Noname manuscript No.
(will be inserted by the editor)

Higher-order, generically complete, continuous, and polynomial-time isometry invariants of periodic sets

Daniel Widdowson · Vitaliy Kurlin

Received: date / Accepted: date

Abstract Periodic point sets model all solid crystalline materials (crystals) whose atoms can be considered zero-sized points with or without atomic types. This paper addresses the fundamental problem of checking whether claimed crystals are novel, not noisy perturbations of known materials obtained by unrealistic atomic replacements. Such near-duplicates have skewed ground-truth because past comparisons relied on unstable cells and symmetries. The proposed Lipschitz continuity under noise is a new essential requirement for machine learning on any data objects that have ambiguous representations and live in continuous spaces. For periodic point sets under isometry (any distance-preserving transformation), we designed invariants that distinguish all known counter-examples to the completeness of past descriptors and detect thousands of (near-)duplicates in large high-profile databases of crystals within two days on a modest desktop computer.

Keywords Periodic set · rigid motion · isometry · invariant · metric · continuity

Mathematics Subject Classification (2020) 52C07 · 52C25 · 51N20 · 11H06
Communicated by

Acknowledgements This research was supported by the EPSRC New Horizons grant 'Inverse design of periodic crystals' (2022-2025) and the Royal Society APEX fellowship 'New geometric methods for mapping the space of periodic crystals' (2023-2025).

1 The key questions of mathematical data science for real applications

Many real data objects have infinitely many different representations. For example, any rigid object such as a solid crystalline material can be given by atomic coordinates that strongly depend on a chosen basis in Euclidean space $\mathbb{R}^3$. Hence the **first question** that mathematical data science should ask about any objects is *Same or different?* [54]. To make this question meaningful, we should rigorously define what objects can be called *the same* (or *equivalent*) as formalized below.

D.Widdowson and V.Kurlin
Computer Science, University of Liverpool, UK E-mail: vitaliy.kurlin@liverpool.ac.uk

An *equivalence* is a binary relation (denoted by $S \sim Q$) satisfying three axioms: (1) *reflexivity:* any object $S \sim S$; (2) *symmetry:* if $S \sim Q$ then $Q \sim S$; (3) *transitivity:* if $S \sim Q$ and $Q \sim T$ then $S \sim T$. Any classification needs an equivalence satisfying these axioms to split all objects into disjoint classes: the *equivalence class* $[S]$ of an object S consists of all Q equivalent to S. If two classes $[S]$ and $[T]$ share a common object Q, then $[S] = [T]$ by the transitivity axiom.

For any collection of objects, one can consider many different equivalences. For instance, any finite or periodic configurations of atoms (molecules or crystals) can be called equivalent if they have the same chemical composition. However, we know many *polymorphic* materials (such as diamond and graphite) that have the same composition but differ by other properties. In this case, we need a stronger equivalence that would split all atomic configurations into as many different classes as practically necessary to uniquely identify all physical and chemical properties.

The following equivalence is crucial for many real objects, including molecules and materials whose structures are determined in a rigid form [5]: a *rigid motion* is a composition of translation and rotations in $\mathbb{R}^n$, which preserves all object properties under the same ambient conditions such as temperature and pressure. Indeed, there is no sense in distinguishing atomic configurations that can be exactly matched by rigid motion, but it is important to see differences in *rigid shapes* (equivalence classes under rigid motion) that can affect their properties.

If we consider compositions of a rigid motion with mirror reflections, we get a slightly weaker equivalence: an *isometry* (denoted by $S \simeq Q$) is any distance-preserving transformation. Since mirror images can be distinguished by a sign of orientation, we focus on isometries, which form the full Euclidean group $\mathrm{E}(n)$.

After an equivalence (isometry in our case) is fixed, objects can be distinguished by an isometry *invariant* I that is a function mapping a given object S to a numerical value (vector or a matrix) $I(S)$ preserved under any isometry, i.e. if $S \simeq Q$, then $I(S) = I(Q)$. An example invariant of a finite set S is its size (the number of points). Any non-constant invariant I can distinguish some (not necessarily) all non-isometric sets, i.e. if $I(S) \neq I(Q)$ then $S \not\simeq Q$ by definition.

The invariance is stronger than the *equivariance* requiring that any isometry f maps $I(S)$ to $T_f(I(S))$, where a transformation T_f depends on f. For example, any linear combination $e(S)$ of coordinates of a finite set $S \subset \mathbb{R}^n$ is equivariant, not invariant, and hence allows a *false negative* that is a pair of objects $S \simeq Q$ with $e(S) \neq e(Q)$. The invariance is much stronger by requiring that T_f is the identity. Then $I(S) \neq I(Q)$ always guarantees that $S \not\simeq Q$ are not isometric.

A full answer to the question '*Same or different?*' requires a *complete* invariant I satisfying the much harder inverse implication: if $I(S) = I(Q)$ then $S \simeq Q$. In other words, I has no *false positives* that are pairs $S \not\simeq Q$ with $I(S) = I(Q)$. All triangles S (sets of three points) have a complete invariant $I(S)$ of three inter-point distances due to the side-side-side (SSS) theorem. Any complete invariant is similar to a DNA-style code that uniquely identifies any object under isometry.

A simple input of real objects is a discrete set of points, which can represent corners, edge pixels, or atomic centers in a molecule or a material. In the finite case, if given points $p_1, \dots, p_m \in \mathbb{R}^n$ are ordered, they are uniquely determined under isometry [55,35] by the matrix of pairwise Euclidean distances $|p_i - p_j|$ or the Gram matrix of scalar products $p_i \cdot p_j$, see [62, chapter 2.9] and [61].

However, most points in real objects are unordered, e.g. many materials consist of indistinguishable atoms. A brute-force extension of distance matrices to m unordered points is impractical due to the exponential cost of $m!$ permutations. In this unordered case, [9] proved that the vector of sorted pairwise distances is *generically complete* meaning that this invariant distinguishes all non-isometric finite sets in $\mathbb{R}^n$ outside some measure 0 subspace of singular sets of points.

After the case of 3 points was settled by the SSS theorem 2000+ years ago, even $m = 4$ unordered points in $\mathbb{R}^2$ did not have a better than a brute-force complete isometry invariant based on $4! = 24$ permutations, partially due to infinitely many pairs of non-isometric 4-point clouds with the same 6 pairwise distances [12]. The finite case was solved in 2023 [67] for any number m of unordered points under rigid motion in $\mathbb{R}^n$, see [65, Theorem 5.3] for a simpler complete invariant for 4 points under isometry in $\mathbb{R}^n$. We now focus on the much harder periodic case.

Definition 1.1 (lattice, motif, l-periodic set) *Vectors $v_1, \dots, v_n \in \mathbb{R}^n$ form a* basis *if any vector in $\mathbb{R}^n$ can be written as $v = \sum\limits_{i=1}^{n} t_i v_i$ for unique $t_1, \dots, t_n \in \mathbb{R}$. For $1 \leq l \leq n$, the first l vectors define the* lattice $\Lambda = \{\sum\limits_{i=1}^{l} c_i v_i \mid c_1, \dots, c_l \in \mathbb{Z}\}$ *and the* unit cell $U = \{\sum\limits_{i=1}^{n} x_i v_i \mid x_1, \dots, x_l \in [0,1), x_{l+1}, \dots, x_n \in \mathbb{R}\} \subset \mathbb{R}^n$. *If $l = n$, then U is an n-dimensional parallelepiped. If $l < n$, then U is an infinite slab over an l-dimensional parallelepiped on $v_1, \dots, v_l$. For any finite* motif *of points $M \subset U$, the sum $S = M + \Lambda = \{p + v \mid p \in M, v \in \Lambda\}$ is an l-*periodic point set*.*

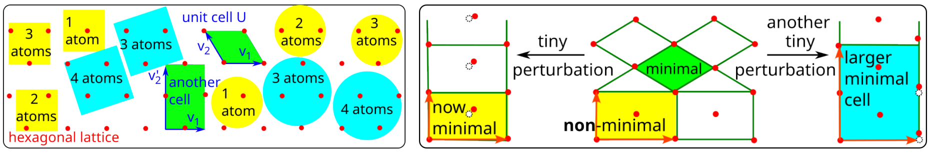


Fig. 1 **Left**: any periodic point set can be given by many pairs (cell, motif), see Definition 1.1. Any periodic set has vastly different finite subsets within boxes or balls of the same cut-off size. **Right**: almost any perturbation can arbitrarily scale up a unit cell and break the symmetry.

A classification of periodic point sets under isometry cannot be easily reduced to the finite case. Indeed, the hexagonal lattice of red points in Fig. 1 (left) has many non-isometric finite subsets of points within differently positioned boxes or balls of the same cut-off radii. A motif of points within a unit cell is also ambiguous because any lattice can be generated by infinitely many different bases, which span *primitive* (minimal by volume) unit cells of various shapes. Crystallographers developed a unique Niggli cell [45] but any such cell discontinuously scales by an arbitrary factor [66, Theorem 15] under almost all perturbations because of experimental noise [39] and atomic vibrations, see Fig. 1 (right).

Even if a complete invariant distinguishes all different objects, the space of equivalence classes is often continuous in the sense that a small perturbation produces a near-duplicate of a slightly different class. One past approach was to ignore all perturbations up to a small threshold $\varepsilon > 0$. Then the transitivity axiom can make all sets of the same size equivalent through a long enough chain of perturbations $S_1 \sim \cdots \sim S_k$, each time shifting points up to a fixed Euclidean distance $\varepsilon > 0$. Similarly, adding a single outlier should make finite sets non-equivalent, otherwise all sets of different sizes become equivalent by the transitivity axiom.

This *sorites* paradox [30] has been discussed from ancient times: while removing grains from a heap of sand one by one, when will a heap of sand suddenly stop being a heap? The discontinuity problem remained unresolved for materials [70] because isometry classes of periodic crystals still have no well-defined continuous metric. The challenges of continuous measurements are important for real objects under many equivalences and motivate the **second question** '*If different, by how much?*' in Geometric Data Science, which we formalize below for periodic sets.

Problem 1.2 *For all periodic sets $S \subset \mathbb{R}^n$ with up to m of points in a unit cell, find an invariant I with values in a metric space satisfying the conditions below.*

(a) ***Completeness*** *(injectivity): any periodic point sets $S, Q \subset \mathbb{R}^n$ are isometric if and only $I(S) = I(Q)$, i.e. I has* no false negatives *and* no false positives.

(b) ***Invertibility*** *(reconstruction): any periodic point set $S \subset \mathbb{R}^n$ can be reconstructed from its invariant $I(S)$, uniquely under isometry of $\mathbb{R}^n$.*

(c) ***Lipschitz continuity****: there is a distance metric d on invariant values satisfying all metric axioms (1) $d(a,b) = 0$ if and only if $a = b$, (2) $d(a,b) = d(b,a)$, (3)* triangle inequality *$d(a,b) + d(b,c) \geq d(a,c)$ for all a, b, c; and a constant λ such that, for any $\varepsilon > 0$, if a periodic point set Q is obtained by perturbing every point of a periodic point set S up to Euclidean distance ε, then $d(I(S), I(Q)) \leq \lambda\varepsilon$.*

(d) ***Computability****: for a fixed dimension n, the invariant $I(S)$, the metric d and the reconstruction of $S \subset \mathbb{R}^n$ can be obtained in polynomial time of the motif size.*

The reconstruction in condition 1.2(b) is stronger than the completeness in 1.2(a) because a complete invariant can be too complicated with no explicit inversion to an original object. For example, a DNA code is practically used for identifying humans, but cannot (yet) grow a genetic replica of a living person.

Conditions 1.2(a,b) become practically meaningful only with a Lipschitz continuous metric in condition 1.2(c) because any noise makes all real objects at least slightly different as in Fig. 1 (right). This discontinuity allowed anyone to *claim known materials as new* [13] by perturbing atomic positions, scaling up a minimal cell, and changing atomic types to make comparisons by symmetries, unit cells, and chemical compositions unreliable. As a result, many simulated crystals can be artificially generated, e.g. the report of "2.2 million new crystals – equivalent to nearly 800 years' worth of knowledge" from [26] was rebutted by experts [14,64].

The metric axioms are essential for recognizing isometric sets $S \simeq Q$ by checking if a complete invariant I satisfies $d(I(S), I(Q)) = 0$. If the triangle inequality in 1.2(c) fails with any positive error, outputs of k-means and DBSCAN clustering may be pre-determined for a non-metric and hence are not trustworthy [52]. Polynomial-time condition 1.2(d) makes Problem 1.2 notoriously hard, else one can design a complete infinite-size invariant by taking all isometric images of S.

An invariant I satisfying all the conditions above is similar to geographic coordinates that continuously parametrize the surface of Earth. Hence, Problem 1.2 is interpreted as geographic-style mapping of the *Crystal Isometry Space* $\mathrm{CIS}(\mathbb{R}^n; m)$ defined as the *moduli* space of all periodic sets with up to m points in a unit cell under isometry in $\mathbb{R}^n$. An invariant I can be considered a function on the union $\bigcup\limits_{m\geq 1}$ CIS$(\mathbb{R}^n; m)$ with values in a metric space, where all computations should be faster than in CIS$(\mathbb{R}^n; m)$, i.e. in polynomial time in m for a fixed dimension n.

Contributions. We extend the Pointwise Distance Distribution (PDD) [63] to stronger (also generically complete) invariants PDD$^{(h)}$ for higher orders $h > 1$ by keeping the Lipschitz continuity under bounded noise and polynomial-time computability for fixed n, h. The invariants PDD$^{\{2\}}$ distinguish all known examples $S \not\cong Q$ with PDD$(S) =$ PDD(Q) in $\mathbb{R}^3$ and experimentally confirm thousands of near-duplicates in the world's largest databases of periodic materials in section 6.

2 A review of open challenges in representations of periodic crystals

Problem 1.2 makes sense for many real objects (finite point sets, embedded graphs, surfaces or complexes in $\mathbb{R}^n$) under other practical equivalences (affine and projective transformations). The graph isomorphism problem [27] considers only conditions 1.2(a,d) without a continuous metric, which is needed for real lengths of edges. Since pairwise distances [9] distinguish all generic sets of m unordered points under isometry in $\mathbb{R}^n$ and the more recent complete invariants [67] continuously distinguish all finite sets under rigid motion in $\mathbb{R}^n$, we focus on periodic sets.

For $n = 1$, Theorem 4 in [28] justified complete invariants for periodic sequences given by rational angles of the unit circle (in the complex plane $\mathbb{C}$) by using 6-factor products of complex numbers. Since the circle (a period) was fixed, these invariants are discontinuous under perturbations. Indeed, the sequence $\mathbb{Z}$ of integers is infinitely close to $S = \{\varepsilon, 1, \dots, m\} + (m+1)\mathbb{Z} \subset \mathbb{R}$ for any small $\varepsilon > 0$, though their minimum periods 1 and $m + 1$ are arbitrarily different. The much simpler complete invariant of a periodic sequence $S = \{p_1, \dots, p_m\} + L\mathbb{Z} \subset \mathbb{R}$ with a period L, where $0 \leq p_1 < \dots < p_m < L$, is the list of inter-point distances $p_{i+1} - p_i$ (under cyclic permutations) for $i = 1, \dots, m$ and $p_{m+1} = p_1 + L$.

