similarly expanded to m isometry classes of weight $\frac{1}{m}$.

For any point $p \in S \cap U$, the image $g(p) \in Q$ has a unique point $h(p) \in Q \cap U$ such that $h(p)$ is equivalent to $g(p)$ modulo the lattice of Q. Then the α-clusters of $g(p)$ and $h(p)$ in Q are isometric for any $\alpha \geq 0$. The bijection $p \mapsto h(p)$ between the expanded motifs of S, Q induces the bijection between the expanded sets of m isometry classes. Each correspondence $\sigma_l \mapsto \xi_l$ in the latter bijection can be visualized as the flow $f_{ll} = \frac{1}{m}$ for $l = 1,\dots,m$, so $\sum_{l=1}^{m} f_{ll} = 1$.

To show that the Earth Mover's Distance (EMD) between any initial isoset and its expansion is 0, we collapse all identical isometry classes in the expanded isosets, but keep the arrows with the flows above. Only if both tail and head of two (or more) arrows are identical, we collapse these arrows into one arrow that gets the total weight. All equal weights $\frac{1}{m}$ correctly add up at heads and tails of final arrows to the initial weights w_i, v_j of isometry classes. So the total sum of flows is $\sum_{i=1}^{m}\sum_{j=1}^{m} f_{ij} = 1$ as required by Definition 4.4. It suffices to consider below the EMD only for the expanded isosets of exactly m classes.

We will estimate the boundary tolerant metric between isometry classes σ_l, ξ_l whose centers p and $g(p)$ are ε-close within the common unit cell U. For any fixed point $p \in S \cap U$, shift S by the vector $g(\vec{p}) - \vec{p}$. This shift makes $p \in S$ and $g(p) \in Q$ identical and keeps all pairs $q, g(q)$ for $q \in C(S,p;\alpha)$ within 2ε of each other. Using the identity map f in Definition 4.1, we get the upper bound $\mathrm{BT}([C(S,p;\alpha)],[C(Q,g(p);\alpha)]) \leq 2\varepsilon$. Then $\mathrm{EMD}(I(S;\alpha),I(Q;\alpha)) \leq$

$\sum_{l=1}^{m} f_{ll}\mathrm{BT}([C(S,p;\alpha)],[C(Q,g(p);\alpha)]) \le 2\varepsilon \sum_{l=1}^{m} f_{ll} = 2\varepsilon$ as required. □

Corollary 4.10a justifies that the EMD satisfies all metric axioms for periodic point sets that have a stable radius α. Corollary 4.10b avoids this dependence on α and scales any periodic point set S to the minimum stable radius $\alpha(S) = 1$.

Corollary 4.10. **(a)** For $\alpha > 0$, $\mathrm{EMD}(I(S;\alpha), I(Q;\alpha))$ is a metric on the space of isometry classes of all periodic point sets with a stable radius α in $\mathbb{R}^n$.

(b) For a periodic point set $S \subset \mathbb{R}^n$, let $S/r(S) \subset \mathbb{R}^n$ denote S after uniformly dividing all vectors by the packing radius $r(S)$. Then $|r(S) - r(Q)| + \mathrm{EMD}(I(S/r(S);1), I(Q/r(Q);1))$ is a metric on all periodic point sets. ■

Proof. **(a)** Lemma 4.5 proved the metric axioms for the EMD on isosets. The equality $I(S;\alpha) = I(Q;\alpha)$ is equivalent to isometry $S \simeq Q$ by Theorem 3.10.

(b) By part (a), $\mathrm{EMD}(I(S/r(S);1), I(Q/r(Q);1))$ satisfies the symmetry and triangle inequality, which are preserved by adding the Euclidean distance $d = |r(S) - r(Q)|$ between the packing radii. The equality $\mathrm{EMD} = 0$ means that $S/r(S) \simeq Q/r(Q)$ are isometric. Hence S, Q are isometric up to a uniform factor. Adding the distance $d = |r(S) - r(Q)|$ guarantees that the sum becomes zero only if $r(S) = r(Q)$, so the given sets S, Q should be truly isometric. □

The metric $\mathrm{EMD}(I(S;\alpha), I(Q;\alpha))$ is measured in the same units as atomic coordinates, say in angstroms: $1\text{Å} = 10^{-10}\text{m}$, and hence is physically meaningful. By Theorem 4.9, a small value $\delta = \mathrm{EMD}(I(S;\alpha), I(Q;\alpha))$ means that atoms of S should be perturbed by at least 0.5δ on average for a complete match with Q. Since crystals are practically compared within a finite dataset, we can take any common upper bound of $\alpha(S)$ from Lemma 3.6, also in Corollary 4.10(b).

5. Algorithms to test isometry and to approximate metrics on isosets

This section describes time complexities for computing the complete invariant isoset (Theorem 5.3), comparing isosets (Corollary 5.4), approximating the

boundary tolerant metric BT and Earth Mover's Distance on isosets (Corollary 5.10). All estimates will use the geometric complexity $\mathrm{GC}(S)$ below.

Definition 5.1 (geometric complexity GC)**.** Let a periodic point set $S \subset \mathbb{R}^n$ have an asymmetric unit of m points in a cell U of volume $\mathrm{vol}[U]$. Let L be the symmetry characteristic for $\alpha_0 = 2R(S)$ in Lemma 3.6(c), where $R(S)$ is the covering radius. The *geometric complexity* is $\mathrm{GC}(S) = \frac{(10(L+m+2)R(S)/n)^n}{2\mathrm{vol}[U]}$. ■

Let $V_n = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)}$ be the volume of the unit ball in $\mathbb{R}^n$, where the Gamma function Γ has $\Gamma(k) = (k-1)!$ and $\Gamma(\frac{k}{2}+1) = \sqrt{\pi}(k-\frac{1}{2})(k-\frac{3}{2})\cdots\frac{1}{2}$ for any integer $k \geq 1$. Set $\nu(U,\alpha,n) = \frac{(\alpha+d)^n V_n}{\mathrm{vol}[U]}$, where $d = \sup\limits_{p,q\in U} |\vec{p}-\vec{q}|$ is a longest diagonal of a unit cell U. All complexities assume the real Random-Access Machine (RAM) model and a fixed dimension n of Euclidean space $\mathbb{R}^n$.

The main input size of a periodic set is the number m of motif points because the length of a standard Crystallographic Information File is linear in m. For a fixed dimension n, the big O notation $O(m^n)$ in all complexities means a function $t(m)$ such that $t(m) \leq Cm^n$ for a fixed constant C independent of m. We will include all other parameters depending on a periodic point set S.

Lemma 5.2 (a local cluster)**.** Let a periodic point set $S \subset \mathbb{R}^n$ have m points in a unit cell U. For any stable radius $\alpha \geq 0$ and $p \in M = S \cap U$, the cluster $C(S,p;\alpha)$ has at most $k = \nu m$ points and can be found in time $\nu O(m)$, where $\nu \leq \mathrm{GC}(S)$ for geometric complexity $\mathrm{GC}(S)$ from Definition 5.1. ■

Theorem 5.3 (computing an isoset)**.** For any periodic point set $S \subset \mathbb{R}^n$ given by a motif M of m points in a unit cell U, the isoset $I(S;\alpha)$ at a stable radius α can be found in time $O(m^2 k^{\lceil n/3 \rceil} \log k)$, where $k = \nu m$ for $\nu \leq \mathrm{GC}(S)$. ■

Proof. Lemma 5.2 computes the α-clusters of m points $p \in M$ in time $O(k)$. To verify a congruence (isometry) of finite sets $A, B \subset \mathbb{R}^n$, the algorithm from [35] first moves the centers of mass of A, B to $0 \in \mathbb{R}^n$. We instead move the centers of given clusters A, B to the origin and then follow [35] to check if the shifted clusters are related by an isometry $f \in \mathrm{O}(\mathbb{R}^n;0)$ in time $O(k^{\lceil n/3 \rceil} \log k)$.

The isoset $I(S;\alpha)$ is obtained after identifying isometric clusters for m points through $O(m^2)$ pairwise comparisons. The total time is $O(m^2k^{\lceil n/3\rceil}\log k)$. □

Corollary 5.4 (comparing isosets)**.** There is an algorithm to check if any periodic point sets $S, Q \subset \mathbb{R}^n$ with motifs of at most m points are isometric in total time $O(m^2k^{\lceil n/3\rceil}\log k)$, where $k = \nu m$ for $\nu \le \max\{\mathrm{GC}(S), \mathrm{GC}(Q)\}$. ■

Proof. Theorem 5.3 finds $I(S;\alpha), I(Q;\alpha)$ with a common stable radius in time $O(m^2k^{\lceil n/3\rceil}\log k)$, where each cluster has $k = \nu m$ points by Lemma 5.2. Any classes from $I(S;\alpha), I(Q;\alpha)$ are compared [35] in time $O(k^{\lceil n/3\rceil}\log k)$. Then $O(m^2)$ comparisons suffice to check if there is a bijection $I(S;\alpha) \leftrightarrow I(Q;\alpha)$. □

Definition 5.5 (directed distances $d_{\vec{R}}$ and $d_{\vec{M}}$)**. (a)** For any sets $C, D \subset \mathbb{R}^n$, the directed *rotationally invariant* distance $d_{\vec{R}}(C,D) = \min\limits_{f\in \mathrm{O}(\mathbb{R}^n)} d_{\vec{H}}(C, f(D))$ is minimized over all maps $f \in \mathrm{O}(\mathbb{R}^n;0)$, which fix the origin $0 \in \mathbb{R}^n$.

(b) For any finite sets $C, D \subset \mathbb{R}^n$, order all points $p_1 \dots, p_k \in C$ by increasing distance to the origin 0. The *radius* of C is $R(C) = \max\limits_{p\in C} |p|$. Define the directed *max-min distance* $d_{\vec{M}}(C,D) = \max\limits_{i=1,\dots,k} \min\{\ \alpha - |p_i|,\ d_{\vec{R}}(\{p_1,\dots,p_i\}, D)\ \}$. ■

If $C' \subset C$, then $d_{\vec{R}}(C', D) \le d_{\vec{R}}(C, D)$. Let $C, D \subset \bar{B}(0;\alpha)$ be finite sets including the origin 0. If $C = \{0\}$, then $d_{\vec{R}}(C,D) = 0$ because $C \subset D$, but $d_{\vec{R}}(D,C) = R(D)$ is the radius of D because $D \subset \{0\} + \bar{B}(0;\varepsilon)$ only for $\varepsilon \ge R(D)$. Definition 5.5, Lemma 5.6 and hence all further results work for rigid motion by restricting all maps to the special orthogonal group $\mathrm{SO}(\mathbb{R}^n;0)$.

Lemma 5.6 (max-min formula for $d_{\vec{R}}$ via $d_{\vec{M}}$)**.** For any finite sets $C, D \subset \mathbb{R}^n$, $\alpha \ge R(C)$, the distance $d_{\vec{R}}(C \cup \partial\bar{B}(0;\alpha), D \cup \partial\bar{B}(0;\alpha))$ equals $d_{\vec{M}}(C,D)$. ■

Example 5.7 (max-min formula)**.** Consider the subcluster $C \subset C(\Lambda_4, 0; 2)$ of the points $p_1 = (1,0)$, $p_2 = (1,1)$, $p_3 = (1,-1)$, $p_4 = (2,0)$ from the square lattice Λ_4 in Fig. 6. Let $\alpha = 2$ and $D = C(\Lambda_6, 0; 2)$ be the 2-cluster of the hexagonal lattice Λ_6. Then $d_{\vec{R}}(p_1, D) = 0$ because p_1 coincides with $(1,0) \in D$. Then $d_{\vec{R}}(\{p_1,p_2\}, D) = \sqrt{2} - 1$, because the cloud D after the clockwise rotation through 15° has the points $(\cos 15°, -\sin 15°)$ and $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$

at distances $\sqrt{(\cos 15^\circ - 1)^2 + \sin^2 15^\circ} \approx 0.26$, $\sqrt{2} - 1 \approx 0.41$ to p_1, p_2, respectively. Then $d_{\vec{R}}(\{p_1, p_2, p_3\}, D) = \sqrt{2} - 1$ because the same rotated image of D has $(\sqrt{\frac{3}{2}}, -\sqrt{\frac{3}{2}})$ at the distance $\sqrt{3} - \sqrt{2} \approx 0.32$ to p_3. For $i = 1$, $\min\{\alpha - |p_1|, d_{\vec{R}}(p_1, D)\} = \min\{2 - 1, 0\} = 0$. For $i = 2, 3$, $\min\{\alpha - |p_2|, d_{\vec{R}}(\{p_1, p_2\}, D)\} = \min\{\alpha - |p_3|, d_{\vec{R}}(\{p_1, p_2, p_3\}, D)\} = \min\{2 - \sqrt{2}, \sqrt{2} - 1\} = \sqrt{2} - 1$. For $i = 4$, $\min\{\alpha - |p_4|, d_{\vec{R}}(C, D)\} = 0$ since $\alpha = 2 = |p_4|$. The maximum value is $\sqrt{2} - 1$, so Example 4.3 fits Lemma 5.6. ■

Lemma 5.8 extends [36, section 2.3] from $n = 3$ to any dimension $n > 1$.