A continuous metric $d(S, Q)$ on these cyclic classes of distance lists was introduced in [38] but such a metric requires an expansion to the least common multiple of the sizes $|S|, |Q|$ of motifs and doesn't come with a polynomial-time invariant. The resulting brute force invariant for all periodic sequences S with motifs up to m points needs an expansion to at least 2^m points [24, Theorem 5(1)], which violates condition 1.2(d). Problem 1.2 remained open even in dimension $n = 1$.

A finite approach to measuring the similarity between periodic point sets is to compare their finite subsets within a box or a ball of a large but fixed cut-off radius. However, any periodic point set has many non-isometric finite subsets within differently positioned boxes or balls of the same size as in Fig. 1 (left).

Local clusters centered at all points in a motif M can be converted by Gaussian blurring into smooth functions [8], which can be decomposed in the infinite basis of spherical harmonics [56] and hence considered complete in the limit. [18] discusses challenges of choosing several parameters (blurring, approximation, interaction order), including a cut-off radius that can discontinuously change these clusters due to new neighbors outside a smaller cut-off. Even if this cut-off is smoothed out, a manually chosen value may not suffice or slow down computations [47, 51].

Atomic vibrations are natural to measure by deviations of atoms from their initial positions, but a sum of small deviations over infinitely many points can be infinite and also can give different values for different finite subsets. However, a maximum deviation of atoms is well-defined as the bottleneck distance between any sets via bijections between atoms, which can be displaced but cannot vanish.

Definition 2.1 (bottleneck distance d_B) *The* bottleneck distance $d_B(S,Q) = \inf\limits_{g:S\to Q} \sup\limits_{p\in S} |p-g(p)|$ *for any sets $S,Q\subset\mathbb{R}^n$ of the same cardinality is minimized for all bijections $g:S\to Q$ and maximized for all points $p\in S$.*

Here $|p-q|$ denotes Euclidean distance between points $p,q\in\mathbb{R}^n$. Though Definition 2.1 is impractical because of infinitely many bijections, d_B can be efficiently computed [21] for finite sets if $|p-q|$ is replaced with $L_\infty(p,q) = \max\limits_{i=1,\dots,n} |p_i-q_i|$.

If periodic point sets S,Q have different densities (motif size $|S|$ divided by the cell volume), then $d_B(S,Q)$ is infinite [63, Example 2.1]. Also, $d_B(S,Q)$ is discontinuous under perturbations of 2D lattices [37]whose *primitive* cells have the same minimum volume [63, Example 2.2]. Hence condition 1.2(c) of a Lipschitz continuous metric made Problem 1.2 exceptionally hard in the periodic case.

Definition 2.2 (*metrics* vs *pseudo-metrics*) *A distance d between objects under an equivalence relation $\sim$ is a* metric *if the following axioms hold:*
(1) coincidence: $d(S,Q)=0$ *if and only if $S\sim Q$;*
(2) symmetry: $d(S,Q)=d(Q,S)$ *for any objects S,Q;*
(3) triangle inequality: $d(S,Q)+d(Q,T)\geq d(S,T)$ *for any S,Q,T.*

If the coincidence axiom (1) is replaced with $(1')$ $d(S,S)=0$ *for any S, then non-equivalent $S\not\sim Q$ can have $d(S,Q)=0$, and d is called a* pseudo-metric.

Definition 2.2 guarantees positivity: $2d(S,Q) = d(S,Q)+d(Q,S)\geq d(S,S) = 0$. Many descriptors or invariants are compared by distances (such as Euclidean) that satisfy all metric axioms on descriptor values but define only pseudo-metrics on isometry classes due to the incompleteness of these invariants. If $d(S,Q)>0$, then $S\not\sim Q$ by $(1')$, so a fast pseudo-metric can distinguish between some but not all objects. Pseudo-metrics are weaker than metrics, e.g. the difference $||S|-|Q||$ of set sizes is a pseudo-metric not distinguishing any sets $S\not\cong Q$ of the same size.

Hence metrics satisfying all axioms (similar to complete invariants) are much more valuable than pseudo-metrics (similar to non-invariants or incomplete invariants). Any algorithm using an incomplete invariant I cannot predict different properties of a *false positive* pair of non-isometric sets $S\not\cong Q$ with $I(S)=I(Q)$.

That is why the *discriminative* problem should be solved first (at least in general position) by designing complete and Lipschitz continuous invariants before *generative* attempts can succeed. Any non-complete invariant I is not invertible in the sense that different sets $S\not\cong Q$ (false positives) can have $I(S)=I(Q)$.

Now we review recent continuous invariants in the periodic case. Continuous metrics on lattices under rigid motion are known for dimension $n=2$ [11,10], not yet for $n=3$ [36]. A generically complete and Lipschitz continuous invariant of periodic point sets $S\subset\mathbb{R}^3$ [20] is the sequence of *density functions* $\psi_k(S;t)$ measuring the fractional volume of k-fold intersections $\bigcup\limits_{p_1,\dots,p_k\in S} (\bar{B}(p_1;t)\cap\dots\cap\bar{B}(p_k;t)\cap U)$

for any $k \geq 1$, where U is a unit cell of S, and $\bar{B}(p;t)$ is the closed ball with a center $p \in S$ and a variable radius $t \geq 0$. The infinite sequence $\{\psi_k\}_{k=1}^{+\infty}$ allows only an approximate distance and turned out to be incomplete [3, Example 11], but was analytically described for all periodic sequences of intervals in $\mathbb{R}$ [4]. The *periodic merge tree* [19] is a continuous isometry invariant of periodic graphs with a slow interleaving pseudo-metric and a faster distance on simpler periodic 0-th barcodes. The invariant below solved a weaker version of Problem 1.2 for finite and periodic sets when completeness in (1.2a) is replaced with generic completeness.

Definition 2.3 (Pointwise Distance Distribution PDD) *Let $S \subset \mathbb{R}^n$ be any l-periodic point set with a motif M of m points. For any integer $k \geq 1$ and $p \in M$, let $d_1(p) \leq \cdots \leq d_k(p)$ be the list of Euclidean distances from p to its k nearest neighbors within the whole set S. These lists become rows of the $m \times k$ matrix $D(S, M; k)$. Any $c > 1$ identical rows are collapsed into a single row with the* weight *c/m, which is written in the extra first column. The resulting matrix* $\mathrm{PDD}(S;k)$ *of unordered rows with weights is called the* Pointwise Distance Distribution *[63].*

For finite sets, the PDD was studied under the name of a *local distribution of distances* [42]. The PDD can be considered a multiset of rows and a discrete probability *distribution* with normalized weights interpreted as probabilities.

If a unit cell of S is extended by a factor of c, then any point p in the original motif has c translationally equivalent copies in the extended motif M_c. Then $D(S, M_c; k)$ has c times more rows because each original row is expanded into c identical rows but the invariant $\mathrm{PDD}(S;k)$ is the same weighted distribution of rows, independent of an initial cell of S. The equality between weighted distributions is interpreted as a bijection between unordered sets respecting all weights. This equality is best checked not by considering all bijections but by a metric that vanishes only on equal distributions due to the first metric axiom. The PDD is Lipschitz continuous, computable (for a fixed dimension) in a near-linear time of k, m, and distinguishes all non-isometric sets in *general position* (away from a measure 0 subspace), see [63, Theorems 3.2, 4.3, 4.4, 5.1] and proofs in [65].

Definition 2.4 (homometric sets) *Finite or l-periodic sets $S, Q \subset \mathbb{R}^n$ are called* homometric *[48] if they have the same* Pair Distribution Function (PDF)*, which is a single distribution of all inter-point distances of S (without considering their periodic copies), equivalent to a powder diffraction pattern without a cut-off radius.*

The PDF is easily extractable from X-ray diffraction patterns and can be split into several distributions by fixing an atomic type (chemical element), say by listing average distances from all (say) carbon atoms to their neighbors in the full crystal. The PDD does this splitting by geometry (all identical distances to neighbors) and is stronger than the PDF even for 1-dimensional periodic sequences in Fig. 2.

Almost any perturbation, as in Fig. 1 (right), can split every inter-point distance (say) d into many $d_1, \ldots, d_c$, which are all close to d but are not copies of each other because the initial minimal cell was scaled by the factor c. One attempt to resolve this discontinuity was to blur each distance by a Gaussian deviation and a smoothed PDF as a normalized sum of Gaussians around all distance values.

Discretizing the smoothed PDF for comparisons reduces its strength and creates the counter-intuitive pipeline: a discrete set $S \to$ a smoothed PDF $\to$ a discrete sample of $\mathrm{PDF}(S)$. The discontinuity can be resolved by continuous metrics [33, 60] on PDDs interpreted as a probability distribution of rows of k distances.

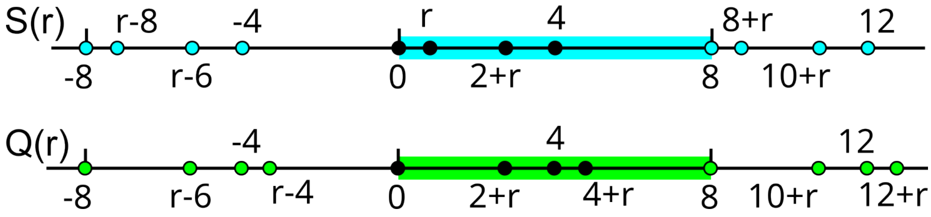
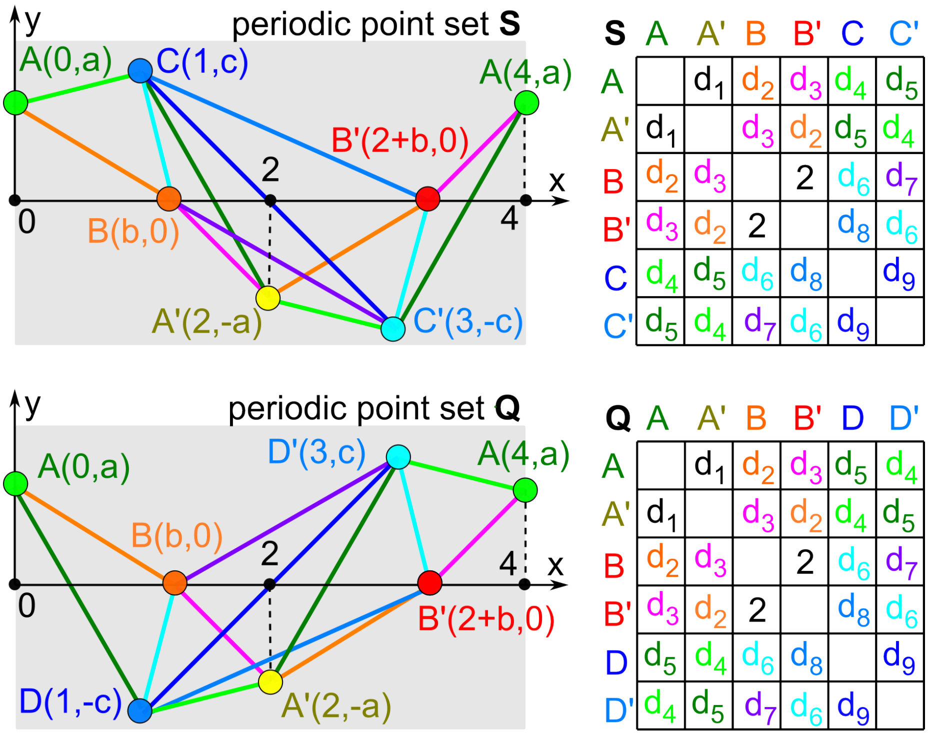


Fig. 2 For any $0 < r \leq 1$, the homometric sets $S(r) = \{0, r, 2+r, 4\} + 8\mathbb{Z} \not\cong Q(r) = \{0, r, 2+r, 4\} + 8\mathbb{Z}$ have identical PDFs from Definition 2.4 but different PDDs whose first columns we write as unordered sets: $\text{PDD}(S(r); 1) = \{r, r, 2-r, 2-r\} \neq \text{PDD}(Q(r); 1) = \{r, r, 2-r, 2+r\}$.



S	A	A'	B	B'	C	C'
A		d_1	d_2	d_3	d_4	d_5
A'	d_1		d_3	d_2	d_5	d_4
B	d_2	d_3		2	d_6	d_7
B'	d_3	d_2	2		d_8	d_6
C	d_4	d_5	d_6	d_8		d_9
C'	d_5	d_4	d_7	d_6	d_9	

Q	A	A'	B	B'	D	D'
A		d_1	d_2	d_3	d_5	d_4
A'	d_1		d_3	d_2	d_4	d_5
B	d_2	d_3		2	d_6	d_7
B'	d_3	d_2	2		d_8	d_6
D	d_5	d_4	d_6	d_8		d_9
D'	d_4	d_5	d_7	d_6	d_9	

Fig. 3 The sets S, Q are 1-periodic in the x-axis with period 4, e.g. A denotes both $(0, a)$, $(4, a)$. **Right**: distances between closest points from classes modulo shifts by 4 in x. Then $\text{PDD}(S; k) = \text{PDD}(Q; k)$ by Example 2.5 but $\text{PDD}^{\{2\}}(S; 1) \neq \text{PDD}^{\{2\}}(Q; 1)$ by Example 3.4.

Example 2.5 (sets with equal PDDs) *The 1-periodic sets $S \not\cong Q$ in [50, Fig. 4] were designed to fail all iterations of the Weisfeiler-Leman test [58]. Fig. 3 shows their 2D versions with period 4 in the x-axis and free parameters $a, b, c > 0$.*

The distances in Fig. 3 (right) are for the closest representatives of 6 points.
$d_1 = 2\sqrt{a^2+1}$, $d_2 = \sqrt{a^2+b^2}$, $d_3 = \sqrt{a^2+(2-b)^2}$,
$d_4 = \sqrt{1+(a-c)^2}$, $d_5 = \sqrt{1+(a+c)^2}$, $d_6 = \sqrt{(1-b)^2+c^2}$,
$d_7 = \sqrt{(3-b)^2+c^2}$, $d_8 = \sqrt{(1+b)^2+c^2}$, $d_9 = 2\sqrt{c^2+1}$.

Then $\mathrm{PDD}(S;k) = \mathrm{PDD}(Q;k)$ *because the equalities between distances (shown in the same color) in Fig. 3 (right) hold after adding any periodic translation, so if* $d_1 = d_2$ *then* $\sqrt{d_1^2 + (4n)^2} = \sqrt{d_2^2 + (4n)^2}$ *for any* $n \in \mathbb{Z}$.

Simpler non-isometric finite sets in $\mathbb{R}^3$ with equal PDDs were distinguished by stronger invariants in [67], which extended PDD by recording distances to subsets of more than one point. In the periodic case, pairs of points behave discontinuously under cell extensions in Fig. 1. Doubling a motif M of m points leads to $(2m)^2$ pairs including new distant neighbors from adjacent cells. This obstacle motivated a 'pointwise' approach to both finite and periodic sets in the next section.

Another 'pointwise' *isoset* [2] was proved to be complete for all periodic point sets in any $\mathbb{R}^n$. A Lipschitz continuous metric on isosets was only approximated in polynomial time [6, 41], but condition (1.2d) requires an exact computation.

3 The new isometry invariants of finite and periodic sets of points

This section extends the PDD to higher order $h > 1$ in Definition 3.1 motivated by pairs of non-isometric sets $S \not\cong Q$ with $\mathrm{PDD}(S;k) = \mathrm{PDD}(Q;k)$ in Example 2.5. Definition 3.1 makes sense for a finite set $S = M$ in any metric space.