Lemma 5.8 (approximating $d_{\vec{R}}$)**.** Let a cloud $C \subset \mathbb{R}^n$ consist of $k = |C|$ points ordered by distances $|p_1| \le \ldots \le |p_k|$ from the origin and $\langle C \rangle$ denote the number of different vectors $\vec{p}/|\vec{p}|$ for $p \in C$. For each $j = 1, \ldots, k$, consider the subcloud $C_j = \{p_1, \ldots, p_j\}$. For any cloud $D \subset \mathbb{R}^n$ of $|D|$ points, all distances $d_j = d_{\vec{R}}(C_j, D)$ from Definition 5.5 for $j = 1, \ldots, k$ can be approximated by some d'_j in time $O(|C|\langle C \rangle^{n-1}|D|)$ so that $d_j \le d'_j \le \omega d_j$, $\omega = 1 + \frac{1}{2}n(n-1)$. ■

The proof of Lemma 5.8 uses only orientation-preserving isometries from $\mathrm{SO}(\mathbb{R}^n, 0)$. Hence the upper bounds from Lemma 5.8, Theorem 5.9, and Corollary 5.10 work for both cases of rigid motion and general isometry in $\mathbb{R}^n$.

Theorem 5.9 (approximating BT)**.** Let periodic point sets $S, Q \subset \mathbb{R}^n$ have isometry classes σ, ξ represented by clusters C, D of a radius α, respectively. In the notations of Lemma 5.8, $\mathrm{BT}(\sigma, \xi)$ from Definition 4.1 can be approximated with the factor $\omega = 1 + \frac{1}{2}n(n-1)$ in time $O(|C|(\langle C \rangle^{n-1} + \langle D \rangle^{n-1})|D|)$. ■

Proof. By Definitions 4.1 and 5.5, the boundary tolerant metric $\mathrm{BT}(\sigma, \xi)$ is the maximum of $d_{\vec{R}}(C \cup \partial\bar{B}(0; \alpha), D \cup \partial\bar{B}(0; \alpha))$ and $d_{\vec{R}}(D \cup \partial\bar{B}(0; \alpha), C \cup \partial\bar{B}(0; \alpha))$. Lemma 5.6 implies that $\mathrm{BT}(\sigma, \xi) = \max\{d_{\vec{M}}(C, D), d_{\vec{M}}(D, C)\}$. It remains to compute required approximations of the two distances $d_{\vec{M}}$ above.

Let C consist of k points ordered by distances $|p_1| \le \ldots \le |p_k|$ from the origin. For each $j = 1, \ldots, k$, consider the subcloud $C_j = \{p_1, \ldots, p_j\}$. We use the approximation d'_j from Lemma 5.8 to compute $d' = \max\limits_{i=1,\ldots,k} \min\{\alpha - |p_i|,\ d'_i\}$

in the extra time $O(|C|)$. Now we check that this final approximation d' is between $d_{\vec{M}}(C,D) = \max\limits_{i=1,\dots,k} \min\{\alpha - |p_i|,\ d_i\}$ in Lemma 5.6 and $\omega d_{\vec{M}}(C,D)$.

The inequalities $d_j \le d'_j \le \omega d_j$ for $j = 1, \dots, k$ from Lemma 5.8 imply that $\min\{\alpha - |p_j|, d_j\} \le \min\{\alpha - |p_j|, d'_j\} \le \min\{\alpha - |p_j|, \omega d_j\}$. By fixing an index j maximizing the left-hand side minimum above, we conclude that $d_{\vec{M}}(C,D) = \min\{\alpha - |p_j|,\ d_j\} \le \max\limits_{i=1,\dots,k} \min\{\alpha - |p_i|,\ d'_i\} = d'$. By fixing an index j maximizing the middle side minimum above, we get the following upper bound: $d' = \min\{\alpha - |p_j|,\ d'_j\} \le \max\limits_{i=1,\dots,k} \min\{\alpha - |p_i|,\ \omega d_i\} \le \omega d_{\vec{M}}(C,D)$. The extra time $O(|C| + |D|)$ for the approximations of $d_{\vec{M}}(C,D)$ and $d_{\vec{M}}(D,C)$ is dominated by the total time $O(|C|(\langle C\rangle^{n-1} + \langle D\rangle^{n-1})|D|)$ from Lemma 5.8. □

Corollary 5.10 (approximating EMD)**.** Let $S, Q \subset \mathbb{R}^n$ be periodic point sets whose motifs have at most m points p and χ different vectors $\vec{p}/|\vec{p}|$. For any $\alpha > 0$, the metric $\mathrm{EMD}(I(S;\alpha), I(Q;\alpha))$ can be approximated with the factor $\omega = 1 + \frac{1}{2}n(n-1)$ in time $O(\nu^2 m^4 \chi^{n-1})$, where $\nu \le \max\{\mathrm{GC}(S), \mathrm{GC}(Q)\}$. ■

Proof. Since S, Q have at most m points in their motifs, their isosets at any radius α have at most m isometry classes. By Theorem 5.9 the distance $\mathrm{BT}(\sigma, \xi)$ between any classes $\sigma \in I(S;\alpha)$ and $\xi \in I(Q;\alpha)$ of α-clusters up to k points can be approximated with the factor ω in time $O(k^2 \chi^{n-1})$. The maximum number of points is $k = \nu m$, where $\nu \le \max\{\mathrm{GC}(S), \mathrm{GC}(Q)\}$ by Lemma 5.2. Since Definition 4.4 uses normalized distributions, ω emerges as a multiplicative upper bound in $\mathrm{EMD}(I(S;\alpha), I(Q;\alpha))$. After computing $O(m^2)$ pairwise distances between sets of m clusters, the exact EMD can be found in the extra time $O(m^3 \log m)$ [37], which is dominated by the time $O(m^2 \nu^2 m^2 \chi^{n-1})$ for all cluster distances. So the total time becomes $O(\nu^2 m^4 \chi^{n-1})$. The EMD can be approximated [38, section 3] with a constant factor in time $O(m)$. □

Counting directions $\vec{p}/|p|$ as points ($\chi \le m$), for dimension $n = 3$, the rough bounds for the isoset and its approximate EMD$'$ in Theorem 5.3 and Corollary 5.10 are $O(m^3 \log m)$ and $O(m^6)$, respectively. Algorithms 1-2 in the appendix describe pseudocodes for Lemma 5.8, Theorem 5.9, and Corollary 5.10.

6. A lower bound for continuous metrics via simpler invariants

Theorem 6.5 gives a lower bound for EMD in terms of the simpler invariant *Pointwise Distance Distribution* [8], see Definition 6.1 below. If S is a lattice or a 1-regular set, then all points are isometrically equivalent, so they have the same distances to all their neighbors. In this case, $\mathrm{PDD}(S;k)$ is a single row of k distances, which is the vector $\mathrm{AMD}(S;k)$ of Average Minimum Distances [7].

Definition 6.1 (Pointwise Distance Distribution PDD)**.** Let a periodic set $S = \Lambda + M$ have points $p_1, \dots, p_m$ in a unit cell. For $k \geq 1$, consider the $m \times k$ matrix $D(S;k)$, whose i-th row consists of the ordered Euclidean distances $d_{i1} \leq \dots \leq d_{ik}$ from p_i to its first k nearest neighbors in the full set S. The rows of $D(S;k)$ are *lexicographically* ordered as follows. A row $(d_{i1}, \dots, d_{ik})$ is *smaller* than $(d_{j1}, \dots, d_{jk})$ if the first (possibly none) distances coincide: $d_{i1} = d_{j1}, \dots, d_{il} = d_{jl}$ for $l \in \{1, \dots, k-1\}$ and the next $(l+1)$-st distances satisfy $d_{i,l+1} < d_{j,l+1}$. If w rows are identical to each other, these rows are collapsed to one row with the *weight* w/m. Put this weight in the extra first column. The final matrix of $k+1$ columns is the *Pointwise Distance Distribution* $\mathrm{PDD}(S;k)$. The Average Minimum Distance $\mathrm{AMD}(S;k)$ is the vector $(\mathrm{AMD}_1, \dots, \mathrm{AMD}_k)$, where AMD_i is the weighted average of the $(i+1)$-st column of $\mathrm{PDD}(S;k)$. ■

Theorem 6.2 (isometry invariance of PDD)**.** For any finite or periodic set $S \subset \mathbb{R}^n$, $\mathrm{PDD}(S;k)$ in Definition 6.1 is an isometry invariant of S for $k \geq 1$. ■

Theorem 6.2 and continuity of PDD in the metric from Definition 6.3 follows from more general results in [8]. The distance between rows $\vec{R}_i(S), \vec{R}_j(Q)$ of PDD matrices below is measured in the metric $L_\infty(\vec{p}, \vec{q}) = \max\limits_{i=1,\dots,k} |p_i - q_i|$.

Definition 6.3 (Earth Mover's Distance on Pointwise Distance Distributions)**.** Let finite or periodic sets $S, Q \subset \mathbb{R}^n$ have $\mathrm{PDD}(S;k)$, $\mathrm{PDD}(Q;k)$ with rows $\vec{R}_i(S), \vec{R}_j(Q)$ of weights $w_i(S), w_(Q)$ for $i = 1, \dots, m(S)$ and $j = 1, \dots, m(Q)$, respectively. A full *flow* from $\mathrm{PDD}(S;k)$ to $\mathrm{PDD}(Q;k)$ is an $m(S) \times m(Q)$ matrix whose element $f_{ij} \in [0,1]$ is called a partial *flow* from $\vec{R}_i(S)$ to $\vec{R}_j(Q)$. The *Earth Mover's Distance* is the minimum value of the *cost* $\mathrm{EMD}(I(S), I(Q)) =$

$\sum_{i=1}^{m(S)}\sum_{j=1}^{m(Q)} f_{ij}L_\infty(\vec{R}_i(S),\vec{R}_j(Q))$ over *flows* $f_{ij}\in[0,1]$ subject to $\sum_{j=1}^{m(Q)} f_{ij}\le w_i(S)$ for $i=1,\dots,m(S)$, $\sum_{i=1}^{m(S)} f_{ij}\le w_j(Q)$ for $j=1,\dots,m(Q)$, $\sum_{i=1}^{m(S)}\sum_{j=1}^{m(Q)} f_{ij}=1$. ■

Lemma 6.4 is a partial case of Theorem 6.5 for 1-regular point sets S,Q.

Lemma 6.4 (lower bound for the tolerant distance BT)**.** Let $S,Q\subset\mathbb{R}^n$ be periodic point sets with a common stable radius α. Choose any points $p\in S$ and $q\in Q$. Let the distance between isometry classes of α-clusters $\varepsilon=\mathrm{BT}([C(S,p;\alpha)],[C(Q,q;\alpha)])$ be smaller than a minimum half-distance between any point of S and Q. Let k be a minimum number of points in the clusters $C(S,p;\alpha-\varepsilon)$ and $C(Q,q;\alpha-\varepsilon)$. Then the L_∞ distance between the rows of the points p,q in $\mathrm{PDD}(S;k),\mathrm{PDD}(Q;k)$, respectively, is at most ε. ■

Theorem 6.5 (lower bound for EMD)**.** Let $S,Q\subset\mathbb{R}^n$ be periodic sets with a common stable radius α. Let $\varepsilon=\mathrm{EMD}(I(S;\alpha),I(Q;\alpha))$ and k be the maximum number of points of S,Q in their $(\alpha-\varepsilon)$-clusters. If ε is less than the half-distance between any points of S,Q, then $\mathrm{EMD}(\mathrm{PDD}(S;k),\mathrm{PDD}(Q;k))\le\varepsilon$. ■

Proof. To prove that $\mathrm{EMD}(\mathrm{PDD}(S;k),\mathrm{PDD}(Q;k))\le\mathrm{EMD}(I(S;\alpha),I(Q;\alpha))$, we choose optimal flows $f_{ij}\in[0,1]$, $i=1,\dots,m(S)$ and $j=1,\dots,m(Q)$, that minimize $\varepsilon=\mathrm{EMD}(I(S;\alpha),I(Q;\alpha))$ in Definition 4.4. For any points $p_i\in S$ and $q_j\in Q$, let $\vec{R}_i(S)$ and $\vec{R}_j(Q)$ be their rows in $\mathrm{PDD}(S;k)$ and $\mathrm{PDD}(Q;k)$, respectively. Lemma 6.4 gives $L_\infty(\vec{R}_i(S),\vec{R}_j(Q))\le\mathrm{BT}([C(S,p_i;\alpha)],[C(Q,q_j;\alpha)])$. These inequalities for all indices i,j and the same flows f_{ij} imply that

$$\sum_{i=1}^{m(S)}\sum_{j=1}^{m(Q)} f_{ij}L_\infty(\vec{R}_i(S),\vec{R}_j(Q))\le\sum_{i=1}^{m(S)}\sum_{j=1}^{m(Q)} f_{ij}\mathrm{BT}(\sigma_i,\xi_j)=\varepsilon$$

by the choice of f_{ij} The left-hand side of the last inequality can become only smaller when minimizing over f_{ij}. Then $\mathrm{EMD}(\mathrm{PDD}(S;k),\mathrm{PDD}(Q;k))\le\varepsilon$. □

7. Real experiments, limitations, significance, and a discussion

In 1930, future Nobel laureate Linus Pauling noticed the ambiguity of crystal structures obtained by diffraction [39]. Such *homometric* crystals with identical

diffraction patterns were only manually distinguished until now because even the generically complete PDDs coincide for the Pauling periodic sets $P(\pm u)$ for all $u \in (0, 0.25)$, see the real overlaid crystals for $u = 0.03$ in Fig. 7 (left).

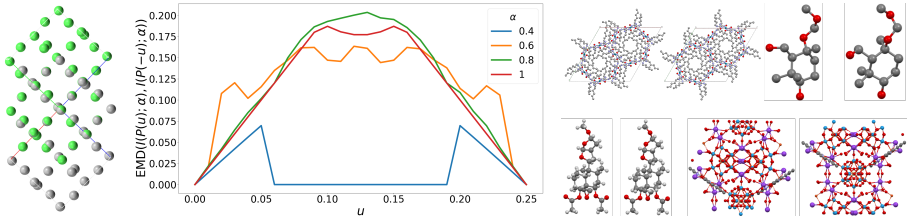


Figure 7: **Left**: a comparison of Pauling's homometric crystals $P(\pm u)$ for $u = 0.03$ [39], by COMPACK [19] aligning subsets of 48 atoms and outputs RMSD, which fails the triangle inequality. The atoms from different $P(\pm 0.03)$ are shown in green and gray. **Middle**: the pairs of $P(\pm u)$ have EMD$' > 0$ for all $u \in (0, 0.25)$ and $\alpha > 0.4$ (running time 50 ms for $u = 0.03$ and $\alpha = 0.5$). **Right**: four pairs of mirror images in the CSD are indistinguishable by all past invariants but have approximate EMD$' > 0$ for all radii $\alpha > 1.5$Å in Fig. 8 (left).

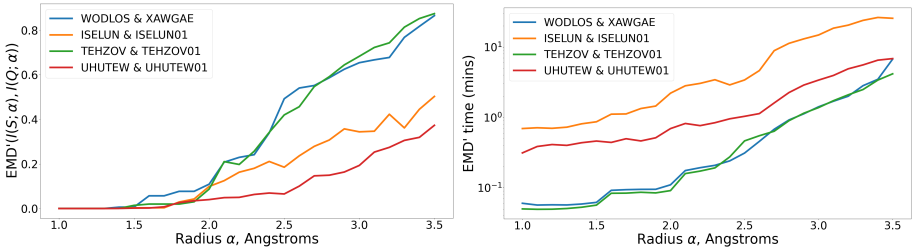


Figure 8: The isosets distinguish all four pairs of mirror images given by their codes in the CSD. **Left**: approximate EMD$'$ for different radii α. **Right**: running times on a desktop.

The strongest past invariant PDD is based on distances and cannot distinguish mirror images. In the CSD, we found four pairs that have identical PDDs but are mirror images shown in Fig. 7 (right), distinguished by isosets with $\alpha \geq 1.5$Å in Fig. 8 (left). For WODLOS vs XAWGAE and $\alpha = 2$, the total time including isosets and EMD is about 4.3 seconds. All experiments were on a modest CPU AMD Ryzen 5 5600X, 32GB RAM. Fig. 9 summarizes times of new invariants and metrics, see Tables A.1 and A.2 for details. The supplementary materials include a Python code with instructions and full tables of metrics and run times for thousands of near-duplicates in the CSD and GNoME.

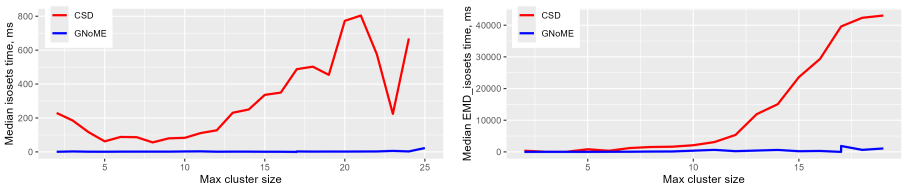


Figure 9: Median run times vs the max size of α-clusters. **Left**: invariants. **Right**: metrics. Times are faster for symmetric (inorganic) crystals in the GNoME than in the CSD.

The limitations of the EMD metric on isosets in Definition 4.4 are a slower running time than for AMD, PDD and the approximate (not exact) algorithm in Corollary 5.10, which are outweighed by the following crucial advantages.

First, all past invariants could not distinguish infinitely many periodic sets (including all mirror images) under rigid motion, e.g. the real crystals in Fig. 7. The new continuous EMD fully solved Problem 1.2, which remained open at least since 1965 [27]. Second, because the proved error factor in the practical dimension $n = 3$ is close to 4, any near-duplicate crystals that differ by atomic deviations of up to ε have an exact distance EMD $\leq 2\varepsilon$ by main Theorem 4.9 and hence an approximate distance up to about 8ε by Corollary 5.10.

Any crystals that can be matched under rigid motion are recognizable since our approximation of EMD $= 0$ is also 0. Any approximate value δ of EMD for real crystals S, Q implies that all atoms of S should be perturbed by at least $\delta/8$ on average for a complete match with Q. This result justifies filtering out near-duplicates [10] in all datasets to avoid machine learning on skewed data.

Future work can use the EMD to continuously quantify changes in material properties under perturbations of atoms and extend Problem 1.2 to metrics on finite or periodic sets of points under affine and projective transformations.

Conclusions. Sections 3 and 4 prepared the complexity results in section 5: algorithms for computing and comparing isosets (Theorems 5.3, Corollary 5.4), and approximating the new boundary tolerant metric BT (Theorem 5.9), and EMD on isosets (Corollary 5.10). The proofs expressed polynomial bounds in terms of the motif *size* $m = |S|$ of a periodic set S because the input size of a

Crystallographic Information File is linear in m, e.g. any lattice has $m = 1$.

The factors depending on the dimension and geometric complexity $\mathrm{GC}(S)$ are inevitable due to the curse of dimensionality and the infinite nature of crystals. In practice, crystal symmetries reduce a motif to a smaller asymmetric part, which usually has less than 20 atoms even for large molecules in the CSD. The lower bound via faster PDD invariants in Theorem 6.5 justifies applying the algorithm of Corollary 5.10 only for a final confirmation of near-duplicates. So the isosets finalized the hierarchy of the faster but incomplete invariants.

The main novelty is the boundary tolerant metric in Definition 4.1 that makes the complete invariant isoset Lipschitz continuous (Theorem 4.9) without extra parameters that are needed to smooth past descriptors such as powder diffraction patterns and atomic environments with fixed cut-off radii. Because the isoset is the only Lipschitz continuous invariant whose completeness under isometry was proved for all periodic point sets in $\mathbb{R}^n$, the isoset was used to confirm the duplicates in the CSD and GNoME, see Tables A.1 and A.2.

The resulting *Crystal Isometry Principle* (CRISP) says that all non-isometric periodic crystals should have non-isometric sets of atomic centers and implies that all known and not yet discovered periodic crystals live in a common *Crystal Isometry Space* (CRIS) whose first continuous maps appeared in [16, 40].

We thank all reviewers for their valuable time and helpful comments.

References

[1] R. Feynman, The Feynman lectures on physics, Vol. 1, 1971.

[2] M. J. Winter, Chemical bonding, Oxford University Press, 2016.

[3] O. Anosova, V. Kurlin, M. Senechal, The importance of definitions in crystallography, IUCrJ 11 (2024) 453–463. doi:10.1107/S2052252524004056.

[4] P. Sacchi, et al., Same or different – that is the question: identification of crystal forms from crystal data, CrystEngComm 22 (2020) 7170–7185.

[5] D. Widdowson, V. Kurlin, Continuous invariant-based maps of the CSD, Crystal Growth and Design 24 (2024) 5627–5636.

[6] D. Bimler, Better living through coordination chemistry: A descriptive study of a prolific papermill that combines crystallography and medicine (2022). doi:10.21203/rs.3.rs-1537438/v1.

[7] D. Widdowson, et al., Average minimum distances of periodic sets - foundational invariants for mapping periodic crystals, MATCH 87 (2022) 529–559.

[8] D. Widdowson, V. Kurlin, Resolving the data ambiguity for periodic crystals, Advances in Neural Information Processing Systems 35 (2022).

[9] A. K. Cheetham, R. Seshadri, Artificial intelligence driving materials discovery? Perspective on the article: Scaling deep learning for materials discovery, Chemistry of Materials 36 (8) (2024) 3490–3495.

[10] P. Zwart, et al., Surprises and pitfalls arising from (pseudo) symmetry, Acta Cryst. D 64 (2008) 99–107.

[11] V. Kurlin, Mathematics of 2-dimensional lattices, Foundations of Computational Mathematics 24 (2024) 805–863.

[12] P. Smith, V. Kurlin, Generic families of finite metric spaces with identical or trivial 1D persistence, J Applied Comp. Topology 8 (2024) 839–855.

[13] V. Kurlin, Polynomial-time algorithms for continuous metrics on atomic clouds of unordered points, MATCH 91 (2024) 79–108.

[14] D. Widdowson, V. Kurlin, Recognizing rigid patterns of unlabeled point clouds by complete and continuous isometry invariants with no false negatives and no false positives, in: CVPR, 2023, pp. 1275–1284.

[15] V. Kurlin, Complete and continuous invariants of 1-periodic sequences in polynomial time, arxiv:2205.04388 (under a minor revision in SIMODS).

[16] M. Bright, A. Cooper, V. Kurlin, Geographic-style maps for 2-dimensional lattices, Acta Crystallographica Section A 79 (2023) 1–13.

[17] O. Anosova, V. Kurlin, An isometry classification of periodic point sets, in: LNCS (DGMM proceedings), Vol. 12708, 2021, pp. 229–241.

[18] Y. Rubner, C. Tomasi, L. Guibas, The Earth Mover's Distance as a metric for image retrieval, International J Computer Vision 40 (2000) 99–121.

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

[20] M. Duneau, C. Oguey, Bounded interpolations between lattices, Journal of Physics A: Mathematical and General 24 (2) (1991) 461.

[21] H.-G. Carstens, et al., Geometrical bijections in discrete lattices, Combinatorics, Probability and Computing 8 (1-2) (1999) 109–129.

[22] M. Laczkovich, Uniformly spread discrete sets in R^d, Journal of the London Mathematical Society 2 (1) (1992) 39–57.

[23] M. Senechal, Quasicrystals and geometry, CUP Archive, 1996.

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

[25] S. Kawano, J. Mason, Classification of atomic environments via the Gromov–Wasserstein distance, Comp. Materials Science 188 (2021) 110144.

[26] S. Rass, S. König, S. Ahmad, M. Goman, Metricizing the euclidean space towards desired distance relations in point clouds, IEEE Transactions on Information Forensics and Security 19 (2024) 7304–7319.

[27] S. Lawton, R. Jacobson, The reduced cell and its crystallographic applications, Tech. rep., Ames Lab, Iowa State University (1965).

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

[29] P. Smith, V. Kurlin, A practical algorithm for degree-k voronoi domains of three-dimensional periodic point sets, in: Lecture Notes in Computer Science (Proceedings of ISVC), Vol. 13599, 2022, pp. 377–391.

[30] O. Anosova, V. Kurlin, Density functions of periodic sequences of continuous events, Journal of Mathematical Imaging and Vision 65 (2023) 689–701.

[31] O. Anosova, V. Kurlin, Density functions of periodic sequences, in: LNCS (DGMM proceedings), Vol. 13493, 2022, pp. 395–408.

[32] B. Delone, N. Dolbilin, M. Shtogrin, R. Galiulin, A local criterion for regularity of a system of points, in: DAN SSSR, Vol. 227, 1976, pp. 19–21.

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

[34] N. Dolbilin, A. Magazinov, Uniqueness theorem for locally antipodal Delaunay sets, Proceedings of the Steklov Institute of Maths 294 (2016) 215–221.

[35] P. Brass, C. Knauer, Testing the congruence of d-dimensional point sets, in: Proceedings of SoCG, 2000, pp. 310–314.

[36] M. T. Goodrich, J. S. Mitchell, M. W. Orletsky, Approximate geometric pattern matching under rigid motions, T-PAMI 21 (1999) 371–379.

[37] J. B. Orlin, A faster strongly polynomial minimum cost flow algorithm, Operations research 41 (2) (1993) 338–350.

[38] S. Shirdhonkar, D. Jacobs, Approximate Earth Mover's Distance in linear time, in: Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[39] L. Pauling, M. Shappell, The crystal structure of bixbyite and the c-modification of the sesquioxides, Zeitschrift für Kristallographie-Cryst. Materials 75 (1930) 128–142.

[40] M. Bright, A. Cooper, V. Kurlin, Continuous chiral distances for 2-dimensional lattices, Chirality 35 (2023) 920–936.

A. Appendix A: detailed proofs of all auxiliary results

This appendix includes more detailed and updated proofs of [17, Lemmas 7 and 11-13]. Fig. 12 visualizes the logical connections between main results.

Fig. 10 (left) shows the 1-regular periodic set $S_1 \subset \mathbb{R}^2$ whose all points (close to vertices of square cells) have isometric global clusters related by translations and rotations through $90°, 180°, 270°$. The set S_2 has extra points at the centers of all square cells. The local α-clusters around these centers are not isometric to α-clusters around the points close to cell vertices for any $\alpha \geq 3\sqrt{2}$.

The 1-regular periodic point set S_1 in Fig. 10 for any $p \in S_1$ has the symmetry group $\mathrm{Sym}(S_1, p; \alpha) = \mathrm{O}(\mathbb{R}^2)$ for $\alpha \in [0, 4)$. Then $\mathrm{Sym}(S_1, p; \alpha)$ stabilizes as $\mathbb{Z}^2$ with one reflection for $\alpha \geq 4$ as soon as $C(S_1, p; \alpha)$ includes one more point.