Definition 3.1 (higher order $\mathrm{PDD}^{\{h\}}(S;k)$) *Let* $S \subset \mathbb{R}^n$ *be any* l*-periodic point set with a motif* M *of* m *points. Fix a point* $p \in M$ *and integers* $h, k \geq 1$. *Consider any* h *distinct points* $p_1, \dots, p_h \in S \setminus \{p\}$ *and the* h*-*order average $\dfrac{2}{h(h+1)} \sum\limits_{0 \leq i < j \leq h} |p_i - p_j|$ *of pairwise distances between the points* $p = p_0, p_1, \dots, p_h$. *Let* $a(p;h,1) \leq \dots \leq a(p;h,k)$ *be the list of* k *smallest averages for the fixed point* p *and variable points* $p_1, \dots, p_h \in S \setminus \{p\}$. *These lists become rows of the* $m \times k$ *matrix* $D(S,M;h,k)$, *where we can collapse any* $c > 1$ *equal rows to one row with the* weight c/m *written in the extra first column. The final matrix of unordered rows with weights is the* h*-*order Pointwise Distance Distribution $\mathrm{PDD}^{\{h\}}(S;k)$.

Lemma 3.3 will prove that $\mathrm{PDD}^{\{h\}}(S;k)$ is independent of a motif M to justify the notation without M. In Definition 3.1, we can keep m rows of average sums with equal weights $1/m$. The matrices $\mathrm{PDD}^{\{1\}}, \dots \mathrm{PDD}^{\{h\}}$ can be concatenated into a single matrix $\mathrm{PDD}^{(h)}(S;k)$ of m unordered rows and kh ordered columns.

Example 3.2 ($\mathrm{PDD}^{(2)}$ for the sequences in Fig. 2) *The sum* $\sum\limits_{0 \leq i < j \leq 2} |p_i - p_j|$ *is the perimeter of the triangle on the points* $p_0 \in M$ *and* $p_1, p_2 \in S$. *The row of a point* $p \in M$ *in* $\mathrm{PDD}^{(2)}(S;k)$ *consists of the* k *shortest distances followed by* k *smallest perimeters (divided by* 3*) of triangles at* p. *In Fig. 2, the point* $p_0 = 0$ *in the motif of* $S(r) = \{0, r, 2+r, 4\} + 8\mathbb{Z}$ *has the nearest neighbors* $p_1 = r$, $p_2 = 2+r$ *at the distances* $r, 2+r$, *and two smallest averaged perimeters* $2(2+r)/3, 8/3$. *The point* $p_0 = 0$ *in* $Q(r) = \{0, r, 2+r, 4\} + 8\mathbb{Z}$ *has the nearest neighbors at the distances* $2+r, 4-r$, *and two smallest averaged perimeters* $\frac{8}{3}, \frac{8}{3}$. *Then* $\mathrm{PDD}^{(2)}(S(r);2) =$
$$\left(\begin{array}{cc|cc} r & 2+r & \frac{2(2+r)}{3} & \frac{8}{3} \\ r & 2 & \frac{2(2+r)}{3} & \frac{2(4-r)}{3} \\ 2-r & 2 & \frac{2(2+r)}{3} & \frac{2(4-r)}{3} \\ 2-r & 4-r & \frac{2(4-r)}{3} & \frac{8}{3} \end{array}\right) \neq \mathrm{PDD}^{(2)}(Q(r);2) = \left(\begin{array}{cc|cc} 2+r & 4-r & \frac{8}{3} & \frac{8}{3} \\ 2-r & 2+r & \frac{4}{3} & \frac{8}{3} \\ r & 2-r & \frac{4}{3} & \frac{8}{3} \\ r & 2 & \frac{4}{3} & \frac{8}{3} \end{array}\right),$$
where all rows have equal weights $\frac{1}{4}$, *so we have skipped these weights for brevity.*

The factor $\dfrac{2}{h(h+1)}$ was chosen to guarantee the Lipschitz continuity with $\lambda = 2$ in (1.2c). Examples 3.4, 4.2 show that $\mathrm{PDD}^{\{2\}}$ distinguishes all known homometric sets for $n = 2, 3$, which have identical PDDs. Any increase in k adds extra columns with larger values to $\mathrm{PDD}^{\{h\}}(S; k)$ without changing any previous values. So the number k is considered a degree of approximation, not a parameter like a cut-off radius whose changes substantially affect local atomic clouds.

Lemma 3.3 proves the invariance of $\mathrm{PDD}^{(h)}$ under isometry in $\mathbb{R}^n$ and under changes of a cell. If k is greater than the number $\binom{m-1}{h-1}$ of h-tuples with a fixed $p \in S$, we set all non-existing sums in Definition 3.1 to the largest existing value.

Lemma 3.3 (invariance of $\mathrm{PDD}^{\{h\}}(S; k)$) *For any integers $h, k \geq 1 \leq l \leq n$ and any finite unordered set S in a metric space or any l-periodic point set $S \subset \mathbb{R}^n$, the higher-order $\mathrm{PDD}^{\{h\}}(S; k)$ from Definition 3.1 is an isometry invariant of S.*

Proof First, for any l-periodic point set $S \subset \mathbb{R}^n$, we show that scaling up a unit cell U to a non-primitive cell keeps PDD invariant. It suffices to scale up U by a factor c, say along the first basis vector v_1 of U, then the number m of motif points of S is multiplied by c. Then the matrix $D(S, S \cap (cU); h, k)$ consisting of k smallest average sums of pairwise distances between $h + 1$ points in Definition 3.1 has the larger size $cm \times k$ in comparison with the original $m \times k$ matrix $D(S, S \cap U; h, k)$ but each row is repeated c times for the shifted points $p + iv_1$, where p is any point from the original motif $M = S \cap U$ of the l-periodic set S, for $i = 0, \ldots, c - 1$.

Second, we show that the matrix $D(S, S \cap U; h, k)$ and hence $\mathrm{PDD}^{\{h\}}(S; k)$ is independent of a primitive cell U. Let U, V be primitive cells of any l-periodic set $S \subset \mathbb{R}^n$ with a lattice Λ. Any point $q \in S \cap V$ can be translated by a vector of Λ to a point $p \in S \cap U$ and vice versa. These translations preserve distances and establish a bijection between the motifs $S \cap U \leftrightarrow S \cap V$, and a bijection between all rows of the matrices $D(S, S \cap U; h, k) \leftrightarrow D(S, S \cap V; h, k)$.

Third, we prove that $\mathrm{PDD}^{\{h\}}(S; k)$ is preserved under any isometry $f : S \to Q$ of l-periodic point sets. Any primitive cell U of S is bijectively mapped by f to the unit cell $f(U)$ of Q, which should be also primitive. Indeed, if Q is preserved by a translation along a vector v that doesn't have all integer coefficients in the basis of $f(U)$, then $S = f^{-1}(Q)$ is preserved by the translation along $f^{-1}(v)$, which doesn't have all integer coefficients in the basis of U, so U was non-primitive. Since U and $f(U)$ have the same number of points from S and $Q = f(S)$, the isometry f gives a bijection between the motifs $S \cap U \leftrightarrow Q \cap f(U)$.

For any discrete sets S, Q, the k smallest average sums of all distances between any point $p \in S \cap U$ and $p_1, \ldots, p_h \in S$, equal the same sums for $f(p) \in Q \cap f(U)$ and $f(p_1), \ldots, f(p_h) \in Q$, respectively. These coincidences imply that $\mathrm{PDD}(S; k_1, \ldots, k_h) = \mathrm{PDD}^{\{h\}}(Q; k_1, \ldots, k_h)$ up to a permutation of rows. □

Example 3.4 ($\mathrm{PDD}^{\{2\}}$ distinguishes S, Q in Example 2.5) *We start with singular cases when S, Q are identical. If $c = 0$, then $C = D$, $C' = D'$, so S, Q are identical in Fig. 3. If $b \in \{0, 1, 2\}$, then the periodic shifts of $B \cup B'$ (hence S, Q) become mirror images with respect to the vertical line $x = 2$. We now assume that $1 < b < 2$. Then $d_2 > d_3$, $d_5 > \max\{d_4, d_6\}$, and $\min\{d_7, d_8, d_9\} > d_6$.*

The set S in Fig. 3 has a motif of 6 points, which generate isometric triangles $\triangle ABC \simeq \triangle A'B'C'$ with the perimeter $d_2 + d_4 + d_6$, see details in Example 2.5.

The other potentially smaller perimeters of triangles on points of S are $d_3+d_5+d_6$, $d_3+d_4+d_7$. The smallest perimeter for S is the minimum of these sums. The smallest perimeter for Q is $\min\{d_2+d_4+d_5,\ d_2+d_5+d_6,\ d_3+d_4+d_6\}$.

If $t = d_2+d_4+d_6$ equals one of the last sums, one of the following cases holds: if $d_2 = d_3$ then $b = 1$, if $d_4 = d_5$ then $c = 0$, if $d_6 = d_7$ then $b = 2$ or 0, so $S \simeq Q$.

If $t = d_3+d_5+d_6$ is a minimal perimeter for S, then t cannot equal any of the three sums for Q. Indeed, if $t = d_2+d_5+d_6$ then $d_2 = d_3$. If $t = d_3+d_4+d_6$ then $d_4 = d_5$. The minimality of the sum t for the set S means that $d_3+d_6 < d_2+d_4$, so $t = d_3+d_5+d_6$ cannot equal $d_2+d_4+d_5$ for Q.

If $t = d_3+d_4+d_7$ is a minimal perimeter for S, then t cannot equal any of the three sums for Q. Indeed, if $t = d_3+d_4+d_6$ then $d_6 = d_7$. The minimality of t for S means that $d_3+d_7 < d_2+d_6 < d_2+d_5$, so $t = d_3+d_4+d_7 < d_2+d_4+d_5$ for Q. Similarly, if $d_4+d_7 < d_5+d_6$ then $t = d_3+d_4+d_7 < d_3+d_5+d_6 < d_2+d_5+d_6$.

In all these cases, S, Q become isometric. Hence the smallest perimeters in $\mathrm{PDD}^{\{2\}}$ *for $k = 1$ distinguish all pairs of the homometric sets S, Q. The same conclusion holds for more general sets obtained from S, Q by periodic translations in other directions (along the y-axis or even in any $\mathbb{R}^n$), see [50, Fig. 10], when extra periods are large and don't affect any triangles with the smallest perimeters.*

The rows of $\mathrm{PDD}^{\{h\}}(S;k)$ are unordered to guarantee the continuity under perturbations, though we can lexicographically order the rows for convenience. Recall that $u = (u_1,\dots,u_n)$ is *lexicographically* smaller than $v = (v_1,\dots,v_n)$ in $\mathbb{R}^n$ (written $u < v$) if $u_i = v_i$ for $i = 1,\dots,k$ and $u_{k+1} < v_{k+1}$ for some $k < n$.

We can convert any $\mathrm{PDD}^{\{h\}}$ into a fixed-size matrix, which can be flattened into a vector for easy comparisons, while keeping the continuity and almost all invariant data. Any distribution of m unordered values can be reconstructed from its m moments defined below. When all weights w_i are rational as in our case, the distribution can be expanded to equal-weighted values $a_1,\dots,a_m$. The m moments can recover all $a_1,\dots,a_m$ as roots of a polynomial of degree m whose coefficients are expressed via the m moments [40]. For example, any reals a, b are the roots of the quadratic polynomial $x^2-(a+b)x+ab$, where $ab = \frac{1}{2}((a+b)^2-(a^2+b^2))$.

Let A be any unordered set of real numbers $a_1,\dots,a_m$ with weights $w_1,\dots,w_m$, respectively, such that $\sum\limits_{i=1}^{m} w_i = 1$. For any integer $t \geq 1$, the t-th *moment* [34, section 2.7] is $\mu_t(A) = \sqrt[t]{m^{1-t}\sum\limits_{i=1}^{m} w_i a_i^t}$, so $\mu_1(A) = \sum\limits_{i=1}^{m} w_i a_i$ is the usual average. For $t \geq 2$, we normalize the sum (before taking the t-th root) by the factor $m^{(1/t)-1}$ to prove continuity of all moments with the Lipschitz constant $\lambda = 2$.

Definition 3.5 (the t-moments matrix $\mu^{(t)}[\mathrm{PDD}^{\{h\}}]$) *Fix any integers $h, k, t \geq 1 \leq l \leq n$, and a finite or l-periodic point set $S \subset \mathbb{R}^n$. For every column A of the matrix* $\mathrm{PDD}^{\{h\}}(S;k)$ *from Definition 3.1, which consists of unordered numbers $a_1,\dots,a_m$ with weights, write the new column $(\mu_1(A),\dots,\mu_t(A))$. The resulting t-*moments *matrix of sizes $t\times k$ is denoted by $\mu^{(t)}[\mathrm{PDD}^{\{h\}}(S;k)]$. For $t = h = 1$, the $1\times k$ matrix $\mu^{(1)}[\mathrm{PDD}(S;k)]$ was called the vector of* Average Minimum Distances *[66] and was also denoted by* $\mathrm{AMD}(S;k) = (\mathrm{AMD}_1,\dots,\mathrm{AMD}_k)$.

The matrix $\mu^{(t)}[\mathrm{PDD}^{\{h\}}(S;k)]$ has t ordered rows and k ordered columns but is a bit weaker than the original distribution $\mathrm{PDD}^{\{h\}}(S;k)$ with the same parameters h,k, because each column is reconstructable from its moments for $t\geq m$ only up to a permutation. However, to faster filter distant crystals, we can flatten any matrix $\mu^{(t)}[\mathrm{PDD}(S;k)]$ with indexed entries to a vector of tk coordinates.

For a finite set $S\subset\mathbb{R}$, a simple complete invariant under translations is the ordered sequence of inter-point distances. However, a naive extension to periodic sets is discontinuous, e.g. $\mathbb{Z}$ is ε-close to $\{0,1+\varepsilon\}+2\mathbb{Z}$ but their periods 1 and 2 are not close. Definition 3.6 introduces a distribution whose completeness and Lipschitz continuity for $n=1$ will be proved by Theorem 3.7 and Lemma 4.6.

Definition 3.6 (*Pointwise Shift Distribution* PSD) *For any periodic point set (sequence) $S\subset\mathbb{R}$ with a motif M of m points, write down distances from each $p\in M$ to its k nearest neighbors $q>p$ in increasing order in a row of an $m\times k$-matrix. Collapse any $c>1$ equal rows to one row with the weight c/m in an extra first column. The resulting matrix $\mathrm{PSD}(S;k)$ is the* Pointwise Shift Distribution *and makes sense for any finite set $S=M\subset\mathbb{R}$ of $m\geq k+1$ unordered points.*

$\mathrm{PSD}(S;k)$ differs from $\mathrm{PDD}(S;k)$ because we consider only neighbors q to the right of a point p in the line $\mathbb{R}$, so PSD consists of shifts (distances to the right).