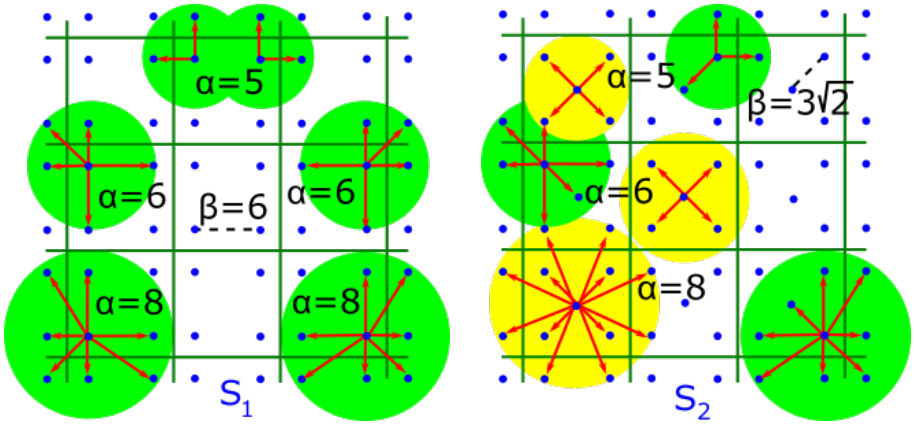


Figure 10: **Left**: in $\mathbb{R}^2$, the periodic point set S_1 has the square unit cell $[0, 10)^2$ containing the four points $(2, 2), (2, 8), (8, 2), (8, 8)$, so S_1 isn't a lattice, but is 1-regular by Definition 3.1, and $\beta(S_1) = 6$. All local α-clusters of S_1 are isometric, shown by red arrows for $\alpha = 5, 6, 8$, see Definition 3.2. **Right**: S_2 has the extra point $(5, 5)$ in the center of the cell $[0, 10)^2$ and is 2-regular with $\beta(S_2) = 3\sqrt{2}$, so S_2 has green and yellow isometry types of α-clusters.

Fig. 11 illustrates the isosets for the periodic sets S_1, S_2 in Fig. 10.

Lemma A.1 (isotree properties)**.** The isotree $\mathrm{IT}(S)$ has the properties below:

(A.1a) for $\alpha = 0$, the α-partition $P(S; 0)$ consists of one class;

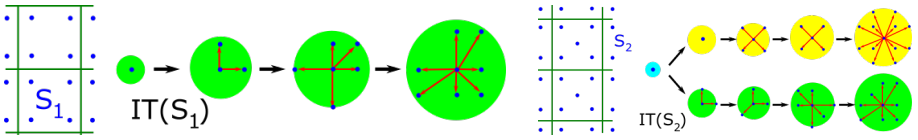


Figure 11: **Left**:The isotree $\mathrm{IT}(S_1)$ from Definition 3.4 of the 1-regular set S_1 in Fig. 10 for any $\alpha \geq 0$ has one isometry class of α-clusters up to rotation. **Right**: the isotree $\mathrm{IT}(S_2)$ of the 2-regular set S_2 in Fig. 10 stabilizes with two non-isometric classes of α-clusters for $\alpha \geq 4$.

(A.1b) if $\alpha < \alpha'$, then $\mathrm{Sym}(S,p;\alpha') \subseteq \mathrm{Sym}(S,p;\alpha)$ for any point $p \in S$;

(A.1c) if $\alpha < \alpha'$, the α'-partition $P(S;\alpha')$ *refines* $P(S;\alpha)$, i.e. any α'-equivalence class from $P(S;\alpha')$ is included into an α-equivalence class from the partition $P(S;\alpha)$. So the cluster count $|P(S;\alpha)|$ is non-strictly increasing in α. ■

Proof. (A.1a) Let $\alpha \geq 0$ be smaller than the minimum distance $2r(S)$ betweens any points of S. Then any cluster $C(S,p;\alpha)$ is the single-point set $\{p\}$. All these 1-point clusters are isometric to each other. So $|P(S;\alpha)| = 1$ for $\alpha < 2r(S)$.

(A.1b) For any $p \in S$, the inclusion of clusters $C(S,p;\alpha) \subseteq C(S,p;\alpha')$ implies that any isometry $f \in \mathrm{O}(\mathbb{R}^n;p)$ that isometrically maps the larger cluster $C(S,p;\alpha')$ to itself also maps the smaller cluster $C(S,p;\alpha)$ to itself. Hence any element of $\mathrm{Sym}(S,p;\alpha') \subseteq \mathrm{O}(\mathbb{R}^n;p)$ belongs to $\mathrm{Sym}(S,p;\alpha)$.

(A.1c) If points $p,q \in S$ are α'-equivalent at the larger radius α', i.e. the clusters $C(S,p;\alpha')$ and $C(S,q;\alpha')$ are related by an isometry from $\mathrm{O}(\mathbb{R}^n;p,q)$, then p,q are α-equivalent at the smaller radius α. Hence any α'-equivalence class of points in S is a subset of an α-equivalence class in S. □

The proofs of Lemmas A.2, A.3, and Theorem 3.10 follow [32], [33, section 4], and extend to the oriented case by taking orientation-preserving isometries. Recall that $\mathrm{O}(\mathbb{R}^n;p,q)$ denotes the set of all isometries of $\mathbb{R}^n$ that map p to q.

Lemma A.2 (local extension)**.** Let $S,Q \subset \mathbb{R}^n$ be periodic point sets and $\mathrm{Sym}(S,p;\alpha-\beta) = \mathrm{Sym}(S,p;\alpha)$ for some point $p \in S$ and $\alpha > \beta$. Assume that there is an isometry $g \in \mathrm{O}(\mathbb{R}^n;p,q)$ such that $g(C(S,p;\alpha)) = C(Q,q;\alpha)$.

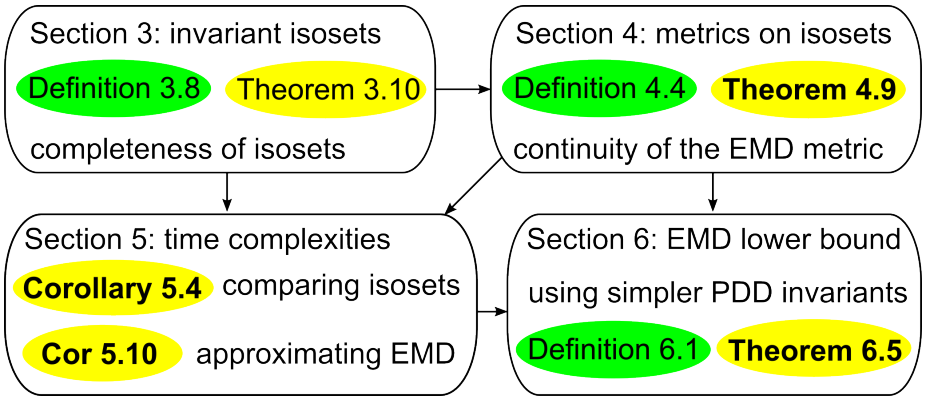


Figure 12: Key definitions and main Theorems 4.9, 6.5 about the continuous metrics on complete invariant isosets with time complexities of algorithms in Corollaries 5.4, 5.10.

Let $f \in \mathrm{O}(\mathbb{R}^n; p, q)$ be any isometry such that $f(C(S, p; \alpha - \beta)) = C(Q, q; \alpha - \beta)$. Then f isometrically maps the larger clusters: $f(C(S, p; \alpha)) = C(Q, q; \alpha)$. ■

Proof. The composition $h = f^{-1} \circ g$ fixes p and isometrically maps $C(S, p; \alpha - \beta)$ to itself, so $h \in \mathrm{Sym}(S, p; \alpha - \beta)$. The condition $\mathrm{Sym}(S, p; \alpha - \beta) = \mathrm{Sym}(S, p; \alpha)$ implies that $h \in \mathrm{Sym}(S, p; \alpha)$, so the isometry $h \in \mathrm{O}(\mathbb{R}^n; p)$ isometrically maps the larger cluster $C(S, p; \alpha)$ to itself. Then the given isometry $f = g \circ h^{-1}$ isometrically maps $C(S, p; \alpha)$ to $f(C(S, p; \alpha)) = g(C(S, p; \alpha)) = C(Q, q; \alpha)$. □

Lemma A.3 (global extension)**.** Let periodic point sets $S, Q \subset \mathbb{R}^n$ have a common stable radius α satisfying Definition 3.5 for an upper bound $\beta \geq \beta(S), \beta(Q)$. Let $I(S; \alpha) = I(Q; \alpha)$ and $p \in S$, $q \in Q$ be any points with an isometry $f \in \mathrm{O}(\mathbb{R}^n; p, q)$ such that $f(C(S, p; \alpha)) = C(Q, q; \alpha)$. Then $f(S) = Q$. ■

Proof. To show that $f(S) \subset Q$, it suffices to check that the image $f(a)$ of any point $a \in S$ belongs to Q. By Definition 3.3 the points $p, a \in S$ are connected by a sequence of points $p = a_0, a_1, \dots, a_k = a \in S$ such that the distances $|a_{i-1} - a_i|$ between any successive points have the upper bound β for $i = 1, \dots, k$.

We will prove that $f(C(S, a_k; \alpha)) = C(Q, f(a_k); \alpha)$ by induction on k, where the base $k = 0$ is given. The induction step below goes from i to $i + 1$.

The ball $\bar{B}(a_i;\alpha)$ contains the smaller ball $\bar{B}(a_{i+1};\alpha-\beta)$ around the closely located center a_{i+1}. Indeed, since $|a_{i+1}-a_i|\le\beta$, the triangle inequality for the Euclidean distance implies that any point $a'_{i+1}\in\bar{B}(a_{i+1};\alpha-\beta)$ with $|a'_{i+1}-a_i|\le\alpha-\beta$ satisfies $|a'_{i+1}-a_i|\le|a'_{i+1}-a_{i+1}|+|a_{i+1}-a_i|\le(\alpha-\beta)+\beta=\alpha$, so $\bar{B}(a_{i+1};\alpha-\beta)\subset\bar{B}(a_i;\alpha)$. Then the inductive assumption $f(C(S,a_i;\alpha))=C(Q,f(a_i);\alpha)$ gives $f(C(S,a_{i+1};\alpha-\beta))=f(C(S,a_i;\alpha))\cap f(\bar{B}(a_{i+1};\alpha-\beta))=C(Q,f(a_i);\alpha)\cap\bar{B}(f(a_{i+1});\alpha-\beta)=C(Q,f(a_{i+1});\alpha-\beta)$.

Due to $I(S;\alpha)=I(Q;\alpha)$, the isometry class of $C(S,a_{i+1};\alpha)$ equals an isometry class of $C(Q,b_{i+1};\alpha)$ for some point $b_{i+1}\in Q$, i.e. there is an isometry $g\in\mathrm{O}(\mathbb{R}^n;a_{i+1},b_{i+1})$ such that $g(C(S,a_{i+1};\alpha))=C(Q,b_{i+1};\alpha)$. Since $f\circ g^{-1}\in\mathrm{O}(\mathbb{R}^n;b_{i+1})$ isometrically maps $C(Q,b_{i+1};\alpha-\beta)$ to $C(Q,f(a_{i+1});\alpha-\beta)$, the points $b_{i+1},f(a_{i+1})\in Q$ are in the same $(\alpha-\beta)$-equivalence class of Q.

By condition (3.5a), the splitting of the periodic point set $Q\subset\mathbb{R}^n$ into α-equivalence classes coincides with its splitting into $(\alpha-\beta)$-equivalence classes. Hence the points $b_{i+1},f(a_{i+1})\in Q$ are in the same α-equivalence class of Q. Then $C(Q,f(a_{i+1});\alpha)$ is isometric to $C(Q,b_{i+1};\alpha)=g(C(S,a_{i+1};\alpha))$.

Now we can apply Lemma A.2 for $p=a_{i+1},q=f(a_{i+1})$ and conclude that the given isometry f, which satisfies $f(C(S,a_{i+1};\alpha-\beta))=C(Q,f(a_{i+1});\alpha-\beta)$, isometrically maps the larger clusters: $f(C(S,a_{i+1};\alpha))=C(Q,f(a_{i+1});\alpha)$. The induction step is finished. The inclusion $f^{-1}(Q)\subset S$ is proved similarly. □

Lemma A.4 (all stable radii $\alpha\ge\alpha(S)$)**.** If α is a stable radius of a periodic point set $S\subset\mathbb{R}^n$, then so is any larger radius $\alpha'>\alpha$. Then all stable radii form the interval $[\alpha(S),+\infty)$, where $\alpha(S)$ is the minimum stable radius of S. ■

Proof. Due to Lemma (A.1bc), conditions (3.5ab) imply that the α'-partition $P(S;\alpha')$ and the symmetry groups $\mathrm{Sym}(S,p;\alpha')$ remain the same for all $\alpha'\in[\alpha-\beta(S),\alpha]$, where $\beta(S)$ is the bridle length. We need to show that they remain the same for any $\alpha'>\alpha$ and will apply Lemma A.3 for $S=Q$ and $\beta=\beta(S)$. Let points $p,q\in S$ be α-equivalent, i.e. there is an isometry $f\in\mathrm{O}(\mathbb{R}^n;p,q)$ such that $f(C(S,p;\alpha))=C(S,q;\alpha)$. By Lemma A.3, f isometrically maps the

full set S to itself. Then all larger α'-clusters of p, q are matched by f, so p, q are α'-equivalent and $P(S;\alpha) = P(S,\alpha')$. Similarly, any isometry $f \in \mathrm{Sym}(S,p;\alpha)$ by Lemma A.3 for $S = Q$ and $p = q$, isometrically maps the full set S to itself. Then $\mathrm{Sym}(S,p;\alpha')$ coincides with $\mathrm{Sym}(S,p;\alpha)$ for any $\alpha' > \alpha$. □

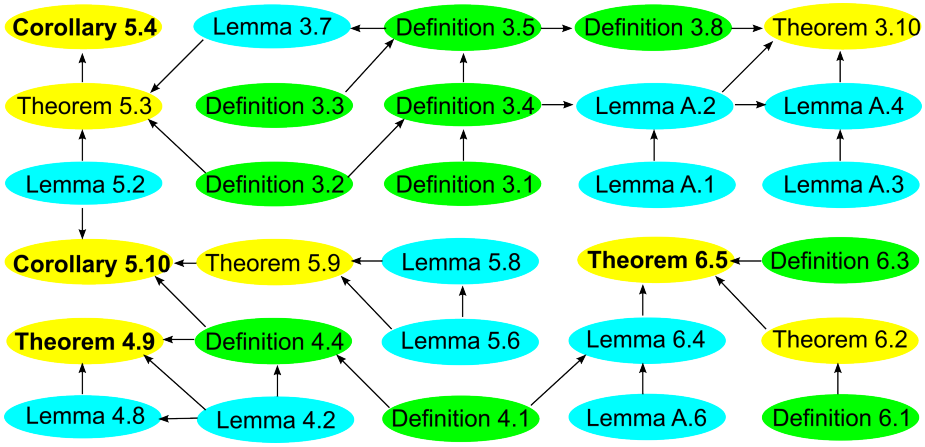


Figure 13: Logical steps towards main Theorems 4.9, 6.5 and Corollaries 5.4 and 5.10.

Proof of Theorem 3.10. The part *only if* $\Rightarrow$. Let f be an isometry of $\mathbb{R}^n$, which isometrically maps one periodic point set S to another Q. For any point p in a motif $M(S)$ of S, the image $f(p) \in Q$ is equivalent to a unique point $g(p)$ in a motif $M(Q)$ of Q modulo a translation along a vector from the lattice of Q.

Then, for any $p \in M(S)$ and $\alpha \geq 0$, the clusters $C(S,p;\alpha)$ and $C(Q,g(p);\alpha)$ are related by an isometry of $\mathbb{R}^n$. Hence the bijection $g : M(S) \to M(Q)$ induces a bijection $I(S;\alpha) \to I(Q;\alpha)$ between all isometry classes with weights.

The part *if* $\Leftarrow$. Fix a point $p \in S$. The cluster $C(S,p;\alpha)$ represents a class $\sigma \in I(S;\alpha)$. Due to $I(S;\alpha) = I(Q;\alpha)$, the class σ equals some $\xi \in I(Q;\alpha)$. Hence there is an isometry f of $\mathbb{R}^n$ such that the cluster $f(C(S,p;\alpha)) = C(Q,f(p);\alpha)$ represents ξ. By Lemma A.3, f isometrically maps S to Q. □

Proof of Lemma 4.2. Since the set $\mathrm{O}(\mathbb{R}^n;p,q)$ is compact, the minimum $\varepsilon \geq 0$ is achieved in the inclusions from (4.1b) for some isometries $f \in \mathrm{O}(\mathbb{R}^n;p,q)$ and

$g \in \mathrm{O}(\mathbb{R}^n; q, p)$. Then, for any clusters $C(S, \tilde{p}; \alpha)$ and $C(Q, \tilde{q}; \alpha)$ isometric to $C(S, p; \alpha)$ and $C(Q, q; \alpha)$ via $f_S \in \mathrm{O}(\mathbb{R}^n; \tilde{p}, p)$ and $g_Q \in \mathrm{O}(\mathbb{R}^n; \tilde{q}, q)$, respectively, the same minimum $\varepsilon \geq 0$ is achieved in the following inclusions (and vice versa), which proves the independence of $\mathrm{BT}(\sigma, \xi)$ under a choice of clusters.

$C(Q, \tilde{q}; \alpha - \varepsilon) \subseteq \tilde{f}(C(S, \tilde{p}; \alpha)) + \bar{B}(0; \varepsilon)$ for $\tilde{f} = g_Q^{-1} \circ f \circ f_S \in \mathrm{O}(\mathbb{R}^n; \tilde{p}, \tilde{q})$, and $C(S, \tilde{p}; \alpha - \varepsilon) \subseteq \tilde{g}(C(Q, \tilde{q}; \alpha)) + \bar{B}(0; \varepsilon)$ for $\tilde{g} = f_S^{-1} \circ g \circ g_Q \in \mathrm{O}(\mathbb{R}^n; \tilde{p}, \tilde{q})$.

Now we prove the coincidence axiom. By Definition 4.1, $\mathrm{BT}(\sigma, \xi) = 0$ means that some representatives of given classes σ, ξ satisfy $C(Q, q; \alpha) \subseteq f(C(S, p; \alpha))$ for some $f \in \mathrm{O}(\mathbb{R}^n; p, q)$ and $C(S, p; \alpha) \subseteq g(C(Q, q; \alpha))$ for some $g \in \mathrm{O}(\mathbb{R}^n; q, p)$. Combining these inclusions, we get $C(Q, q; \alpha) \subseteq f \circ g(C(Q, q; \alpha))$. Since $f \circ g \in \mathrm{O}(\mathbb{R}^n; q)$ is an isometry fixing q and both clusters in the inclusion above consist of the same number of points the surjection $a \mapsto f \circ g(a)$ for $a \in C(Q, q; \alpha)$ should bijective, so $C(Q, q; \alpha) = f \circ g(C(Q, q; \alpha))$. Then the initial inclusions are equalities. Hence $C(S, p; \alpha), C(Q, a; \alpha)$ are related by the isometry $f \in \mathrm{O}(\mathbb{R}^n; p, q)$, so $\sigma = \xi$. The symmetry axiom holds because the inclusions in condition (4.1b) are symmetric to each other under swapping the arguments.

To prove the triangle inequality, let clusters $C(S, p; \alpha)$, $C(Q, f(p); \alpha)$, $C(T, g \circ f(p); \alpha)$ represent σ, ξ, ζ, respectively, so that $\varepsilon_1 = \mathrm{BT}(\sigma, \xi)$ and $\varepsilon_2 = \mathrm{BT}(\xi, \zeta)$ are achieved for inclusions $C(Q, f(p); \alpha - \varepsilon_1) \subseteq f(C(S, p; \alpha)) + \bar{B}(0; \varepsilon_1)$ and $C(T, g \circ f(p); \alpha - \varepsilon_2) \subseteq g(C(Q, f(p); \alpha)) + \bar{B}(0; \varepsilon_2)$ for isometries f, g of $\mathbb{R}^n$.

The last inclusion gives $C(T, g \circ f(p); \alpha - \varepsilon_1 - \varepsilon_2) \subseteq g(C(Q, f(p); \alpha - \varepsilon_1)) + \bar{B}(0; \varepsilon_2)$ because we can reduce the radius α in the cluster $g(C(Q, f(p); \alpha))$ to $\alpha - \varepsilon_1$. Indeed, if a point $t \in C(T, g \circ f(p); \alpha - \varepsilon_1 - \varepsilon_2)$ is covered by a closed ball $\bar{B}(q; \varepsilon_2)$ for some $q \in g(C(Q, f(p); \alpha))$, then $|q - t| \leq \varepsilon_2$ and

$$|q - g \circ f(p)| \leq |q - t| + |t - g \circ f(p)| \leq \varepsilon_2 + (\alpha - \varepsilon_1 - \varepsilon_2) = \alpha - \varepsilon_1.$$

Hence q belongs to the smaller cluster $g(C(Q, f(p); \alpha - \varepsilon_1)$ as required. Now we apply the isometry g to the inclusion $C(Q, f(p); \alpha - \varepsilon_1) \subseteq f(C(S, p; \alpha)) + \bar{B}(0; \varepsilon_1)$ to get $C(T, g \circ f(p); \alpha - \varepsilon_1 - \varepsilon_2) \subseteq g(C(Q, f(p); \alpha - \varepsilon_1)) + \bar{B}(0; \varepsilon_2) \subseteq g \circ f(C(S, p; \alpha)) + \bar{B}(0; \varepsilon_1 + \varepsilon_2)$ as $(q + \bar{B}(0; \varepsilon_1)) + \bar{B}(0; \varepsilon_2) = q + \bar{B}(0; \varepsilon_1 + \varepsilon_2)$.

Swapping the roles of S, T in the arguments above, we similarly prove that if $C(S,p;\alpha-\varepsilon_1) \subseteq f(C(Q,f^{-1}(p);\alpha)) + \bar{B}(0;\varepsilon_1)$ and $C(Q,f^{-1}(p);\alpha-\varepsilon_2) \subseteq g(C(T,g^{-1}\circ f^{-1}(p);\alpha)) + \bar{B}(0;\varepsilon_2)$ for some isometries f, g of $\mathbb{R}^n$, then

$$C(S,p;\alpha-\varepsilon_1-\varepsilon_2) \subseteq f\circ g(C(T,g^{-1}\circ f^{-1}(p);\alpha)) + \bar{B}(0;\varepsilon_1+\varepsilon_2).$$

Definition 4.1 implies that $\mathrm{BT}(\sigma,\zeta) \le \varepsilon_1+\varepsilon_2 = \mathrm{BT}(\sigma,\xi)+\mathrm{BT}(\xi,\zeta)$. □

Example A.5 (detailed computations for Example 4.3)**.** Fig. 6 shows the stable 2-clusters $C(\Lambda_4,0;2)$ and $C(\Lambda_6,0;2)$ of the square (Λ_4) and hexagonal (Λ_6) lattices. Without rotations, the 1st picture of Fig. 6 shows the directed Hausdorff distance $d_{\vec{H}} = \sqrt{(1-\frac{\sqrt{3}}{2})^2+(\frac{1}{2})^2} = \sqrt{2-\sqrt{3}} \approx 0.52$ between clusters with the added boundary circle $\partial B(0;2)$. Due to high symmetry, it suffices to consider rotations of the square vertex $(1,1)$ for angles $\gamma \in [45°, 60°]$ because all other ranges can be isometrically mapped to this range for another vertex of the square. We find the squared distances $s_1(\gamma)$ and $s_2(\gamma)$ from the vertex $(\sqrt{2}\cos\gamma, \sqrt{2}\sin\gamma)$ rotated from $(1,1)$ at $\gamma = 45°$ through the angle $\gamma - 45°$ to its closest neighbors $(\frac{1}{2}, \frac{\sqrt{3}}{2})$ and $(\frac{3}{2}, \frac{\sqrt{3}}{2})$ in $C(\Lambda_6,0;2)$.

$s_1(\gamma) = \left|(\sqrt{2}\cos\gamma, \sqrt{2}\sin\gamma) - \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right)\right|^2 = \left(\sqrt{2}\cos\gamma - \frac{1}{2}\right)^2 + \left(\sqrt{2}\sin\gamma - \frac{\sqrt{3}}{2}\right)^2 = 3 - \sqrt{2}\cos\gamma - \sqrt{6}\sin\gamma$, $\frac{ds_1}{d\gamma} = \sqrt{2}\sin\gamma - \sqrt{6}\cos\gamma = 0$, $\tan\gamma = \sqrt{3}$, $\gamma = 60°, s_1 = (\sqrt{2}-1)^2$ is minimal for the points in $y = \sqrt{3}x$ at distances $1, \sqrt{2}$ from 0.

$s_2(\gamma) = \left|(\sqrt{2}\cos\gamma, \sqrt{2}\sin\gamma) - \left(\frac{3}{2}, \frac{\sqrt{3}}{2}\right)\right|^2 = \left(\sqrt{2}\cos\gamma - \frac{3}{2}\right)^2 + \left(\sqrt{2}\sin\gamma - \frac{\sqrt{3}}{2}\right)^2 = 5 - 3\sqrt{2}\cos\gamma - \sqrt{6}\sin\gamma$, $\frac{ds_2}{d\gamma} = 3\sqrt{2}\sin\gamma - \sqrt{6}\cos\gamma = 0, \gamma = 30°, s_2 = (\sqrt{3}-\sqrt{2})^2$ is minimal for the points in $y = \frac{x}{\sqrt{3}}$ at distances $\sqrt{2}, \sqrt{3}$ from 0.

It might look that the second minimum is smaller. However, for the angle $\gamma = 30°$, another vertex $(-1,1)$ rotated through $\gamma - 45° = -15°$ has distance $\sqrt{2}-1$ to its closest neighbor $(-\frac{1}{2}, \frac{\sqrt{3}}{2}) \in C(\Lambda_6,0;2)$. For any angle $\gamma \in [45°, 60°]$, the second function has the minimum $s_2(45°) = 2-\sqrt{3} = d^2_{\vec{H}}$ in the 1st picture of Fig. 6. Hence the vertex $(1,1)$ has the minimum distance $\sqrt{2}-1 \approx 0.41 < \sqrt{2-\sqrt{3}} \approx 0.52$ in the 3rd picture of Fig. 6. All other points

of the square cluster $C(\Lambda_4, 0; 2)$ are even closer to their neighbors in $C(\Lambda_6, 0; 2)$. For example, the point $(1, 0)$ rotated by 15° has the distance to $(1, 0)$ equal to $\sqrt{(\cos 15^\circ - 1)^2 + \sin^2 15^\circ} \approx 0.26$. The final picture in Fig. 6 confirms that all points of the hexagonal cluster $C(\Lambda_6, 0; 2)$ are covered by the $(\sqrt{2} - 1)$-offset of $C(\Lambda_4, 0; 2)$ and the boundary circle. So $\mathrm{BT} = \sqrt{2} - 1 \approx 0.41$. ■

Proof of Lemma 4.5. The symmetry axiom holds because Definition 4.4 is symmetric under swapping S, Q. The triangle axiom was proved for any weighted distributions in [18, Appendix A]. We will prove that if $\mathrm{EMD}(I(S;\alpha), I(Q;\alpha)) = 0$ then $I(S;\alpha) = I(Q;\alpha)$ as isosets. Indeed, $\sum_{i=1}^{m(S)} \sum_{j=1}^{m(Q)} f_{ij} \mathrm{BT}(\sigma_i, \xi_j) = 0$ means that, for any i, j, if $f_{ij} > 0$ then $\mathrm{BT}(\sigma_i, \xi_j) = 0$, so $\sigma_i = \xi_j$ by the coincidence axiom of BT from Lemma 4.2(a). Hence any flow $f_{ij} > 0$ is always between equal isometry classes. The conditions on weights of σ_i, ξ_j in Definition 4.4 imply that every class σ_i should 'flow' to its equal class ξ_j of the same weight. These flows define a bijection $I(S;\alpha) \to I(Q;\alpha)$ respecting all weights. □

Proof of Lemma 4.8. Choose the origin $0 \in \mathbb{R}^n$ at a point of S. Applying translations, we can assume that primitive unit cells $U(S), U(Q)$ of the given periodic sets S, Q have a vertex at the origin 0. Then $S = \Lambda(S) + (U(S) \cap S)$ and $Q = \Lambda(Q) + (U(Q) \cap Q)$, where $\Lambda(S), \Lambda(Q)$ are lattices of S, Q, respectively.