Theorem 3.7 (completeness for $n=1$) ***(a)*** *A finite set $M\subset\mathbb{R}$ of m unordered points is reconstructable from $\mathrm{PDD}(M;m-1)$ uniquely under isometry.* ***(b)*** *For all periodic sets $S\subset\mathbb{R}$ with m points in a motif, $\mathrm{PSD}(S;m)$ is a complete invariant under rigid motion and can be computed in time $O(m^2)$.*

Proof **(a)** For a finite set $S\subset\mathbb{R}$ of m unordered points, we prove that S can be reconstructed from $\mathrm{PSD}(S;m-1)$ uniquely under isometry. Indeed, the number m can be assumed to be known as one plus the number of columns in $\mathrm{PSD}(S;m-1)$. Find a row R whose last distance d is maximal in $\mathrm{PSD}(S;m-1)$. This maximal distance is achieved exactly for two most distant points of S, else $\mathrm{PSD}(S;m-1)$ is unrealizable by m distinct points. These two most distant points can be fixed at the positions 0 and d up to isometry of $\mathbb{R}$. All other $m-2$ points of S are uniquely determined by the first $m-2$ distances in the row R, which should be distinct.

(b) The time to compute $\mathrm{PSD}(S;k)$ is linear in the size m of a motif and in the number k of neighbors. Let S have a motif M of m points $0=p_0<p_1<\cdots<p_{m-1}<p_m$ and period $L=p_m-p_0$. For any point $p_i\in M$, the distance to its k-th neighbor is $p_{i+k-mN}-p_i+LN$, where $N=[k/m]$ is the integer part and $p_j=p_{j-m}+L$ for $m\leq j<2m$. So all k neighbors of p_i are computed in linear time in both k,m, hence the total time over m points of M is quadratic in m.

Now we prove that any periodic point set $S\subset\mathbb{R}$ can be reconstructed (uniquely under translation) from any row $a_1<\cdots<a_{m-1}<a_m$ of $\mathrm{PSD}(S;m)$ by writing the points of a motif as $p_k=a_{k+1}-a_1$ for $k=0,\dots,m-1$, where $p_0=0$, and setting the period of S to d_m. The number m is given as the number of columns of $\mathrm{PSD}(S;m)$. The completeness can be stated as follows: any periodic sequences $S,Q\subset\mathbb{R}$ whose motifs have at most m points are related by translation if and only if $\mathrm{PSD}(S;m)=\mathrm{PSD}(Q;m)$ as weighted distributions of unordered rows. □

The invariant $\mathrm{PSD}(S;k)$ can be enhanced to a complete invariant under isometry (including reflections) in $\mathbb{R}$ as follows. Let $\bar{S}$ be the mirror image of S under reflection $x \mapsto -x$. In any row $a_1 < \cdots < a_k$ of $\mathrm{PSD}(S;k)$ for $k \geq m$, we can use the m-th distance a_m equal to the period L to write the corresponding row

$$L - a_{m-1} < \cdots < L - a_1 < 2L - a_{m-1}$$

in the matrix $\mathrm{PSD}(\bar{S};k)$. Then any periodic sequences S, Q are related by isometry in $\mathbb{R}$ if and only if $\mathrm{PSD}(S;m) = \mathrm{PSD}(Q;m)$ or $\mathrm{PSD}(\bar{S};m) = \mathrm{PSD}(Q;m)$.

Theorem 3.7 with the Lipschitz continuity in Lemma 4.6 will show that the PSD solves Problem 1.2 for all periodic sets (under rigid motion) for $n = 1$.

The generic completeness of $\mathrm{PDD}(S;k)$ (with a motif size $|S|$ and a lattice of S) in [63, Theorem 4.4] and Examples 3.4, 4.2 motivate the following conjecture.

Conjecture 3.8 (completeness of $\mathrm{PDD}^{(h)}$ under isometry in $\mathbb{R}^h$) *For $h \geq 1$, any periodic point set $S \subset \mathbb{R}^h$ can be reconstructed (uniquely under isometry) from the invariant $\mathrm{PDD}^{(h)}(S;k)$ for a sufficiently large k in Definition 3.1.*

4 Lipschitz continuous metrics on higher-order invariants

This section introduces metrics on $\mathrm{PDD}^{\{h\}}$ invariants and proves their Lipschitz continuity. Any vectors $u, v \in \mathbb{R}^m$ of distances or their average sums can be compared by the *Minkowski* metric $L_q(u,v) = (\sum\limits_{i=1}^{m} |u_i - v_i|^q)^{1/q}$ for $q \in [1, +\infty)$ and $L_\infty(u,v) = \max\limits_{i=1,\dots,m} |u_i - v_i|$ in the limit case $q = +\infty$. The *Root Mean Square* metric $\mathrm{RMS}(u,v) = \dfrac{L_2(u,v)}{\sqrt{m}}$ is the Euclidean metric normalized by the (square root of the) number m of coordinates. These metrics L_q and RMS controllably change under perturbations of distances and will play the role of a 'ground' metric d to compare unordered distributions $\mathrm{PDD}^{\{h\}}$ by the EMD metric below.

Definition 4.1 (Earth Mover's Distance EMD **[53])** ***(a)*** *Let X be a space with a ground metric d. Any unordered set $\{(R_i, w_i)\}_{i=1}^m$ of objects $R_i \in X$ with weights $w_i > 0$ such that $\sum\limits_{i=1}^{m} w_i = 1$ is called a (normalized)* weighted distribution*. For any such weighted distributions $A = \{(R_i(A), w_i(A))\}_{i=1}^{m(A)}$ and $B = \{(R_j(B), w_j(B))\}_{j=1}^{m(B)}$, the* Earth Mover's Distance *is defined as*

$$\mathrm{EMD}(A,B) = \min_{f_{ij} \in [0,1]} \sum_{i=1}^{m(A)} \sum_{j=1}^{m(B)} f_{ij} d(R_i(A), R_j(B))$$

subject to $\sum\limits_{i=1}^{m(A)} f_{ij} \leq w_j(B)$, $\sum\limits_{j=1}^{m(B)} f_{ij} \leq w_i(A)$, and $\sum\limits_{i=1}^{m(A)} \sum\limits_{j=1}^{m(B)} f_{ij} = 1$.

(b) *For any real $q \in [1, +\infty]$, any integers $n, h, k \geq 1$, and any periodic point sets $S, Q \subset \mathbb{R}^n$, the distance $\mathrm{EMD}_q^{\{h\}}[k](S,Q)$ is the* EMD *from part (a) between the distributions $\mathrm{PDD}^{\{h\}}(S;k)$ and $\mathrm{PDD}^{\{h\}}(Q;k)$ with the ground metric L_q. Define the distance $\mathrm{EMD}_q^{(h)}[k](S,Q) = \max\limits_{i=1,\dots,h} \{\mathrm{EMD}_q^{\{i\}}[k](S,Q)\}$. The notation* EMD *without a subscript q is used for the (default) ground metric* RMS *instead of L_q.*

Our experiments use RMS or the Minkowski metric L_∞, because these ground metrics will give the Lipschitz constant 2 for the EMD on $\mathrm{PDD}^{\{h\}}$ in Theorem 4.3. Definition 4.1(b) introduced the distance $\mathrm{EMD}_q^{(h)}[k](S,Q)$ as the maximum of EMDs over orders $1,\dots,h$ to keep the Lipschitz constant small. This maximum can be replaced with a sum or another metric transform, see [17, section 4.1].

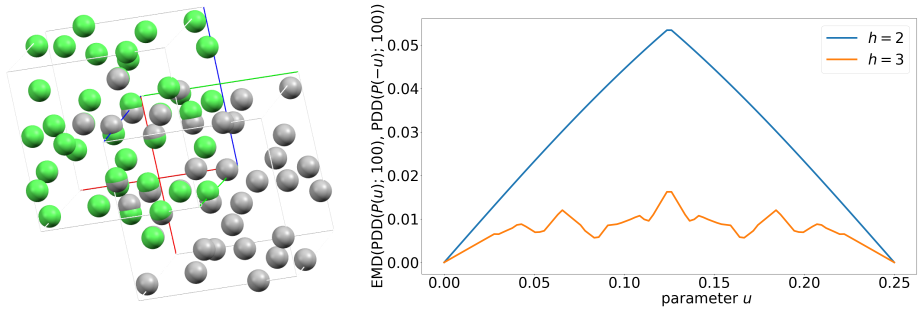


Fig. 4 **Left**: a comparison of Pauling's crystals $P(\pm u)$ for $u = 0.03$ [49], by COMPACK [15], which aligns subsets of 15 atoms. The atoms from different $P(\pm 0.03)$ are shown in green and gray. **Right**: EMD_∞ from Definition 4.1(b) is between $\mathrm{PDD}^{\{h\}}$ for $k = 100$ and Pauling's crystals $P(\pm u)$, which depend on $u \in [0, 0.25]$ and are identical at the boundary values.

Example 4.2 ($\mathrm{PDD}^{\{2\}}$ **distinguishes Pauling's crystals**) *Fig. 4 (left) shows a pair of overlaid Pauling's crystals* $P(\pm 0.03)$ *with 24 atoms in a cubic cell [49]. The importance of* $\mathrm{PDD}^{\{2\}}$ *in comparison with* PDD *is demonstrated by the infinite series of periodic sets* $P(\pm u) \subset \mathbb{R}^3$*, which have the same* $\mathrm{PDD}(P(u);k) = \mathrm{PDD}(P(-u);k)$ *for all parameters* $u \in (0, 0.25)$ *and* $k \geq 1$ *but are distinguished by* $\mathrm{PDD}^{\{h\}}(S;100)$ *due to* $\mathrm{EMD}_\infty^{\{h\}}[100] > 0$ *for* $h = 2, 3$ *in Fig. 4 (right).*

For any discrete set $S \subset \mathbb{R}^n$, the *packing radius* $r(S)$ is the minimum half-distance between any points of S. Recall the brief notation from Definition 4.1(b):

$$\mathrm{EMD}_q^{\{h\}}[k](S,Q) = \mathrm{EMD}_q\big(\,\mathrm{PDD}^{\{h\}}(S;k),\ \mathrm{PDD}^{\{h\}}(Q;k)\,\big).$$

Theorem 4.3 (Lipschitz continuity of $\mathrm{PDD}^{\{h\}}$**)** *Fix integers* $h, k \geq 1 \leq l \leq n$*. Let* Q *be a finite or an* l*-periodic point set obtained from a finite or an* l*-periodic point set* $S \subset \mathbb{R}^n$*, respectively, by perturbing every point of* S *up to a Euclidean distance* $\varepsilon \in [0, r(S))$*. Then* $\mathrm{EMD}_q^{\{h\}}[k](S,Q) \leq 2\varepsilon\sqrt[q]{k}$*, where* $\sqrt[q]{k} = 1$ *for* $q = +\infty$*, and* $\mathrm{EMD}\big(\,\mathrm{PDD}^{\{h\}}(S;k),\ \mathrm{PDD}^{\{h\}}(Q;k)\,\big) \leq 2\varepsilon$*, where the ground metric is* RMS.

Fig. 5 shows how $\mathrm{EMD}_\infty^{\{2\}}[100]$ continuously changes under perturbations.

Lemma 4.4 (perturbation of an ordered vector) *Let* $v_1 \leq \cdots \leq v_k$ *be a vector* v *of ordered real numbers. For some* $\varepsilon \geq 0$*, let a map* g *perturb each* a_i *to* $\tilde{v}_i = g(v_i)$ *so that* $|v_i - \tilde{v}_i| \leq \varepsilon$ *for* $i = 1,\dots,k$*. Let* $\tilde{v}$ *be the vector obtained by putting* $\tilde{v}_1,\dots,\tilde{v}_k$ *in increasing order. Then* $L_\infty(v,\tilde{v}) \leq \varepsilon$*.*

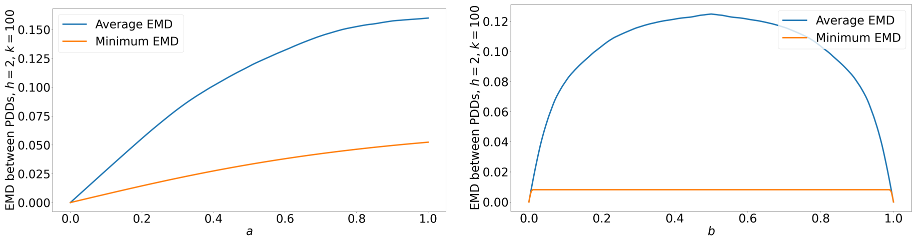


Fig. 5 The distance $\text{EMD}_\infty^{\{2\}}[100]$ between the 1-periodic sets S, Q in Fig. 3, which have identical PDDs. The average and minimum of $\text{EMD}_\infty^{\{2\}}[100]$ were computed for uniformly sampled parameters a, b, c from Example 3.4. These sets S, Q are isometric for $b \in \{0, 1\}$ but $\text{EMD}_\infty^{\{2\}}[100] > 0$ for $0 < b < 1$ experimentally confirms that $S \not\cong Q$, see Example 3.4.

Proof It suffices to prove that the i-th number u_i in the ordered vector $\tilde{v}$ is ε-close to the i-th number v_i in the original vector v, so $v_i - \varepsilon \leq u_i \leq v_i + \varepsilon$ for $i = 1, \ldots, k$. Assume by contradiction that $u_i < v_i - \varepsilon$. Since every component of v was perturbed by at most ε, the i numbers $u_1 \leq \cdots \leq u_i < v_i - \varepsilon$ can be obtained only as perturbations of components from v that are strictly less than v_i. However, the ordered vector A has at most $i-1$ numbers that are less v_i. This contradiction proves that $u_i \geq v_i - \varepsilon$. A similar argument proves that $u_i \leq v_i + \varepsilon$. □

Lemma 4.5 (upper bound of EMD**)** *Consider any weighted distributions* $A = \{(R_i(A), w_i)\}_{i=1}^m$ *and* $B = \{(R_i(B), w_i)\}_{i=1}^m$ *of matched objects with equal weights and ground distances* $d(R_i(A), R_i(B)) \leq \varepsilon$ *for* $i = 1, \ldots, m$. *Then* $\text{EMD}(A, B) \leq \varepsilon$.

Proof Define the flows f_{ij} from the m objects of A to the corresponding m objects of B by setting $f_{ii} = \dfrac{1}{m}$ and $f_{ij} = 0$ for $i \neq j$, $i, j = 1, \ldots, m$. Then

$$\text{EMD}(A, B) \leq \sum_{i,j=1}^{m} f_{ij} d(R_i(S), R_j(Q)) = \frac{1}{m} \sum_{i=1}^{m} d(R_i(S), R_i(Q)) \leq \frac{1}{m} \sum_{i=1}^{m} \varepsilon = \varepsilon$$

since $\text{EMD}(A, B)$ is the minimum over all $f_{ij} \in [0, 1]$, see Definition 4.1(a). □

Proof of Theorem 4.3. In the periodic case, if the perturbation satisfies $\varepsilon < r(S)$, [20, Lemma 4.1] and [6, Lemma 4.8] proved that S, Q have a common lattice with a unit cell U such that $S = \Lambda + (S \cap U)$ and $Q = \Lambda + (Q \cap U)$. Then S, Q share a unit cell U and have the same number $m = m(S) = m(Q)$ of points in U. The arguments below also work for any finite sets S, Q in a large enough U.

Expand $\text{PDD}^{\{h\}}$ of both S, Q to the matrices with m equally weighted rows. Reorder all m rows of $D(S, S \cap U; h, k)$ and $D(Q, Q \cap U; h, k)$ according to the bijection $g : S \cap U \to Q \cap U$. Since any $p \in S$ is perturbed up to ε, any distance $L_q(p, q)$ between $p, q \in S$ and hence any average sum a from Definition 3.1 changes by at most 2ε due to the triangle inequality for the Minkowski metric L_q.