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

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

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

These points for $m \in \mathbb{Z}$ lie in a straight line with gaps $|q(j) - q(i)| < \delta$. The open balls with the packing radius $r(Q)$ and centers at all points of Q do not overlap. Hence all closed balls with the radius $d_B(S,Q) < r(Q)$ and the same centers are at least 2δ away from each other. Due to $|q(j) - q(i)| < \delta = r(Q) - d_B(S,Q)$, there is $m \in \mathbb{Z}$ such that $q(i) + m(q(j) - q(i))$ is outside the union $Q + \bar{B}(0; d_B(S,Q))$ of all these smaller balls. Then $q(i) + m(q(j) - q(i))$ has a distance more than $d_B(S,Q)$ from any point of Q. The translations along all vectors of the lattice $\Lambda(Q)$ preserve the union of balls $Q + \bar{B}(0; d_B(S,Q))$. Then the point $(i + m(j-i))p \in S$, which is equivalent to $q(i) + m(q(j) - q(i))$ modulo $\Lambda(Q)$, has a distance more than $d_B(S,Q)$ from any point of Q. This conclusion contradicts Definition 2.1(b) of the bottleneck distance $d_B(S,Q)$. $\square$

Proof of Lemma 5.2. To find all points in $C(S,p;\alpha)$, we will extend U by adding adjacent cells in'spherical' shells around U. After considering the initial cell U with a basis $\vec{v}_1, \ldots, \vec{v}_n$, we take $3^n - 1$ cells $U + \vec{v}$ for vectors $\vec{v} = \sum_{i=1}^{n} c_i \vec{v}_i \in \Lambda - \{0\}$ with integer coordinates $c_i \in \{-1, 0, 1\}$. The next 'spherical' shell consists of $5^n - 3^n$ cells $U + \vec{v}$ and so on. For any shifted cell $U + \vec{v}$ with $v \in \Lambda$, if all vertices have distances more than α to p, this cell is discarded. Otherwise, we check if any translated points $M + \vec{v}$ are within the closed ball $\bar{B}(p;\alpha)$ of radius α. The upper union $\bar{U} = \bigcup \{(U + \vec{v}) : v \in \Lambda, (U + \vec{v}) \cap \bar{B}(p;\alpha) \neq \emptyset\}$ consists of $\frac{\mathrm{vol}[\bar{U}]}{\mathrm{vol}[U]}$ cells and is contained in the larger ball $B(p; \alpha + d)$, because any shifted cell $U + \vec{v}$ within $\bar{U}$ has the longest diagonal d and intersects $B(p;\alpha)$. Since each $U + \vec{v}$ contains m points of S, we check at most $m\frac{\mathrm{vol}[\bar{U}]}{\mathrm{vol}[U]}$ points. So

$$|C(S,p;\alpha)| \leq m\frac{\mathrm{vol}[\bar{U}]}{\mathrm{vol}[U]} \leq m\frac{\mathrm{vol}[B(p;\alpha+d)]}{\mathrm{vol}[U]} = m\frac{(\alpha+d)^n V_n}{\mathrm{vol}[U]} = \nu(U,\alpha,n)m,$$

where $\nu(U,\alpha,n) = \frac{(\alpha+d)^n V_n}{\mathrm{vol}[U]}$. We will estimate $\nu(U,\alpha,n)$ using the upper bound $\alpha \leq (L+m+1)2R(S)$ from Lemma 3.6(c). Since the longest diagonal has the upper bound $2R(S) \geq d$ because the closed balls with the radius $\frac{d}{2}$ and centers at the vertices of a unit cell U cover U, so $\alpha + d \leq (L+m+2)2R(S)$.

Since $\Gamma(\frac{n}{2}+1) = \sqrt{\pi}\frac{(2n-1)(2n-3)\ldots 1}{2^n} = \sqrt{\pi}\frac{(2n)(2n-1)(2n-2)(2n-3)\ldots 1}{2^n(2n)(2n-2)\ldots 2} = \sqrt{\pi}\frac{(2n)!}{2^{2n}n!}$, the volume of the unit ball becomes $V_n = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)} = (\sqrt{\pi})^{n-1}\frac{2^{2n}n!}{(2n)!}$. The

bounds $\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \exp(\frac{1}{12n+1}) < n! < \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \exp(\frac{1}{12n})$, imply that $V_n = (\sqrt{\pi})^{n-1}\frac{2^{2n}n!}{(2n)!} \le \frac{(\sqrt{\pi})^{n-1}}{\sqrt{2}} 2^{2n}\left(\frac{n}{e}\right)^n\left(\frac{e}{2n}\right)^{2n}\exp(\frac{1}{12n}-\frac{1}{24n+1}) \le \frac{\exp(\frac{1}{22})}{\sqrt{2\pi}}\left(\frac{e\sqrt{\pi}}{n}\right)^n$ because $\frac{1}{12n}-\frac{1}{24n+1} = \frac{(24n+1)-12n}{12n(24n+1)} \le \frac{12n+1}{12\times 24n} \le \frac{13n}{12\times 24n} < \frac{1}{22}$ for $n \ge 1$. Then $V_n \le \frac{\exp(\frac{1}{22})}{\sqrt{2\pi}}\left(\frac{e\sqrt{\pi}}{n}\right)^n$ implies that $\nu(U,\alpha,n) = \frac{(\alpha+d)^n V_n}{\mathrm{vol}[U]} \le \frac{((L+m+2)2R(S))^n V_n}{\mathrm{vol}[U]} \le \frac{\exp(\frac{1}{22})}{\sqrt{2\pi}}\frac{((L+m+2)2R(S)e\sqrt{\pi}/n)^n}{\mathrm{vol}[U]} \le \frac{(10(L+m+2)R(S)/n)^n}{2\mathrm{vol}[U]} = \mathrm{GC}(S)$ as required. □

Proof of Lemma 5.6. The directed distance $d_{\vec{R}}(C\cup\partial\bar{B}(0;\alpha), D\cup\partial\bar{B}(0;\alpha))$ is the minimum $\varepsilon \in [0,\alpha]$ such that, for some $f \in \mathrm{O}(\mathbb{R}^n)$, all points of $C \cap B(0;\alpha-\varepsilon)$ are covered by $f(D)+\bar{B}(0;\varepsilon)$ as all points of $C \setminus B(0;\alpha-\varepsilon)$ are ε-close to the boundary $\partial\bar{B}(0;\alpha)$. Let $j \in \{1,\dots,k\}$ be the largest index so that $|p_j| < \alpha-\varepsilon$. Then $C\cap B(0;\alpha-\varepsilon) = \{p_1,\dots,p_j\}$ and $d_{\vec{R}}(\{p_1,\dots,p_i\},D) \le d_{\vec{R}}(C\cap B(0;\alpha-\varepsilon),D) \le \varepsilon$ for all $i=1,\dots,j$. By the above choice of j, if $j<i\le k$ then $\alpha-|p_i| \le \varepsilon$. Hence, for all $i=1,\dots,k$, both terms in the minimum $\min\{\ \alpha-|p_i|,\ d_{\vec{R}}(\{p_1,\dots,p_i\},D)\ \}$ are at most ε. Then $d_{\vec{M}}(C,D) \le \varepsilon$.

Using the brief notation $d_{\vec{M}} = d_{\vec{M}}(C,D)$, to prove the converse inequality $d_{\vec{R}}(C\cup\partial\bar{B}(0;\alpha), D\cup\partial\bar{B}(0;\alpha)) \le d_{\vec{M}}$, we check below that $C\cap\bar{B}(0;\alpha-d_{\vec{M}})$ is covered by $f(D)+\bar{B}(0;d_{\vec{M}})$ for some $f\in\mathrm{O}(\mathbb{R}^n)$ or, equivalently, the inequality $d_{\vec{R}}(C\cap\bar{B}(0;\alpha-d_{\vec{M}}),D) \le d_{\vec{M}}$ holds. Let $j\in\{1,\dots,k\}$ be the largest index so that $|p_j| \le \alpha-d_{\vec{M}}$. Since $\alpha-|p_j| \ge d_{\vec{M}}$ and $\min\{\ \alpha-|p_j|,\ d_{\vec{R}}(\{p_1,\dots,p_j\},D)\ \} \le \max\limits_{i=1,\dots,k}\min\{\alpha-|p_i|,\ d_{\vec{R}}(\{p_1,\dots,p_i\},D)\} = d_{\vec{M}}$, the term $d_{\vec{R}}(\{p_1,\dots,p_j\},D)$ in the minimum above is at most $d_{\vec{M}}$. Due to $C\cap\bar{B}(0;\alpha-d_{\vec{M}}) = \{p_1,\dots,p_j\}$, we get $d_{\vec{R}}(C\cap\bar{B}(0;\alpha-d_{\vec{M}}),D) \le d_{\vec{M}}$, which proves the required equality. □

Proof of Lemma 5.8. Let $q_1 \in D$ be a point that has a maximum distance to the origin $0\in\mathbb{R}^n$. If there are several points at the same maximum distance, choose any of them. Similar choices below do not affect the estimates. For any $1<i<n$, let q_i be a point of D that has a maximum perpendicular distance to the linear subspace spanned by the previously defined vectors $\vec{q}_1,\dots,\vec{q}_{i-1}$.

The key idea is to replace the minimization in $d_{\vec{R}}(C_j,D)$ over infinitely many $f\in\mathrm{O}(\mathbb{R}^n)$ by a finite minimization over compositions $f_{n-1}[s_{n-1}]\circ\ldots\circ f_1[s_1] \in \mathrm{O}(\mathbb{R}^n)$ depending on finitely many unknown points $s_1,\dots,s_{n-1}\in C_j$, which

will be exhaustively checked in time $O(\langle C \rangle^{n-1})$.

If the point q_1 belongs to the infinite straight line $L(s_1)$ through the points s_1 and 0, then set $f_1[s_1]$ to be the identity map. Otherwise, let $f_1[s_1] \in \mathrm{SO}(\mathbb{R}^n)$ fix the linear subspace orthogonal to the plane spanned by $\vec{s}_1, \vec{q}_1$, and then rotate the point q_1 to $L(s_1)$ through the smallest possible angle. Since q_1 is a furthest point of D from the origin 0 and $|s_1 - q_1| \leq d_j$, the rotation $f_1[s_1]$ moves q_1 and hence any other point of D by at most d_j. Then any point in $f_1[s_1](D)$ is at most $2d_j$ away from its closest neighbor in C_j, so $d_{\vec{H}}(C_j, f_1[s_1](D)) \leq 2d_j$.

For any $1 < i < n$, if $\vec{q}_i$ belongs to the linear subspace $L(q_1, \ldots, q_{i-1}, s_i)$ spanned by $\vec{q}_1, \ldots, \vec{q}_{i-1}, \vec{s}_i$, set $f_i[s_i]$ to be the identity map. Else let the rotation $f_i[s_i] \in \mathrm{SO}(\mathbb{R}^n)$ fix the linear subspace orthogonal to $\vec{q}_1, \ldots, \vec{q}_i, \vec{s}_i$, and rotate q_i to $L(q_1, \ldots, q_{i-1}, s_i)$ through the smallest possible angle. Since $f_1[s_1](q_2)$ is at most $2d_j$ away from $s_2 \in C_j$, the map $f_2[s_2]$ moves $f_1[s_1](q_2)$ and hence any other point of $f_1[s_1](D)$, by at most $2d_j$. Since q_2 had a maximum perpendicular distance from the line through $\vec{q}_1$, the composition $f_2[s_2] \circ f_1[s_1]$ moves any point of D by at most $d_j + 2d_j = 3d_j$. For $2 < i < n$, the composition $f_{n-1}[s_{n-1}] \circ \ldots \circ f_1[s_1]$ moves any point of D by the maximum distance $d_j + 2d_j + \ldots + (n-1)d_j = \frac{n(n-1)}{2} d_j = (\omega - 1)d_j$. Define the rotated image $D' = f_{n-1}[s_{n-1}] \circ \cdots \circ f_1[s_1](D)$ based on the neighbors $s_1, \ldots, s_{n-1} \in C_j$ of $q_1, \ldots, q_{n-1} \in D$, respectively. Since the subcloud C_j is covered by the d_j-offset of D and hence by the ωd_j-offset of D', the approximation $d_{\vec{H}}(C_j, D')$ is non-strictly between the exact distance $d_j = d_{\vec{H}}(C_j, D)$ and its upper bound ωd_j.

The algorithm starts by finding $n - 1$ 'farthest' points $q_1, \ldots, q_{n-1} \in D$, which are independent of j, in time $O(n|D|)$. Here q_1 is a point of D with a maximum distance $|q_1|$ to the origin, q_2 is a next 'farthest' point of D from 0 and so on. If some of these points have equal distances to 0, they can be chosen in any order. Though we do not know which points $s_1, \ldots, s_{n-1} \in C_j$ become nearest neighbors of $q_1, \ldots, q_{n-1} \in D$ after an optimal rotation, we will consider all (unit vectors of) points $s_1, \ldots, s_{n-1} \in C_j$ to compute the approximate distance $d'_j = \min\limits_{s_1, \ldots, s_{n-1} \in C_j} d_{\vec{H}}(C_j, D')$ for $j = 1, \ldots, k$. The minimization for all such

points keeps both bounds: $d_j \le d'_j \le \omega d_j$.

To minimize choices for $s_1, \ldots, s_{n-1}$, we remove from the ordered list $p_1, \ldots, p_k$ all points p_i whose unit vectors $\vec{p}_i/|\vec{p}_i|$ appear earlier with smaller indices. Then we consider each variable point s_i from the remaining list C' in increasing order of distances from the origin. For any chosen points $s_1, \ldots, s_{n-1}$, we compute the rotated image $D' = f_{n-1}[s_{n-1}] \circ \cdots \circ f_1[s_1](D)$ by computing matrix products in time $O(|D|)$, where we skip polynomial factors of the fixed dimension n.

To compute the approximation $d'_j = d'_{\vec{R}}(C_j, D) = \min_{s_1,\ldots,s_{n-1} \in C_j} d_{\vec{H}}(C_j, D')$, we keep the current minimum of d'_j, which will be updated after getting the distance $d_{\vec{H}}(C_j, D')$ for every new choice of $s_1, \ldots, s_{n-1} \in C_j$. For each rotated image $D' = f_{n-1}[s_{n-1}] \circ \cdots \circ f_1[s_1](D)$, we run the internal loop for $j = 1, \ldots, k$. For each point p_j from the ordered full cloud C, we compute the distance $d(p_j, D') = \min_{q \in D'} |p_j - q|$ to its nearest neighbor in D' in time $O(|D|)$. We use the previous iteration for $j-1$ to get $d_{\vec{H}}(C_j, D') = \max\{d_{\vec{H}}(C_{j-1}, D'), d(p_j, D')\}$, where we set $d_{\vec{H}}(C_0, D') = 0$. If all $s_1, \ldots, s_{n-1} \in C_j$ and the current value of d'_j is larger than $d_{\vec{H}}(C_j, D')$, we update $d'_j := d_{\vec{H}}(C_j, D')$. Because C has $O(\langle C \rangle^{n-1})$ points $s_1, \ldots, s_{n-1}$ with distinct normalized vectors, which determine $D' = f_{n-1}[s_{n-1}] \circ \cdots \circ f_1[s_1](D)$, the total time is $O(|C|\langle C \rangle^{n-1}|D|)$. □

Lemma A.6 implies that ordering lists of pairs of ε-perturbations will keep ε-closeness of corresponding ordered values. This result will help prove Lemma 6.4.

Lemma A.6 (re-ordering of ε-close values)**.** For any $\varepsilon \ge 0$, let $C = \{c_1, \ldots, c_k\}$ and $D = \{d_1, \ldots, d_k\}$ satisfy $|c_i - d_i| \le \varepsilon$, $i = 1, \ldots, k$. For any $i = 1, \ldots, k$, let $c_{(i)}, d_{(i)}$ be the i-th largest values in C, D, respectively. Then $|c_{(i)} - d_{(i)}| \le \varepsilon$. ■

Proof. We consider any real d_i an ε-perturbation of the corresponding value c_i for $i = 1, \ldots, k$. Assume towards contradiction that $c_{(i)} < d_{(i)} - \varepsilon$, so all i smallest values of C are less than $d_{(i)} - \varepsilon$. Then all ε-perturbations of these i values in D are less than $d_{(i)}$, so D has i values that are strictly smaller than $d_{(i)}$. This conclusion contradicts that $d_{(i)}$ is the i-th largest value in D. The

assumption $c_{(i)} > d_{(i)} + \varepsilon$ similarly leads to contradiction. Hence, after writing both C, D in increasing order, their i-th largest values remain ε-close. □

Proof of Lemma 6.4. Definition 4.1c of $\varepsilon = \mathrm{BT}([C(S,p;\alpha)],[C(Q,q;\alpha)])$ implies that, for a suitable isometry $f \in \mathrm{O}(\mathbb{R}^n)$, the image $f(C(S,p;\alpha-\varepsilon)-\vec{p})$ is covered by the ε-offset of $C(Q;q;\alpha)-\vec{q}$ shifted by q to the origin. Since ε is smaller than a minimum half-distance between points of S, Q, the above covering establishes a bijection g with all (at least k) neighbors of p and q in their $(\alpha-\varepsilon)$-clusters.

The covering condition above means that the corresponding neighbors are at a maximum distance ε from each other. The triangle inequality implies that the distances from corresponding neighbors to their centers p, q differ by at most ε. The ordered distances from p, q to their k neighbors in the $(\alpha-\varepsilon)$-clusters form the rows of p, q in $\mathrm{PDD}(S;k), \mathrm{PDD}(Q;k)$. The bijection g may not respect their order. By Lemma A.6 the ordered distances with the same indices are ε-close. So the L_∞ distance between the rows of p, q is at most ε. □

The proof of Lemma 3.6(c) referred to Lemma A.7, which was briefly proved in the 2nd paragraph in [32, p. 20] without a formal statement. We stated and prove this auxiliary result below to make all arguments complete.

Lemma A.7 (finite symmetry group)**.** Let a periodic point set $S \subset \mathbb{R}^n$ be n-dimensional, i.e. S is not contained a lower-dimensional affine subspace of $\mathbb{R}^n$. Then the symmetry group $\mathrm{Sym}(S,p;2R(S))$ is finite for any point $p \in S$.

Proof. Recall that the covering radius $R(S)$ is the largest radius R of an open ball $B(q;R)$ within the complement $\mathbb{R}^n \setminus S$ for $q \in \mathbb{R}^n$. Consider any such ball $\bar{B}(q;R(S))$ whose boundary sphere passes through the given point $p \in S$ and whose interior contains no points of S. Then the closed ball $\bar{B}(q;R(S))$ should include at least one more point $p_1 \in S \setminus \{p\}$ with $|p-p_1| \le 2R(S)$. Otherwise, the ball $\bar{B}(q;R(S))$ can be slightly expanded from $p \in S$ without including any points of $S \setminus \{p\}$, which contradicts the definition of the covering radius $R(S)$.

If $n \ge 2$, consider another open ball $B(q_1;R(S)) \subset \mathbb{R}^n \setminus S$ that touches at p the straight line $L(p,p_1)$ through p, p_1. Then the closed ball $\bar{B}(q_1;R(S))$

should include at least one more point $p_2 \in S$ outside the line $L(p, p_1)$, so $|p - p_2| \le 2R(S)$ and $p, p_1, p_2 \in S$ span the 2-dimensional plane $L(p, p_1, p_2)$. Otherwise the closed ball $\bar{B}(q_1; R(S))$ can be slightly expanded from $p \in S$ on the boundary $\partial B(q_1; R(S))$ without including any points of $S \setminus \{p, p_1\}$.

If $n \ge 3$, consider another open ball $B(q_2; R(S)) \subset \mathbb{R}^n \setminus S$ that touches at p the plane $L(p, p_1, p_2)$. Then the closed ball $\bar{B}(q_2; R(S))$ should include at least one more point $p_3 \in S$ outside the plane $L(p, p_1, p_2)$. Then $p, p_1, p_2, p_3 \in S$ span the 3-dimensional subspace $L(p, p_1, p_2, p_3)$ and so on until we find $n+1$ affinely independent points $p, p_1, \dots, p_n \in S$ such that each p_i is at a maximum distance $2R(S)$ from p for $i = 1, \dots, n$. Since the cluster $C(S, p; 2R(S))$ contains $n+1$ affinely independent points, its symmetry group $\mathrm{Sym}(S, p; 2R(S))$ is finite. □

Example A.8 (Earth Mover's Distance for lattices with bottleneck distance $d_B = +\infty$)**.** The 1D lattices $S = \mathbb{Z}$ and $Q = (1+\delta)\mathbb{Z}$ with the bottleneck distance $d_B(S, Q) = +\infty$ have PDD consisting of a single row (as for any lattice). For instance, $\mathrm{PDD}(S; 4) = (1, 1, 2, 2)$ and $\mathrm{PDD}(Q; 4) = (1+\delta, 1+\delta, 2+2\delta, 2+2\delta)$. For the common stable radius $\alpha = 2 + 2\delta$, Example 4.6 computed $\mathrm{EMD}(I(S; \alpha), I(Q; \alpha)) = 2\delta$. Theorem 6.5 considers the maximum number k of points in clusters of S, Q with the radius $\alpha - 2\delta = 2$, so $k = 2$.

Then $\mathrm{EMD}(\mathrm{PDD}(S; 2), \mathrm{PDD}(Q; 2))$ equals the L_∞ distance δ between the short rows $(1, 1)$ and $(1+\delta, 1+\delta)$. The above computations illustrate the lower bound $\mathrm{EMD}(\mathrm{PDD}(S; 2), \mathrm{PDD}(Q; 2)) = \delta \le \mathrm{EMD}(I(S; \alpha), I(Q; \alpha)) = 2\delta$. This inequality becomes equality for the larger stable radius $\alpha = 2 + 4\delta$, because the clusters of S, Q with the radius $\alpha - 2\delta = 2 + 2\delta$ contain $k = 4$ points. The L_∞ distance between $(1, 1, 2, 2)$ and $(1+\delta, 1+\delta, 2+2\delta, 2+2\delta)$ is 2δ for $\delta < \frac{1}{8}$, so $\mathrm{EMD}(\mathrm{PDD}(S; 4), \mathrm{PDD}(Q; 4)) = 2\delta = \mathrm{EMD}(I(S; 2+4\delta), I(Q; 2+4\delta))$. ■

Example A.9 (lower bound for a distance between square and hexagonal lattices)**.** The square lattice Λ_4 and hexagonal lattice Λ_6 with minimum inter-point distance 1 have a common stable radius $\alpha = 2$ as shown in Fig. 6. The maximum number of points in the stable 2-clusters is $k = 12$. The

rows $\mathrm{PDD}(\Lambda_4;12) = (1,1,1,1,\sqrt{2},\sqrt{2},\sqrt{2},\sqrt{2},2,2,2,2)$ and $\mathrm{PDD}(\Lambda_6;12) = (1,1,1,1,1,1,\sqrt{3},\sqrt{3},\sqrt{3},\sqrt{3},\sqrt{3},\sqrt{3})$ have the L_∞ distance $\max\{\sqrt{2}-1, 2-\sqrt{3}\} = \sqrt{2}-1$, which coincides with $\mathrm{EMD}(I(\Lambda_4;2), I(\Lambda_6;2))$ in Example 4.3. ■

The latest version includes Algorithms 1 and 2, which implement the approximations from section 5 of new metrics defined in section 4. Tables A.1-A.2 with near-duplicates and run times on the CSD and GNoME are new.

Algorithm 1: Pseudocode for the boundary tolerant metric BT between α-clusters and EMD between isosets in Theorem 5.9 and Corollary 5.10, where the max-min distance $d_{\vec{M}}$ is approximated by Algorithm 2.

Input : Isosets `is1` $= I(S;\alpha)$, `is2` $= I(Q;\alpha)$, with weights `w1`, `w2`

Output: $\mathrm{EMD}(I(S;\alpha), I(Q;\alpha))$

```
distance_matrix = zeros(len(is1), len(is2))
for i in range(len(is1)) do
    for j in range(len(is2)) do
        BT = max(d_M(C, D), d_M(D, C))
        distance_matrix[i, j] = BT
emd = EMD(w1, w2, distance_matrix)
return emd
```

To confirm near-duplicates in the CSD and GNoME database by complete isosets, we first filtered out pairs of crystals that have $L_\infty \geq 10^{-4}$Å on faster invariants ADA100. The full table of the resulting 4385 pairs with all distances and running times is in the supplementary materials. In most pairs, crystals belong to the same 6-letter code family because their structures are either polymorphs (different phases with the same composition) or slightly different versions determined under different temperatures or pressures. However, 398 pairs consist of geometric near-duplicates that were assigned to (unexpectedly) different families. In almost all these cases, the EMD metric on isosets was 0 after rounding to 10^{-10} (floating point error) in Angstroms. Table A.1 shows all 25 crystals where the EMD metric was only slightly above 0.

Algorithm 2: Pseudocode for the directed max-min distance $d_{\vec{M}}$ in Definition 5.5(b) by using an approximation of $d_{\vec{R}}$ in Lemma 5.8.