Some of the average sums from the original matrix $D(S, S \cap U; h, k)$ can increase up to 2ε and will be outside the k smallest average sums in the new matrix $D(S, S \cap U; h, k)$. In this case, for each row $i = 1, \ldots, m$, let $k_i \geq k$ be the maximum index such that the k_i-th smallest average sum (of pairwise distances between $h + 1$ points including $p_i \in S$) for S is at most 2ε plus the largest average sum on $h+1$ points from the original matrix $D(S, S\cap U; h,k)$ in the i-th row. Set $k'=\max\limits_{i=1,\dots,m} k_i \geq k$. Then the i-th row of $D(Q, Q\cap U; h, k')$ is obtained from the i-th row of $D(S, S\cap U; h, k')$ of k' numbers by changing every value by at most 2ε, putting them in increasing order, and taking the first $k\leq k'$ smallest values.

For each $i=1,\dots,m$, Lemma 4.4 implies that the corresponding components in the extended i-th rows of k' numbers in $D(S, S\cap U; h, k')$ and $D(Q, Q\cap U; h, k')$ differ by at most 2ε. The same conclusion holds for the shorter i-th rows $R_i(S)$ and $R_i(Q)$ of k values in the matrices $D(S, S\cap U; h,k)$ and $D(Q, Q\cap U; h,k)$, respectively. Then $L_q(R_i(S), R_i(Q)) \leq \sqrt[q]{k(2\varepsilon)^q} = 2\varepsilon\sqrt[q]{k}$. The Euclidean L_2 normalized with the factor $\frac{1}{\sqrt{k}}$ has the upper bound $\mathrm{RMS}(R_i(S), R_i(Q)) \leq 2\varepsilon$. By Lemma 4.5, the distributions of rows $R_i(S)$ and $R_i(Q)$ have the same upper bound for their EMD metrics: $\mathrm{EMD}_q\big(\mathrm{PDD}^{\{h\}}(S;k), \mathrm{PDD}^{\{h\}}(Q;k)\big) \leq 2\varepsilon\sqrt[q]{k}$ and $\mathrm{EMD}\leq 2\varepsilon$. □

Lemma 4.6 (Lipschitz continuity of PSD**)** *For all finite or periodic sequences $S\subset\mathbb{R}$, for the ground metrics* RMS *and L_q, define the* EMD *and* EMD_q *respectively, on distributions* $\mathrm{PSD}(S;k)$ *for $k\geq 1$, see Definition 3.6. Let $Q\subset\mathbb{R}$ be a finite or periodic sequence obtained by perturbing every point of S up to $\varepsilon\in[0, r(S))$. Then* $\mathrm{EMD}_q\big(\mathrm{PSD}(S;k), \mathrm{PSD}(Q;k)\big) \leq 2\varepsilon\sqrt[q]{k}$*,* $\mathrm{EMD}\big(\mathrm{PSD}(S;k), \mathrm{PSD}(Q;k)\big) \leq 2\varepsilon$*.*

Proof Since ε is less than the packing radius $r(S)$, the given ε-perturbation defines a bijection $g: S\to Q$, which changes inter-point distance by at most 2ε. The bijection g induces a 1-1 correspondence between rows $R_i(S)$ and $R_i(Q)$ of $\mathrm{PSD}(S;k)$ and $\mathrm{PSD}(Q;k)$, respectively with ground distances $L_q(R_i(S), R_i(Q)) \leq 2\varepsilon\sqrt[q]{k}$ and $\mathrm{RMS}(R_i(S), R_i(Q)) \leq 2\varepsilon$, which guarantee the required bounds by Lemma 4.5.

Recall the brief notation of a maximum metric from Definition 4.1(b):

$$\mathrm{EMD}_q^{(h)}[k](S,Q) = \max_{i=1,\dots,h}\Big\{\mathrm{EMD}_q\big(\,\mathrm{PDD}^{\{i\}}(S;k),\ \mathrm{PDD}^{\{i\}}(Q;k)\,\big)\Big\}.$$

Lemma 4.7 (lower bounds of EMD**)** *Fix any real $q\in[1,+\infty]$ and integers $h,k\geq 1\leq l\leq n$. Let $S,Q\subset\mathbb{R}^n$ be any finite or l-periodic point sets. Then*

(a) $\mathrm{EMD}_q^{(h)}[k](S,Q) \geq \mathrm{EMD}_q\big(\,\mathrm{PDD}^{(g)}(S;k),\ \mathrm{PDD}^{(g)}(Q;k)\,\big)$ *for $1\leq g\leq h$;*

(b) $\mathrm{EMD}_q^{\{h\}}[k](S,Q) \geq \mathrm{EMD}_q\big(\,\mathrm{PDD}^{\{h\}}(S;k'),\ \mathrm{PDD}^{\{h\}}(Q;k')\,\big)$ *for $1\leq k'\leq k$;*

(c) $\mathrm{EMD}_q^{\{h\}}[k](S,Q) \geq L_q\big(\,\mu^{(1)}[\mathrm{PDD}^{\{h\}}(S;k)],\ \mu^{(1)}[\mathrm{PDD}^{\{h\}}(Q;k)]\,\big)$*.*

The same inequalities hold for the ground metric RMS *instead of L_q.*

Proof **(a)** If the order h drops to g, the maximum of a fewer number of distances cannot become larger by Definition 4.1(b): $\mathrm{EMD}_q^{(h)}[k](S,Q) \geq \mathrm{EMD}_q^{(g)}[k](S,Q)$.

(b) Let $f_{ij}\in[0,1]$ be the parameters that minimize the EMD in Definition 4.1(b):

$$\mathrm{EMD}_q^{\{h\}}[k](S,Q) = \sum_{i=1}^{m(S)}\sum_{j=1}^{m(Q)} f_{ij}L_q(R_i(S), R_j(Q)),$$

where $R_i(S), R_j(Q)$ are rows in the distributions $\mathrm{PDD}^{\{h\}}(S;k), \mathrm{PDD}^{\{h\}}(Q;k)$, respectively. If k drops to k', the smaller distributions $\mathrm{PDD}^{\{h\}}(S;k'), \mathrm{PDD}^{\{h\}}(Q;k')$

are obtained from $\mathrm{PDD}^{\{h\}}(S;k),\mathrm{PDD}^{\{h\}}(Q;k)$ by removing the last $k-k'$ columns. The shortened rows $R'_i(S),R'_j(Q)$ of $k'\leq k$ components in the smaller distributions $\mathrm{PDD}^{\{h\}}(S;k'),\mathrm{PDD}^{\{h\}}(Q;k')$ satisfy $L_q(R_i(S),R_j(Q))\geq L_q(R'_i(S),R'_j(Q))$. Then

$$\mathrm{EMD}_q^{\{h\}}[k](S,Q)\geq\sum_{i=1}^{m(S)}\sum_{j=1}^{m(Q)}f_{ij}L_q(R_i(S),R_j(S))\geq$$
$$\min_{f_{ij}\in[0,1]}\sum_{i=1}^{m(S)}\sum_{j=1}^{m(Q)}f_{ij}L_q(R'_i(S),R'_j(Q))=\mathrm{EMD}_q^{\{h\}}[k'](S,Q).$$

(c) Considering $\mathrm{PDD}^{\{h\}}(S;k)$ as a weighted distribution of rows, $\mu^{(1)}[\mathrm{PDD}^{\{h\}}(S;k)]$ is its centroid from [16, section 3]. The argument below follows the proof for $q=+\infty$ of [16, Theorem 1]. Below we use the inequality $||u||_q+||v||_q||\geq||u+v||_q$ for the q-norm $||v||_q=\big(\sum\limits_{i=1}|v_i|^q\big)^{1/q}$ of the Minkowski metric L_q. Let $f_{ij}\in[0,1]$ be the parameters that minimize the EMD in Definition 4.1(b):

$$\mathrm{EMD}_q(\mathrm{PDD}^{\{h\}}(S;k),\mathrm{PDD}^{\{h\}}(Q;k))=\sum_{i=1}^{m(S)}\sum_{j=1}^{m(Q)}f_{ij}L_q(R_i(S),R_j(Q))=$$
$$\sum_{i=1}^{m(S)}\sum_{j=1}^{m(Q)}||f_{ij}\big(R_i(S)-R_j(Q)\big)||_q\geq||\sum_{i=1}^{m(S)}\sum_{j=1}^{m(Q)}f_{ij}(R_i(S)-R_j(Q))||_q=$$
$$||\sum_{i=1}^{m(S)}\big(\sum_{j=1}^{m(Q)}f_{ij}R_i(S)\big)-\sum_{j=1}^{m(Q)}\big(\sum_{i=1}^{m(S)}f_{ij}R_j(Q)\big)||_q=$$
$$||\sum_{i=1}^{m(S)}w_i(S)R_i(S)-\sum_{j=1}^{m(Q)}w_j(Q)R_j(Q)||_q=$$
$$L_q\big(\mu^{(1)}[\mathrm{PDD}^{\{h\}}(S;k)],\,\mu^{(1)}[\mathrm{PDD}^{\{h\}}(Q;k)]\big).$$

All proofs are the same for the ground metric $\mathrm{RMS}=\dfrac{L_2}{\sqrt{k}}$ instead of L_q. □

Corollary 4.8 extends the case $h=1$ from [66, Theorem 9], where $\mathrm{AMD}(S;k)=\mu^{(1)}[\mathrm{PDD}(S;k)]$ is the vector of Average Minimum Distances, to any order $h\geq 1$.

Corollary 4.8 (Lipschitz continuity of $\mu^{(1)}[\mathrm{PDD}^{\{h\}}]$) *Fix integers $h,k\geq 1\leq l\leq n$. Let Q be a finite or an l-periodic set obtained from a finite or an l-periodic point set $S\subset\mathbb{R}^n$, respectively, by perturbing every point of S up to a Euclidean distance $\varepsilon\in[0,r(S))$. Then $L_q\big(\mu^{(1)}[\mathrm{PDD}^{\{h\}}(S;k)],\,\mu^{(1)}[\mathrm{PDD}^{\{h\}}(Q;k)]\big)\leq 2\varepsilon\sqrt[q]{k}$ and $\mathrm{RMS}\big(\mu^{(1)}[\mathrm{PDD}^{\{h\}}(S;k)],\,\mu^{(1)}[\mathrm{PDD}^{\{h\}}(Q;k)]\big)\leq 2\varepsilon$.*

Proof The required bounds follow from Theorem 4.3 and Lemma 4.7(c). □

We conjecture that higher moments $\mu^{(t)}[\mathrm{PDD}^{\{h\}}]$ for $t>1$ are continuous under perturbations of points, possibly in a weaker (than Lipschitz) sense.

5 The asymptotic curves and computational complexity of $\mathrm{PDD}^{\{h\}}$

To analyze the asymptotic of $\mathrm{PDD}^{\{h\}}(S;k)$ as $k\to+\infty$, we choose a real $b\geq h$ such that $\dbinom{b}{h}=\dfrac{b(b-1)\dots(b-h+1)}{h!}$ belongs to the interval $(k-1,k]$. Then set $b(h,k) = b+1$ e.g. $b(1,k) = k+1$, $b(2,k) = 1.5+\sqrt{2k}$. Let V_n be the unit ball volume in $\mathbb{R}^n$, e.g. $V_2 = \pi$. Any periodic set $S \subset \mathbb{R}^n$ with a motif of m points and unit cell of volume $\mathrm{vol}[U]$ has the *point packing coefficient* $\mathrm{PPC}(S) = \sqrt[n]{\dfrac{\mathrm{vol}[U]}{mV_n}}$.

Theorem 5.1 (asymptotic of $\mathrm{PDD}^{\{h\}}(S;k)$ as $k \to +\infty$) *Let a periodic point set $S \subset \mathbb{R}^n$ have a cell with a longest diagonal d. For any integers $h,k \geq 1$, let $a(h,k)$ be the average sum of the k-th column of* $\mathrm{PDD}^{\{h\}}(S;k)$ *from Definition 3.1.*

$$\text{Then } \frac{2}{h+1}\Big(\mathrm{PPC}(S)\sqrt[n]{b(h,k)} - d\Big) \leq a(h,k) \leq \frac{2h}{h+1}\Big(\mathrm{PPC}(S)\sqrt[n]{b(h,k)} + d\Big)$$

for any $k \geq 1$. If $h = 1$, then $\lim\limits_{k\to+\infty} \dfrac{a(1,k)}{\sqrt[n]{k}} = \mathrm{PPC}(S)$. *If $h = 2$, then we have the bounds* $\dfrac{2}{3}\mathrm{PPC}(S) \leq \dfrac{a(2,k)}{\sqrt[2n]{2k}} \leq \dfrac{4}{3}\mathrm{PPC}(S)$ *for large enough k.*

Since any lattice $\Lambda \subset \mathbb{R}^n$ has a single point in a motif, any Pointwise Distance Distribution $\mathrm{PDD}^{\{h\}}(\Lambda;k)$ is a single row of the k numbers, which coincides with the vector $\mu^{(1)}[\mathrm{PDD}^{\{h\}}(\Lambda;k)]$ and can be visualized as a piecewise linear curve through k points. Fig. 6 shows six 2D lattices illustrating the asymptotic behavior of $\mathrm{PDD}^{\{h\}}$ for $h = 2,3$ in Fig. 7. Theorem 5.1 supports the following conjecture.

Conjecture 5.2 (h-order limit) *In the notations of Theorem 5.1 for any periodic point set $S \subset \mathbb{R}^n$,* $\lim\limits_{k\to+\infty} \dfrac{a(h,k)}{\sqrt[hn]{h!k}}$ *exists for any $h \geq 2$. If this limit differs from* $\mathrm{PPC}(S)$*, it can be called the h-order* point packing coefficient $\mathrm{PPC}(S;h)$.

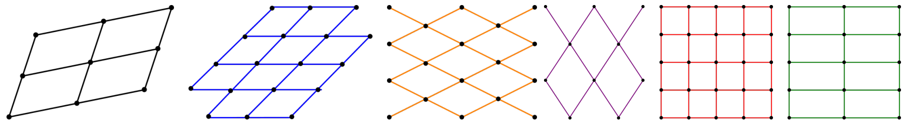


Fig. 6 The six 2-dimensional lattices whose invariants appeared in Fig. 7. **1st**: a generic black lattice Λ_1 with the basis $(1.25, 0.25), (0.25, 0.75)$ and $\mathrm{PPC}(\Lambda_1) = \sqrt{\dfrac{7}{8\pi}} \approx 0.525$. **2nd**: the blue hexagonal lattice Λ_2 with the basis $(1,0), (1/2, \sqrt{3}/2)$ and $\mathrm{PPC}(\Lambda_2) = \sqrt{\dfrac{\sqrt{3}}{2\pi}} \approx 0.528$. **3rd**: the orange rhombic lattice Λ_3 with the basis $(1, 0.5), (1, -0.5)$ and $\mathrm{PPC}(\Lambda_3) = \sqrt{\dfrac{1}{\pi}} \approx 0.564$. **4th**: the purple rhombic lattice Λ_4 with the basis $(1, 1.5), (1, -1.5)$ and $\mathrm{PPC}(\Lambda_4) = \sqrt{\dfrac{3}{\pi}} \approx 0.977$. **5th**: the red square lattice Λ_5 with the basis $(1,0), (0,1)$ and $\mathrm{PPC}(\Lambda_5) = \sqrt{\dfrac{1}{\pi}} \approx 0.564$. **6th**: the green rectangular lattice Λ_6 with the basis $(2,0), (0,1)$ and $\mathrm{PPC}(\Lambda_6) = \sqrt{\dfrac{2}{\pi}} \approx 0.798$.