Input : finite clouds C, D (ordered by distance to the origin 0)

Output: $d_{\vec{M}}(C, D)$

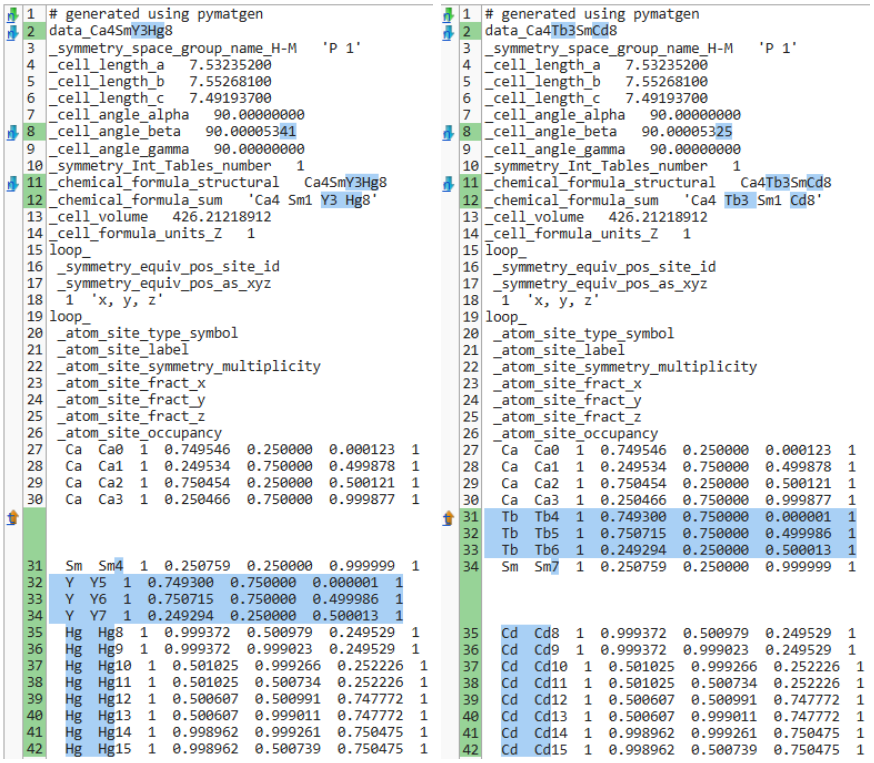
```
alpha = max(norm(C[-1]), norm(D[-1]))
q1, d_R, max_d, res = D[-1], [inf] * len(C), -inf, -inf
for q in D do
    if d := perp_dist(q, q1) > max_d then max_d = d; q2 = q
for i1, p1 in enumerate(C) do
    R1b = rotation_to_align(q1, p1)
    q1_, q2_ = dot(R1b, q1), dot(R1b, q2)
    for i2, p2 in enumerate(C) do
        if i1 == i2 then continue
        src_normal = cross(q1_, q2_)
        R2b = identity(3)
        if norm(src_normal) != 0 then
            tar_normal = cross(p1, p2)
            if norm(tar_normal) == 0 then
                if dot(p1, p2) < 0 then
                    R2b = rotation_to_align(p1, p2)
            else
                R2b = rotation_to_align(src_normal, tar_normal)
        d_H = Hausdorff_dist(C, dot(D, R2b × R1b))
        if v := min(d_H, alpha - norm(p)) > res then res = v
return res
```

Table A.1: The first pair consists of rigidly different mirror images from Fig. 8 (right). All others are geometric near-duplicates from (surprisingly) different families in the CSD, confirmed by tiny values of the EMD metric on isosets. The distance units are in *attometers*: 1 am $= 10^{-8}$Å $= 10^{-18}$ meter. The run times in milliseconds (ms) depend on the cluster size (maximum number of atoms in α-clusters) according to Theorem 5.3 and Corollary 5.10.

CSD id1	CSD id2	EMD_isosets, am	isosets time, ms	EMD_isosets time, ms	cluster size
WODLOS	XAWGAE	85856.22	129.619	1204.58	7
TAFQIA	VAVQIS	952.96	1690.949	321603.86	20
FIJKIU	IPEQUR	728.43	407.579	77455.47	16
JIZMIR01	JIZNAK	496.08	40.454	634.73	5
HIYVUG01	MASPIF	334.62	35.518	543.45	7
KIVXEW10	KIWCEC	125.03	22.456	32.32	5
XAYZOP	ZEMDAZ	89.47	301.217	1697.26	4
KIVXEW10	KIWCEC28	83.07	21.701	32.49	5
AFIBOH	NENCUF	31.67	126.582	1160.58	5
KIVXEW07	KIWCEC09	31.11	22.287	36.95	5
KIVXEW07	KIWCEC11	31.11	22.434	36.75	5
KIVXEW11	KIWCEC26	26.11	21.646	32.33	5
SERKIL	SERKOR	23.78	2444.885	18485.57	6
ADESAG	REWPOB	5.81	54.675	5689.27	15
GEQRAX	IFOQOL	0.05	265.4	2090.42	6
BUKYEN	UYOCES	0.03	398.129	15739.11	11
GOHYOT	VIHCEY	0.01	100.031	940.16	5
JUMCUP	QAHBOT	0.01	179.367	4234.6	5
CALMOV	CALNAI	0.01	128.437	3913.32	4
NABKOT	ZIVSEF	0.01	75.401	796.14	5
LIBGAE	VESJUY	0.01	41.535	403.87	3
AMEVEV	OLERON	0	70.172	558.78	4
SIHFIZ	TEZBUV	0	207.984	1761.39	5
XATCAA	ZAQMEN	0	60.254	394.74	4
PIDREA	XIZNOL	0	94.135	243.46	5

Table A.2: After excluding 3248 exact numerical duplicates from [3, Table 1], the next 25 pairs of closest near-duplicates in the GNoME database are confirmed by tiny values of the EMD metric on isosets. The distance units are *attometers*: 1 am = 10^{-8}Å = 10^{-18} meter. The run times are in milliseconds (ms). The cluster size is the maximum number of atoms in α-clusters, which affects times. The full table of 2858 pairs is in the supplementary materials.

GNoME id1	GNoME id2	EMD_isosets, am	isosets time, ms	EMD_isosets time, ms	cluster size
1547d30046	ddc216e80c	1	1.659	434.362	14
b4065a4798	e78d3559e6	1.7	3.034	13.271	6
98ab164895	df1252bc44	2	1.002	419.142	14
0de9d25713	b1733941a7	2.7	1.971	49.816	6
0e79f7c053	6cf951ac6f	3	1.035	429.487	14
07ece241f0	45cacc8d45	3.2	0.618	14.374	6
a58dc74a92	c16bf63220	4.1	2.532	641.086	14
5023e3a4b8	8f7ffb4d4a	4.6	2.776	10.02	6
3198d1a3ea	35f67abe6d	5	1.031	403.398	14
6826b81efb	76ee112799	5	0.985	407.618	14
6826b81efb	e9be17f0ee	5	1.008	404.306	14
2cff5f2fa0	f470a5f6fa	5.3	0.635	169.911	17
2ce912f039	9de239ee0c	5.5	0.632	3.456	2
c9f5a7a51b	fd9f40e0e1	6	1.14	195.261	10
18078e002b	aca2a892a5	6	1.028	421.009	14
18078e002b	b9722429b1	6	1.18	445.453	14
18078e002b	b702e73db3	6	1.035	414.325	14
34b4204eee	adee17535b	6	1.017	396.855	14
506b8b5646	60d266db80	6	1.174	413.254	14
506b8b5646	ec7b789cb3	6	1.174	403.014	14
780741962f	a19688f106	6.5	1.804	870.629	15
780741962f	c6af1fc763	6.5	2.731	921.496	15
780741962f	c64c3e245c	6.5	1.792	829.979	15
b06353561c	b6d2341d32	6.6	2.387	259.359	12
ebb33e044c	ebc9a4db61	6.8	1.232	450.351	14

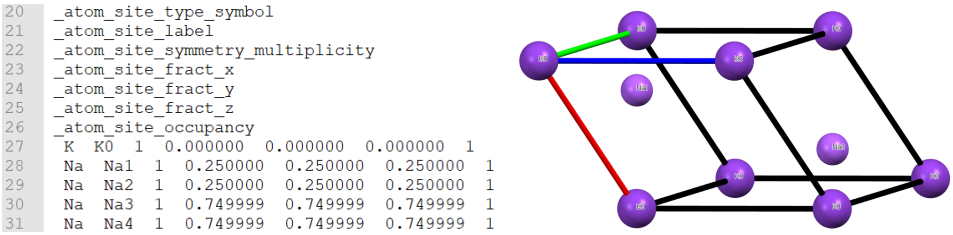


```
# generated using pymatgen
data_Ca4SmY3Hg8
_symmetry_space_group_name_H-M   'P 1'
_cell_length_a   7.53235200
_cell_length_b   7.55268100
_cell_length_c   7.49193700
_cell_angle_alpha   90.00000000
_cell_angle_beta   90.00005341
_cell_angle_gamma   90.00000000
_symmetry_Int_Tables_number   1
_chemical_formula_structural   Ca4SmY3Hg8
_chemical_formula_sum   'Ca4 Sm1 Y3 Hg8'
_cell_volume   426.21218912
_cell_formula_units_Z   1
loop_
 _symmetry_equiv_pos_site_id
 _symmetry_equiv_pos_as_xyz
  1  'x, y, z'
loop_
 _atom_site_type_symbol
 _atom_site_label
 _atom_site_symmetry_multiplicity
 _atom_site_fract_x
 _atom_site_fract_y
 _atom_site_fract_z
 _atom_site_occupancy
  Ca  Ca0  1  0.749546  0.250000  0.000123  1
  Ca  Ca1  1  0.249534  0.750000  0.499878  1
  Ca  Ca2  1  0.750454  0.250000  0.500121  1
  Ca  Ca3  1  0.250466  0.750000  0.999877  1
  Sm  Sm4  1  0.250759  0.250000  0.999999  1
  Y  Y5  1  0.749300  0.750000  0.000001  1
  Y  Y6  1  0.750715  0.750000  0.499986  1
  Y  Y7  1  0.249294  0.250000  0.500013  1
  Hg  Hg8  1  0.999372  0.500979  0.249529  1
  Hg  Hg9  1  0.999372  0.999023  0.249529  1
  Hg  Hg10  1  0.501025  0.999266  0.252226  1
  Hg  Hg11  1  0.501025  0.500734  0.252226  1
  Hg  Hg12  1  0.500607  0.500991  0.747772  1
  Hg  Hg13  1  0.500607  0.999011  0.747772  1
  Hg  Hg14  1  0.998962  0.999261  0.750475  1
  Hg  Hg15  1  0.998962  0.500739  0.750475  1
```

```
# generated using pymatgen
data_Ca4Tb3SmCd8
_symmetry_space_group_name_H-M   'P 1'
_cell_length_a   7.53235200
_cell_length_b   7.55268100
_cell_length_c   7.49193700
_cell_angle_alpha   90.00000000
_cell_angle_beta   90.00005325
_cell_angle_gamma   90.00000000
_symmetry_Int_Tables_number   1
_chemical_formula_structural   Ca4Tb3SmCd8
_chemical_formula_sum   'Ca4 Tb3 Sm1 Cd8'
_cell_volume   426.21218912
_cell_formula_units_Z   1
loop_
 _symmetry_equiv_pos_site_id
 _symmetry_equiv_pos_as_xyz
  1  'x, y, z'
loop_
 _atom_site_type_symbol
 _atom_site_label
 _atom_site_symmetry_multiplicity
 _atom_site_fract_x
 _atom_site_fract_y
 _atom_site_fract_z
 _atom_site_occupancy
  Ca  Ca0  1  0.749546  0.250000  0.000123  1
  Ca  Ca1  1  0.249534  0.750000  0.499878  1
  Ca  Ca2  1  0.750454  0.250000  0.500121  1
  Ca  Ca3  1  0.250466  0.750000  0.999877  1
  Tb  Tb4  1  0.749300  0.750000  0.000001  1
  Tb  Tb5  1  0.750715  0.750000  0.499986  1
  Tb  Tb6  1  0.249294  0.250000  0.500013  1
  Sm  Sm7  1  0.250759  0.250000  0.999999  1
  Cd  Cd8  1  0.999372  0.500979  0.249529  1
  Cd  Cd9  1  0.999372  0.999023  0.249529  1
  Cd  Cd10  1  0.501025  0.999266  0.252226  1
  Cd  Cd11  1  0.501025  0.500734  0.252226  1
  Cd  Cd12  1  0.500607  0.500991  0.747772  1
  Cd  Cd13  1  0.500607  0.999011  0.747772  1
  Cd  Cd14  1  0.998962  0.999261  0.750475  1
  Cd  Cd15  1  0.998962  0.500739  0.750475  1
```

Figure 14: The GNoME crystals 1547d30046 and ddc216e80c in the first row of Table A.2 are compared as texts by https://text-compare.com. All differences are highlighted in blue.

Fig. 15 shows the most striking pair of exact duplicates in the GNoME is cdc06a1a2a and 0e2d8f26d6, whose CIFs are identical symbol by symbol in addition to two pairs of atoms at the same positions (Na1=Na2 and Na3=Na4).



```
_atom_site_type_symbol
_atom_site_label
_atom_site_symmetry_multiplicity
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_occupancy
 K  K0  1  0.000000  0.000000  0.000000  1
 Na  Na1  1  0.250000  0.250000  0.250000  1
 Na  Na2  1  0.250000  0.250000  0.250000  1
 Na  Na3  1  0.749999  0.749999  0.749999  1
 Na  Na4  1  0.749999  0.749999  0.749999  1
```

Figure 15: Different entries cdc06a1a2a and 0e2d8f26d6 in the GNoME database are not only identical symbol by symbol but also contain two pairs of atoms (Na1=Na2 and Na3=Na4) at the same positions. **Left**: a screenshot from the CIF. **Right**: Mercury visualization can show only one atom in each pair of coinciding atoms, e.g. only Na1 and not Na2 from the CIF.