Theorem 5.1 justifies that there is no need to substantially increase the number k of neighbors since $\mathrm{PDD}^{\{h\}}(S;k)$ largely depends on $\mathrm{PPC}(S)$ when $k \to +\infty$. The practical advice is to choose k depending on the size of a motif or constituent

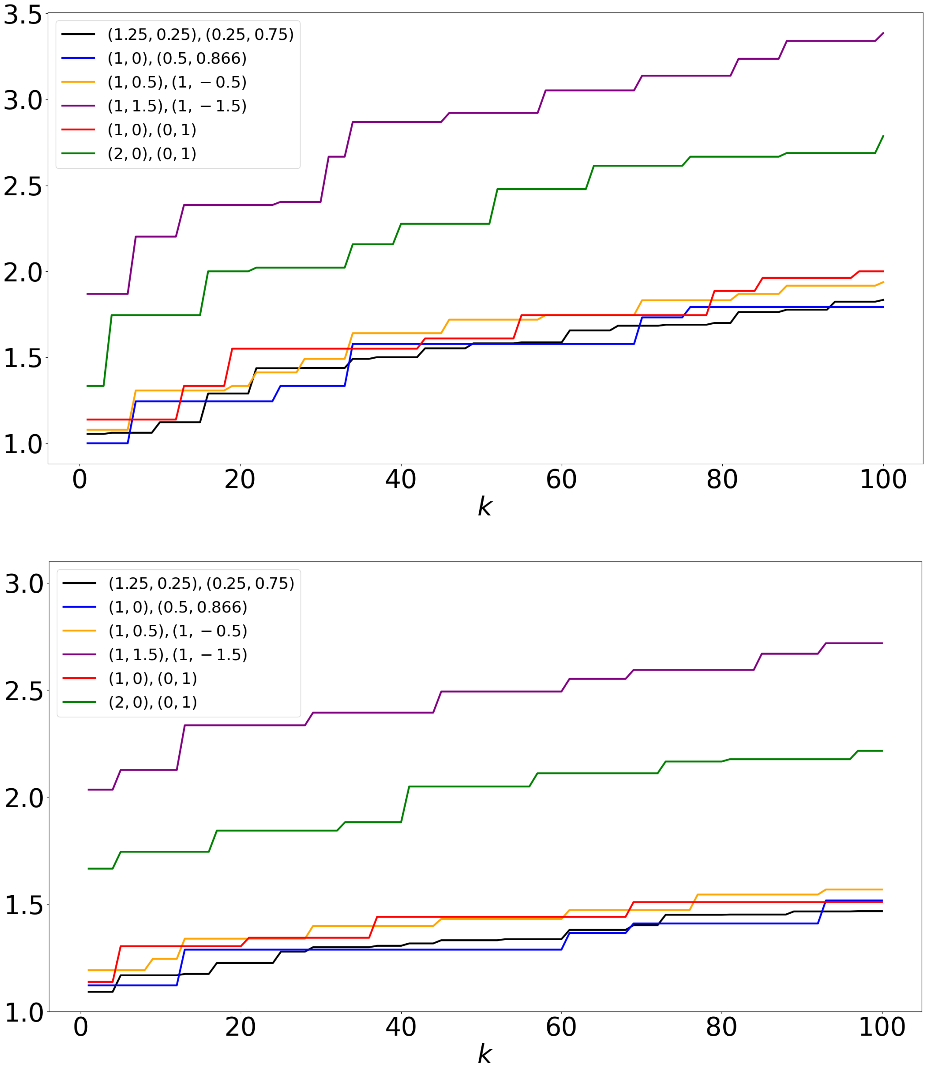


Fig. 7 The asymptotic behavior of the higher-order $\mathrm{PDD}^{\{2\}}(\Lambda; k)$ and $\mathrm{PDD}^{(3)}(\Lambda; k)$ for the six lattices $\Lambda \subset \mathbb{R}^2$ in Fig. 6, see their bases in the legends. **Top**: $h = 2$. **Bottom**: $h = 3$.

molecules so that all atoms have enough neighbors to capture the periodic connectivity. We consider k a degree of approximation similar to the number of decimal places on a calculator. Theorem 5.1 implies similar bounds for all t-moments and means that $\mathrm{PDD}^{\{h\}}(S; k)$ and $\mu^{(t)}[\mathrm{PDD}^{\{h\}}](S; k)$ are most discriminative for small values of k, so we used $k = 100$, $t \leq 10$, and $h \leq 3$ in all experiments later.

Lemma 5.3 (distance bounds) *Let $S \subset \mathbb{R}^n$ be any periodic point set. For any $h, k \geq 1$ and a point $p \in S$, let $a(h, k)$ be the k-th smallest average sum achieved* *for of all pairwise distances between p and h other points $p_1,\dots,p_h \in S$, see Definition 3.1. Set $R = \max\limits_{i=1,\dots,h} |p_i - p|$. Then $\dfrac{2R}{h+1} \le a(h,k) \le \dfrac{2hR}{h+1}$.*

Proof After translating $p \in S$ to the origin $0 \in \mathbb{R}^n$, one can assume that $p = 0$. Let $p_1 \in S$ be a point such that $R = |p_1| = \max\limits_{i=1,\dots,h} |p_i|$. For any other point $p_i \neq p_1$, the triangle inequalities $|p_i| + |p_1 - p_i| \ge |p_1| = R$ imply that

$$a(h,k) = \frac{2}{h(h+1)} \sum_{0 \le i < j \le h} |p_i - p_j| \ge$$

$$\ge \frac{2}{h(h+1)} \left(|p_1| + \sum_{i=2}^{h} (|p_i| + |p_1 - p_i|) \right) \ge \frac{2}{h(h+1)} \left(R + \sum_{i=2}^{h} R \right) = \frac{2R}{h+1}.$$

For the upper bound of $a(h,k)$, we use $|p_i| \le R$ and the triangle inequalities $|p_i - p_j| \le |p_i| + |p_j| \le 2R$ as follows:

$$a(h,k) = \frac{2}{h(h+1)} \left(\sum_{i=1}^{h} |p_i| + \sum_{1 \le i < j \le h} |p_i - p_j| \right) \le$$

$$\le \frac{2}{h(h+1)} \left(\sum_{i=1}^{h} R + \sum_{1 \le i < j \le h} 2R \right) = \frac{2}{h(h+1)} \left(hR + \frac{h(h-1)}{2} 2R \right) = \frac{2hR}{h+1},$$

which finishes the proof of the upper bound. □

For $h = 1$, the bounds of Lemma 5.3 give the exact equality $a(1,k) = R$. Lemma 5.4 was proved in a slightly more general form in Lemma 11 from [66].

Lemma 5.4 (number of points in a ball) *Let $S \subset \mathbb{R}^n$ be any periodic point set with a unit cell U, which has m points of S, generates a lattice Λ, and has a longest diagonal of a length d. For any point $p \in S \cap U$ and a radius r, consider*

$$U_-(p;r) = \bigcup_{v \in \Lambda} \{(U+v) \text{ such that } (U+v) \subset \bar{B}(p;r)\},$$

$$U_+(p;r) = \bigcup_{v \in \Lambda} \{(U+v) \text{ such that } (U+v) \cap \bar{B}(p;r) \neq \emptyset\}.$$

Then the number of points of S in the closed ball $\bar{B}(p;r)$ has the bounds

$$\left(\frac{r-d}{\mathrm{PPC}(S)} \right)^n \le m \frac{\mathrm{vol}[U_-(p;r)]}{\mathrm{vol}[U]} \le |S \cap \bar{B}(p;r)| \le m \frac{\mathrm{vol}[U_+(p;r)]}{\mathrm{vol}[U]} \le \left(\frac{r+d}{\mathrm{PPC}(S)} \right)^n.$$

Lemma 5.5 (increasing binomial coefficient) *For any fixed integer $h \ge 1$, the binomial coefficient $\dbinom{b}{h} = \dfrac{b(b-1)\dots(b-h+1)}{h!}$ is strictly increasing for any real $b \ge h$ so that if $h \le b < c$ then $\dbinom{b}{h} < \dbinom{c}{h}$.*

Proof The derivative $\dfrac{d}{dx}\dbinom{x}{h} > 0$ for any $x \geq h$.

Proof of Theorem 5.1. To prove the lower bound for the k-th smallest sum $a(h,k)$, set $r = \dfrac{h+1}{2}a(h,k)$. For any point p in a motif of S, consider the closed ball $\bar{B}(p;r)$ with the center p and radius r. By the lower bound of Lemma 5.3, all points $p_1,\dots,p_h \in S$ that are used for computing $a(h,k)$ have the maximum distance $R = \max\limits_{i=1,\dots,h} |p_i - p| \leq \dfrac{h+1}{2}a(h,k) = r$ and hence belong to $\bar{B}(p;r)$.

By the upper bound of Lemma 5.4, if this ball contains c points of S (excluding p), then $c+1 \leq \left(\dfrac{r+d}{\mathrm{PPC}(S)}\right)^n$. By using p and any other h distinct points $p_1,\dots,p_h$ among c points in $S \cap \bar{B}(p;r)$, we can form $\dbinom{c}{h} = \dfrac{c(c-1)\dots(c-h+1)}{h!}$ tuples $p, p_1,\dots,p_h$ whose average sums of all pairwise distances should include all k smallest values up to the k-th sum $a(h,k)$. Hence $\dbinom{c}{h} \geq k$.

For $c \geq h = 2$, the last inequality is $\dfrac{c(c-1)}{2} \geq k$, $c^2 - c - 2k \geq 0$, $c \geq \dfrac{1+\sqrt{1+8k}}{2} \geq 0.5 + \sqrt{2k}$. For any $h \geq 1$, let $b(h,k) = b+1$ satisfy $b \geq h$ and $\dbinom{b}{h} = \dfrac{b(b-1)\dots(b-h+1)}{h!} \in (k-1,k]$, e.g. one can set $b(2,k) = 1.5 + \sqrt{2k}$. By Lemma 5.5, $\dbinom{c}{h} \geq k$ for $c \geq h$ implies that $c \geq b = b(h,k) - 1$. Then

$$\left(\frac{r+d}{\mathrm{PPC}(S)}\right)^n \geq c+1 \geq b(h,k), \quad \frac{r+d}{\mathrm{PPC}(S)} \geq \sqrt[n]{b(h,k)},$$

$$\frac{h+1}{2}a(h,k) = r \geq \mathrm{PPC}(S)\sqrt[n]{b(h,k)} - d, \quad a(h,k) \geq \frac{2}{h+1}\Big(\mathrm{PPC}(S)\sqrt[n]{b(h,k)} - d\Big).$$

To prove the upper bound for the k-th sum $a(h,k)$, set $R = \dfrac{h+1}{2h}a(h,k)$ and consider any $r < R$. By the upper bound of Lemma 5.3, p with any other h points $p_1,\dots,p_h \in S \cap \bar{B}(p;r)$ have average sums that are at most $\dfrac{2hr}{h+1} < \dfrac{2hR}{h+1} = a(h,k)$, which is less than the k-th smallest sum $a(h,k)$. If the ball $\bar{B}(p;r)$ contains c points of S (excluding p), then these points can form at most $k-1$ tuples consisting of p and h (of c) other vertices, so $\dbinom{c}{h} \leq k-1$. Since $\dbinom{b}{h} = \dfrac{b(b-1)\dots(b-h+1)}{h!} \in (k-1,k]$ says that $\dbinom{b}{h} > k-1 \geq \dbinom{c}{h}$, Lemma 5.5 for $b = b(h,k) - 1 \geq h$ implies that $b > c$. Lemma 5.4 gives

$$\left(\frac{r-d}{\mathrm{PPC}(S)}\right)^n \leq c+1 < b+1 = b(h,k), \quad \frac{r-d}{\mathrm{PPC}(S)} < \sqrt[n]{b(h,k)}.$$

Since the resulting inequality $r < \mathrm{PPC}(S)\sqrt[n]{b(h,k)} + d$ holds for all $r < R$, where $R = \dfrac{h+1}{2h}a(h,k)$ is fixed, we get $\dfrac{h+1}{2h}a(h,k) = R \leq \mathrm{PPC}(S)\sqrt[n]{b(h,k)} + d$ and

$a(h,k) \leq \dfrac{2h}{h+1}\Big(\mathrm{PPC}(S)\,\sqrt[n]{b(h,k)}+d\Big)$. If $h=1$, both bounds have the same term:

$$\mathrm{PPC}(S)\,\sqrt[n]{b(1,k)} - d \leq a(h,k) \leq \mathrm{PPC}(S)\,\sqrt[n]{b(1,k)} + d.$$

If we divide both sides of the last inequality by $\sqrt[n]{k}$, we get $\lim\limits_{k\to+\infty} \dfrac{a(1,k)}{\sqrt[n]{k}} = \mathrm{PPC}(S)$. We replaced $k+1$ with k in $b(1,k)$, because $\lim\limits_{k\to+\infty} \dfrac{\sqrt[n]{k+1}}{\sqrt[n]{k}} = 1$ for any fixed dimension n. For similar reasons and $h=2$, the ratio $\dfrac{a(2,k)}{\sqrt[2n]{2k}}$ has the asymptotic bounds $\dfrac{2}{3}\mathrm{PPC}(S)$ and $\dfrac{4}{3}\mathrm{PPC}(S)$ as $k \to +\infty$. □

Since the average sums $a(h,k)$ are increasing according to Theorem 5.1 for $h=1,2$, comparing raw distances or sums for large k is affected by deviating asymptotics. To neutralize the effect of increasing deviations in k, [65, Definition 3.7] adjusted $\mathrm{PDD}(S;k)$ by subtracting $\mathrm{PPC}(S)\,\sqrt[n]{k}$ from distances to k-th neighbors. While Conjecture 5.2 remains open for $h \geq 2$, supported by Fig. 8, we find a coefficient c minimizing the sum of squared deviations $E(c) = \sum\limits_{j=1}^{k} (a(h,k) - c\,\sqrt[hn]{h!k})^2$. The polynomial $E(c)$ has the derivative $E'(c) = 2\sum\limits_{j=1}^{k} \sqrt[hn]{h!k}(c\,\sqrt[hn]{h!k} - a(h,k))$ and a global minimum at $c = \dfrac{\sum_{j=1}^{k} a(h,j)\,\sqrt[hn]{h!j}}{\sum_{j=1}^{k}(\,\sqrt[hn]{h!j})^2}$. Definition 5.6 uses this c to subtract from every row of $\mathrm{PDD}^{\{h\}}$ the fitted sequence $C(j) = c\,\sqrt[hn]{h!j}$ for $j=1,\dots,k$.

Definition 5.6 (Average/Pointwise Deviations from Asymptotic) ***(a)*** *Fix any integers $n,k \geq 1$ and $h \geq 2$. For a finite or periodic point set $S \subset \mathbb{R}^n$ and a point $p \in S$, let $a(h,1) \leq \dots \leq a(h,k)$ be the k column averages of the higher-order distribution* $\mathrm{PDD}^{\{h\}}(S;k)$*, considered a matrix of m unordered rows. Set $c(S;h,k) = \dfrac{\sum_{j=1}^{k} a(h,j)\,\sqrt[hn]{h!j}}{\sum_{j=1}^{k}(\,\sqrt[hn]{h!j})^2}$. Let $A(S;h,k)$ denote the matrix of m identical rows, each consisting of the k ordered elements $c(S;h,k)\,\sqrt[hn]{h!j}$ for $j=1,\dots,k$.*
(b) *The* Pointwise Deviation from Asymptotic $\mathrm{PDA}^{\{h\}}(S;k) = \mathrm{PDD}^{\{h\}}(S;k) - A(S;h,k)$ *is a distribution of unordered rows with the same weights as* $\mathrm{PDD}^{\{h\}}(S;k)$*. The t-moments from Definition 3.5 give the $t \times k$ matrix $\mu^{(t)}[\mathrm{PDA}^{\{h\}}(S;k)]$ of m ordered rows, which can be flattened to a vector of tk coordinates. The* Average Deviation from Asymptotic *is the vector* $\mathrm{ADA}^{\{h\}}(S;k) = \mu^{(1)}[\mathrm{PDA}^{\{h\}}(S;k)]$ *consisting of k column averages (counted with weights) of the $m \times k$ matrix* $\mathrm{PDA}^{\{h\}}(S;k)$.

Definition 5.6 for $h=1$ uses the Point Packing Coefficient $c(S;1,k) = \mathrm{PPC}(S)$, which depends only on S (independent of k), so $\mathrm{PDA}^{\{1\}}(S;k)$ coincides with the previously defined $\mathrm{PDA}(S;k)$ in [65, Definition 3.7]. We adapt EMD from Definition 4.1(a) to $\mathrm{PDA}^{\{h\}}(S;k)$ with the ground metric L_q on rows below.

Definition 5.7 (EMD for $\mathrm{PDA}^{\{h\}}$, $\mathrm{PDA}^{(h)}$ and Local Novelty Distances) ***(a)*** *For any real $q \in [1,+\infty]$, any integers $n,h,k \geq 1$, and any periodic point sets $S,Q \subset \mathbb{R}^n$, Definition 4.1(a) introduces the distances* EMD, EMD_q *between* $\mathrm{PDA}^{\{h\}}(S;k)$ *and* $\mathrm{PDA}^{\{h\}}(Q;k)$ *with the ground metrics* RMS, L_q*, respectively.*

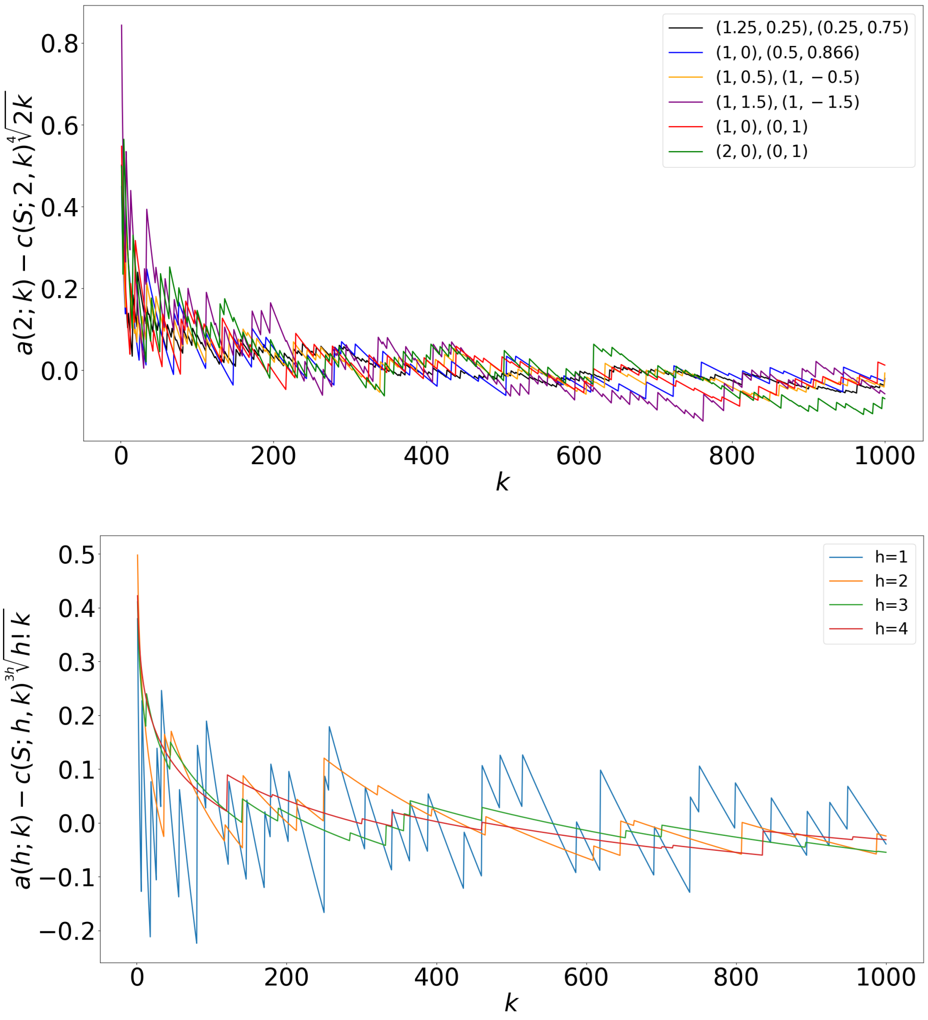


Fig. 8 By Definition 5.6, any lattice $\Lambda \subset \mathbb{R}^n$ has the vector $\mathrm{ADA}^{\{h\}}(\Lambda;k)$ consisting of the differences $a(h;k) - c(S;h,k)\ \sqrt[hn]{h!k}$, which converge to 0 in the plots. Here $a(h,k)$ is the k-th smallest average sum of pairwise distances from $0 \in \Lambda$ to h other points in Λ. The coefficient $c(S;h,k)$ was experimentally fitted for $h \geq 2$ but should be independent of k by Conjecture 5.2. **Top**: six 2D lattices from Fig. 6 and $h = 2$. **Bottom**: cubic lattice $\mathbb{Z}^3$ and $h = 1, 2, 3, 4$.

(b) *The joint invariant* $\mathrm{PDA}^{(h)}(S;k)$ *is obtained by concatenating* $\mathrm{PDA}^{\{i\}}(S;k)$ *for* $i = 1, \dots, h$. *Define the max metric between* $\mathrm{PDA}^{(h)}(S;k)$ *and* $\mathrm{PDA}^{(h)}(Q;k)$ *as the maximum of all distances* $\mathrm{EMD}_q\big(\mathrm{PDA}^{\{i\}}(S;k), \mathrm{PDA}^{\{i\}}(Q;k)\big)$ *for* $i = 1, \dots, h$, *similarly for the* EMD *based on the ground distance* RMS *instead of* L_q.

(c) *Fix an invariant distribution* I *with a metric* d, *e.g.* $I(S) = \mathrm{PDA}^{\{h\}}(S;k)$ *and* $d = \mathrm{EMD}$ *for all periodic point sets* $S \subset \mathbb{R}^n$. *Given a finite dataset* D *of periodic sets, the* $[I,d]$*-based* Local Novelty Distance $\mathrm{LND}[I,d](S;D) = \min\limits_{Q \in D} d(I(S), I(Q))$ *is the shortest distance from* S *to some* $Q \in D$ *in the metric* d *on values of* I.

Lemma 5.8 justifies computations with smaller h, k to filter out distant crystals.

Lemma 5.8 (bounds for metrics on $\mathrm{PDA}^{\{h\}}$**)** *Let $S, Q \subset \mathbb{R}^n$ be any periodic point sets. Fix any real $q \in [1, +\infty]$ and any integers $1 \leq g \leq h$, $1 \leq k' \leq k$. Then* $\mathrm{EMD}_q\big(\mathrm{PDA}^{(h)}(S;k), \mathrm{PDA}^{(h)}(Q;k)\big) \geq \mathrm{EMD}_q\big(\mathrm{PDA}^{(g)}(S;k'), \mathrm{PDA}^{(g)}(Q;k')\big)$*. The same inequality holds for* EMD *with the ground metric* RMS *instead of L_q.*

Proof follows similar to Lemma 4.7ab after replacing $\mathrm{PDD}^{\{h\}}$ with $\mathrm{PDA}^{\{h\}}$. □

Corollary 5.9 (Lipschitz continuity of $\mathrm{PDA}^{\{h\}}$**)** *Fix any integers $n, k, h \geq 1$ and $q \in [1, +\infty]$. Let Q be a periodic point set obtained from a periodic point set $S \subset \mathbb{R}^n$ by perturbing every point of S up to a Euclidean distance $\varepsilon \in [0, r(S))$.*

(a) *We have that $\big|c(S;h,k) - c(Q;h,k)\big| \leq 2\rho_k\varepsilon$ for $\rho_k = \dfrac{\sum_{j=1}^{k} \sqrt[hn]{h!j}}{\sum_{j=1}^{k} (\sqrt[hn]{h!j})^2}$. If $h = 1$, then $c(S;1,k) = \mathrm{PPC}(S)$ and $c(Q;1,k) = \mathrm{PPC}(Q)$ are equal, so we set $\rho_1 = 0$.*

(b) *Let $h = 1$. Then the following identities hold:*
$\mathrm{EMD}(\mathrm{PDA}(S;k), \mathrm{PDA}(Q;k)) = \mathrm{EMD}(\mathrm{PDD}(S;k), \mathrm{PDD}(Q;k))$,
$\mathrm{EMD}_q(\mathrm{PDA}(S;k), \mathrm{PDA}(Q;k)) = \mathrm{EMD}_q(\mathrm{PDD}(S;k), \mathrm{PDD}(Q;k))$,
$\mathrm{RMS}(\mathrm{ADA}(S;k), \mathrm{ADA}(Q;k)) = \mathrm{RMS}(\mathrm{AMD}(S;k), \mathrm{AMD}(Q;k))$,
$L_q(\mathrm{ADA}(S;k), \mathrm{ADA}(Q;k)) = L_q(\mathrm{AMD}(S;k), \mathrm{AMD}(Q;k))$.
Under given ε-perturbations, the Lipschitz constants of the metrics EMD, EMD_q, RMS, L_q *above are $2, 2\sqrt[q]{k}, 2, 2\sqrt[q]{k}$, respectively, for any parameter $q \in [1, +\infty]$.*

(c) *For $h \geq 2$, the upper bounds $\mathrm{EMD}_q(\mathrm{PDA}^{\{h\}}(S;k), \mathrm{PDA}^{\{h\}}(Q;k)) \leq 4\varepsilon\sqrt[q]{k}$ and $\mathrm{EMD}(\mathrm{PDA}^{\{h\}}(S;k), \mathrm{PDA}^{\{h\}}(Q;k)) \leq 4\varepsilon$ hold for the ground metric* RMS.

(d) *To get a known crystal $Q \in D$ from a new crystal S, some atom of S should be perturbed by at least $0.5\mathrm{LND}(S;D)$ for* LND *with the ground metric* EMD_∞.

Proof **(a)** [20, Lemma 4.1] proved that S, Q have a common lattice with a unit cell U such that $S = \Lambda + (S \cap U)$ and $Q = \Lambda + (Q \cap U)$. Then S, Q share a unit cell U and have the same number $m = m(S) = m(Q)$ of points in U, so $\mathrm{PPC}(S) = \mathrm{PPC}(Q)$, which proves the case $h = 1$. For $h \geq 2$, by Definition 5.6(a), we estimate the difference $\big|c(S;h,k) - c(Q;h,k)\big| \leq \dfrac{\sum_{j=1}^{k} \big|a(S;h,j) - a(Q;h,j)\big| \sqrt[hn]{h!j}}{\sum_{j=1}^{k} (\sqrt[hn]{h!j})^2}$. Since every point of S is obtained as an ε-perturbation of a point of Q, there is a bijection $S \to Q$ that shifts every point by at most ε. This bijection induces a 1-1 map between pairwise distances in S, Q, which changes every distance up to 2ε.

By Lemma 4.4, after writing the k smallest 2ε-perturbed average sums in increasing order in every row of $\mathrm{PDD}^{\{h\}}(S;k)$, the corresponding ordered sums still differ by at most 2ε, so $\big|\mathrm{PDD}^{\{h\}}_{i,j}(S;k) - \mathrm{PDD}^{\{h\}}_{i',j}(Q;k)\big| \leq 2\varepsilon$. Then the column averages $a(h,j)$ from Definition 5.6(a) also differ by at most 2ε.

Finally, $|a(S;h,j) - a(Q;h,j)| \leq 2\varepsilon$ gives the required upper bound

$$\big|c(S;h,k) - c(Q;h,k)\big| \leq \frac{\sum_{j=1}^{k} 2\varepsilon \sqrt[hn]{h!j}}{\sum_{j=1}^{k} (\sqrt[hn]{h!j})^2} = 2\rho_k\varepsilon \text{ for } \rho_k = \frac{\sum_{j=1}^{k} \sqrt[hn]{h!j}}{\sum_{j=1}^{k} (\sqrt[hn]{h!j})^2}.$$

(b) For $h = 1$, part (a) proved that $\mathrm{PPC}(S) = \mathrm{PPC}(Q)$. By Definition 5.6, the matrices $\mathrm{PDA}(S;k), \mathrm{PDA}(Q;k)$ are obtained from $\mathrm{PDD}(S;k), \mathrm{PDD}(Q;k)$, respectively, by subtracting the same vector consisting of $\mathrm{PPC}(S)\sqrt[n]{j}$, $j = 1,\dots,k$. Then any RMS or L_q distance between a row in $\mathrm{PDD}(S;k)$ and a row in $\mathrm{PDD}(Q;k)$ has the same value as between the corresponding rows in $\mathrm{PDA}(S;k)$ and $\mathrm{PDA}(Q;k)$.

The minimisation in Definition 4.1 gives the same values EMD and EMD_q when PDD is replaced with PDA. The same argument proves that RMS and L_q remain the same when AMD is replaced with ADA. Hence the Lipschitz constants are the same as in Theorem 4.3 and Corollary 4.8 restricted to order $h = 1$.

(c) By Definition 5.6, any element $\mathrm{PDA}^{\{h\}}_{i,j}(S;k)$ in a row i, and a column j of $\mathrm{PDA}^{\{h\}}(S;k)$ equals $\mathrm{PDD}^{\{h\}}_{i,j}(S;k) - c(S;h,k)\sqrt[hn]{h!j}$. Estimate the difference of i-th elements from the same column j in $\mathrm{PDA}^{\{h\}}(S;k)$ and $\mathrm{PDA}^{\{h\}}(Q;k)$.

$$\begin{aligned}
\varDelta &= \left|\mathrm{PDA}^{\{h\}}_{i,j}(S;k) - \mathrm{PDA}^{\{h\}}_{i,j}(Q;k)\right| = \\
&= \left|\left(\mathrm{PDD}^{\{h\}}_{i,j}(S;k) - c(S;h,k)\sqrt[hn]{h!j}\right) - \left(\mathrm{PDD}^{\{h\}}_{i,j}(Q;k) - c(Q;h,k)\sqrt[hn]{h!j}\right)\right| \\
&= \left|\left(\mathrm{PDD}^{\{h\}}_{i,j}(S;k) - \mathrm{PDD}^{\{h\}}_{i,j}(Q;k)\right) - \left(c(S;h,k) - c(Q;h,k)\right)\sqrt[hn]{h!j}\right| \leq \\
&\leq \left|\mathrm{PDD}^{\{h\}}_{i,j}(S;k) - \mathrm{PDD}^{\{h\}}_{i,j}(Q;k)\right| + \left|c(S;h,k) - c(Q;h,k)\right|\sqrt[hn]{h!j} \leq \\
&\leq 2\varepsilon + 2\rho_k\varepsilon\sqrt[hn]{h!j} = 2(1+\rho_k\sqrt[hn]{h!j})\varepsilon \leq 4\varepsilon,
\end{aligned}$$

where we used the upper bounds from part (a) and also $\rho_k\sqrt[hn]{h!j} \leq 1$ for any $j = 1,\dots,k$. Let $R_i(S), R_i(Q)$ denote the i-th rows of $\mathrm{PDA}^{\{h\}}(S;k)$, $\mathrm{PDA}^{\{h\}}(Q;k)$, respectively. Then $L_q(R_i(S), R_i(Q)) \leq \sqrt[q]{k(4\varepsilon)^q} = 4\varepsilon\sqrt[q]{k}$. The same proof for $\mathrm{RMS} = \dfrac{L_2}{\sqrt{k}}$ multiplies the Lipschitz constant by the factor $\dfrac{1}{\sqrt{k}}$. Lemma 4.5 guarantees the same upper bounds $4\varepsilon\sqrt[q]{k}$ and 4ε for EMD and EMD_q, respectively. If $h = 1$, then $\rho_1 = 0$ by part (a), so 4ε can be replaced with 2ε.

(d) Assume the contrary that Q can be obtained from S by perturbing every atom of S by at most $\varepsilon = 0.5\mathrm{LND}(S;D) = \min\limits_{Q\in D}\mathrm{EMD}_\infty(\mathrm{PDA}^{\{h\}}(S;k), \mathrm{PDA}^{\{h\}}(Q;k))$. Part (c) for $q = +\infty$ implies that $\mathrm{EMD}_\infty(\mathrm{PDA}^{\{h\}}(S;k), \mathrm{PDA}^{\{h\}}(Q;k)) \leq 2\varepsilon = \mathrm{LND}(S;D)$, which contradicts the assumption and hence proves the lemma. □

Theorem 5.10 (time of $\mathrm{PDD}^{\{h\}}$) *For any $h, k \geq 1$ and a periodic point set $S \subset \mathbb{R}^n$ with a motif of m points and a unit cell U with a longest diagonal d, let*

$$a = \max\left\{h\left(1 + \frac{2.5d}{\mathrm{PPC}(S)}\right), \sqrt[h]{16}\right\}, b = \log(2h!) + \log(\mathrm{PPC}(S) + d) - \log r(S),$$

where $r(S)$ is the packing radius of S. Then $\mathrm{PDD}^{\{h\}}(S;k)$ is computable in time

$$O\big(2^{8n}a^n\sqrt[h]{h!k}(b + \log k) + 2^{12n}m(\log k)\log(h!k) + a^{hn}mk\log k\big).$$

Proof Fix the origin $0 \in \mathbb{R}^n$ at the center of the unit cell U. Then any point $p \in M = S \cap U$ is covered by the closed ball $\bar{B}(0, 0.5d)$. By Theorem 5.1, the distance $a(1,k)$ from any point $p \in M$ to its k-th nearest neighbor in S has the upper bound $a(1,k) \leq \mathrm{PPC}(S)\sqrt[n]{k+1} + d$. Then all k neighbors of p in S are covered by the single ball $\bar{B}(0;r)$ of the radius $r = \mathrm{PPC}(S)\sqrt[n]{k+1} + 1.5d$.

For a fixed point p and any $h > 1$, to find a similar ball including all points that are needed to compute the k smallest average sums $a(h,1) \leq \cdots \leq a(h,k)$, we start from the integer number $c = \lceil b(h,k) - 1 \rceil$ of closest neighbors $p_1, \dots, p_c$ of p, where $b(h,k)$ is any real $b+1$ such that $b \geq h$ and $\binom{b}{h} \in (k-1, k]$. Then $\binom{c}{h} \geq k$ by Lemma 5.5. Since the $c+1$ points $p, p_1, \dots, p_c$ are covered by the ball $\bar{B}(p;R)$ of the radius $R = \max\limits_{i=1,\dots,c} |p_i - p|$, the lower bound of Lemma 5.4 gives $\left(\dfrac{R-d}{\mathrm{PPC}(S)}\right)^n \leq c+1 \leq \gamma$, where we set $\gamma = b(h,k)+1$, so $R \leq \mathrm{PPC}(S)\sqrt[n]{\gamma} + d$.

All $\binom{c}{h} \geq k$ average sums of pairwise distances between p and any h of c points from $S \cap \bar{B}(p;R)$ have the upper bound $\dfrac{2hR}{h+1}$ by Lemma 5.3. If the k smallest values of these sums are not greater than $\dfrac{2R}{h+1}$, which clearly holds for $h = 1$, these k smallest values form the required row $a(h,1) \leq \cdots \leq a(h,k)$ of the point $p = p_0$ in $\mathrm{PDD}^{\{h\}}(S;k)$. Indeed, in this case for any h points $p_1, \dots, p_h \in S$ with at least one distance (say) $|p_h - p_0| > R$, the lower bound of Lemma 5.3 implies that the average sum $\dfrac{2}{h(h+1)} \sum\limits_{0 \leq i < j \leq h} |p_i - p_j| > \dfrac{2R}{h+1}$ cannot be among the sought after k smallest values $a(h,1) \leq \cdots \leq a(h,k)$. If we could not find k smallest sums up to $\dfrac{2R}{h+1}$, we extend the radius R to hR.

Similar to the above argument for the smaller radius R, the lower bound of Lemma 5.3 guarantees than any average sum involving at least one point at a distance $|p_h - p_0| > hR$ is greater than $\dfrac{2hR}{h+1}$ and hence cannot be among $k \leq \binom{c}{h}$ smallest sums that were already considered for the smaller ball $\bar{B}(p;R)$. So the larger ball $\bar{B}(p;hR)$ is guaranteed to contain the required k smallest sums.

To cover the necessary neighbors of all points p from a motif $M = S \cap U$, we further increase the radius hR by $0.5d$ and will use the earlier upper bound $R \leq \mathrm{PPC}(S)\sqrt[n]{b} + d$ for $\gamma = b(h,k) + 1 \geq 1$. Let the ball $\bar{B}(p;hR + 0.5d)$ contain ν points of S, including its center p. The upper bound $\nu \leq \left(\dfrac{hR + 1.5d}{\mathrm{PPC}(S)}\right)^n$ from Lemma 5.4 and the earlier upper bound $R \leq \mathrm{PPC}(S)\sqrt[n]{\gamma} + d$ imply that

$$\nu \leq \left(\frac{hR+1.5d}{\mathrm{PPC}(S)}\right)^n \leq \left(h\sqrt[n]{\gamma} + \frac{(h+1.5)d}{\mathrm{PPC}(S)}\right)^n = h^n\gamma\left(1 + \frac{(1+1.5/h)d}{\mathrm{PPC}(S)\sqrt[n]{\gamma}}\right)^n \leq$$
$$\leq h^n\gamma\left(1 + \frac{2.5d}{\mathrm{PPC}(S)}\right)^n \leq a^n\gamma \text{ for } a = \max\left\{h\Big(1 + \frac{2.5d}{\mathrm{PPC}(S)}\Big), \sqrt[h]{16}\right\}.$$

To find ν nearest neighbors of all m points p from the motif $M = S \cap U$, we gradually extend the cell U in spherical layers by adding shifted copies of U until we get the upper union of shifted unit cells from Lemma 5.4:

$$U_+ = U_+(0; \mathrm{PPC}(S)h\sqrt[n]{\gamma} + 1.5d) \supset \bar{B}(0; hR + 0.5d).$$

To estimate the neighbor search time [22], we build a compressed cover tree on ν points of U_+ in time $O(\nu c_{\min}^8 \log \frac{2R}{d_{\min}})$ by [23, Theorem 3.7], where $c_{\min} \leq 2^n$

is the minimized expansion constant of T, and $\dfrac{R}{r(S)}$ is the upper bound for the ratio of max/min inter-point distances. Recall that $\gamma = b(h,k)+1 = b+2$ and $\dbinom{b}{h} \in (k-1,k]$. If $h=1$, then $\gamma = k+2 = O(k)$. For any $h \geq 2$, we have $\dbinom{b}{h} = \dfrac{O(b^h)}{h!} \leq k$. The rough upper bounds are $\gamma \leq O(\sqrt[h]{h!k})$ and $\log\gamma \leq O(\log(h!k))$ for any fixed h and $k \to +\infty$. Then $R \leq \mathrm{PPC}(S)\sqrt[n]{\gamma}+d$ gives

$$\log(2R) \leq \log(\sqrt[n]{\gamma}(2\mathrm{PPC}(S)+2d)) = \log(2\mathrm{PPC}(S)+2d) + \log\gamma.$$

Then $O\left(\log\dfrac{R}{r(S)}\right) \leq b + \log k$ for $b = \log(2h!) + \log(\mathrm{PPC}(S)+d) - \log r(S)$.

Then the time for a compressed cover tree on T is $O\left(\nu c_{\min}^{8}\log\frac{R}{r(S)}\right) \leq O\Big(\nu c_{\min}^{8}(b+\log k)\Big)$. Below we use the upper bounds $\nu \leq a^n\gamma \leq a^n O(\sqrt[h]{h!k})$ and $\log\nu \leq \log\gamma + n\log a \leq O(\log(h!k))$, where the second term was absorbed by the first one. Using [23, Theorem 4.9], we find k neighbors of m points among ν points of T in time $O(mc^2(\log k)(c_{\min}^{10}\log\nu + ck))$, where $c_{\min} \leq c \leq 2^n$ are expansion constants of T. Then we can compute all distances from each of m points from the motif $S \cap U$ to their k nearest neighbors in T in a time bounded as follows:

$$\begin{aligned}
&O\big(\nu c_{\min}^{8}(b+\log k)\big) + O\big(mc^2\log k(c_{\min}^{10}\log\nu + ck)\big) \leq \\
&O\big(2^{8n}a^n\sqrt[h]{h!k}(b+\log k)\big) + O\big(m2^{2n}\log k(2^{10n}(\log(h!k)+2^{2n}k)\big) \leq \\
&O\big(2^{8n}a^n\sqrt[h]{h!k}(b+\log k) + 2^{12n}m(\log k)\log(h!k) + 2^{4n}mk\log k\big). \qquad (*)
\end{aligned}$$

By Definition 3.1, to compute the k smallest average sums $a(h,1) \leq \cdots \leq a(h,k)$, we consider all unordered h-tuples of points among the found ν neighbors. Due to $\nu \leq a^n\gamma \leq a^n O(\sqrt[h]{h!k})$, the number of these h-tuples is $N = \dbinom{\nu}{h} \leq \dfrac{\nu^h}{h!} \leq \dfrac{a^{hn}}{h!}(O(\sqrt[h]{h!k}))^h = a^{hn}O(k)$. For each of m points in the motif $S \cap U$, we sort N average sums in time $O(N\log N) = a^{hn}O(k\log k)$ and select the k smallest average sums in time $a^{hn}O(mk\log k)$. When adding the latest time to the upper bound in $(*)$, we use $a^{hn} \geq 2^{4n}$, $a^h \geq 16$, $a \geq \sqrt[h]{16}$ to get the expected total time:

$$O\big(2^{8n}a^n\sqrt[h]{h!k}(b+\log k) + 2^{12n}m(\log k)\log(h!k) + a^{hn}mk\log k\big). \qquad \square$$

For small dimensions $n = 2,3$ and orders $h = 2,3$, the upper bound for the time of $\mathrm{PDD}^{\{h\}}$ becomes $O(mk\log k + k\sqrt[h]{k})$, which is close to be linear in both key inputs sizes: the motif size m and the number k of smallest average sums.

For any $h \geq 1$ and a periodic set $S \subset \mathbb{R}^n$ of up to m points in a unit cell, $\mathrm{PDD}^{\{h\}}(S;k)$, the exact EMD can be found in time $O(m^3\log m)$ [46]. By [63, Theorem 4.4], $\mathrm{PDD}(S;k)$ for a large enough k (and hence the stronger $\mathrm{PDD}^{(h)}$) together with a lattice of S and the minimum number m of points in a unit cell of S can be inverted to any generic S (outside a subspace of measure 0), uniquely under isometry. Then by Lemma 3.3 and Theorems 4.3, 5.10, $\mathrm{PDD}^{\{h\}}$ satisfies almost all conditions of Problem 1.2 with generic completeness instead of completeness.

6 Big-scale experiments on high-profile databases of inorganic crystals

This section applies new invariants to quantify the novelty of materials reported by A-lab [59] and MatterGen [69], and then reports pairwise comparisons by the hierarchy of new invariants across three large databases of inorganic crystals:

ICSD: Inorganic Crystal Structure Database [68], 170,206 entries
http://icsd.products.fiz-karlsruhe.de/en (version of February 25, 2025).

MP: Materials Project by the Berkeley lab [31], 153,235 entries
http://next-gen.materialsproject.org (version v2023.11.1).

GNoME: Graph Network Materials Exploration [43], 384,938 entries,
http://github.com/google-deepmind/materials_discovery (November 29, 2023).

Only experimentally measured non-disordered inorganic crystals from the ICSD were included. The Materials Project contains theoretical and experimental structures (including some sourced from the ICSD), but all entries undergo simulations which change their geometry before being added to the database. The GNoME dataset was generated by AI trained on crystals from the Materials Project.

We start by comparing the recently released crystals by MatterGen [69] with the already available structures in the ICSD and MP. Tables 1, 2 show several distances (based on the past PDD and new invariants $\mathrm{PDD}^{\{2\}}$) from MatterGen crystals to their three nearest neighbors in the ICSD and MP respectively.

All distances are measured in Angstroms, where 1Å is approximately the smallest inter-atomic distance. The physical meaning of all computed distances is justified by the Lipschitz continuity, which was proved for the past invariants ADA [66, Theorem 9], PDA [64, Theorem 6], and new invariants $\mathrm{PDA}^{\{h\}}$, see Theorem 4.3, as follows. If every atom of a periodic crystal S is perturbed up to $\varepsilon = 0.1$Å, then all our distances between each invariant of S and its perturbation is at most $2\varepsilon = 0.2$Å. Conversely, if a distance is $d = 0.2$Å, to match underlying crystals exactly, at least of their atoms should be shifted by at least $d/2 = 0.1$Å.

[32] suggested that the MatterGen crystal $\mathrm{TaCr_2O_6}$ is "identical" to ICSD entry 9516, which was reported in 1972 [7]. However, these crystals have $L_\infty[100] = 0.089$Å, $\mathrm{EMD}_\infty[100] = 0.098$Å, and $\mathrm{EMD}_\infty^{(2)}[100] = 0.196$Å, which are larger than the distances in the last three rows of Table 1. In fact, entry 9516 is outside the first 1000 neighbors of $\mathrm{TaCr_2O_6}$ by $\mathrm{EMD}_\infty^{(2)}$ in the ICSD. Tables 3 and 4 show the first neighbors of 43 A-lab crystals [59] in the ICSD and MP, respectively.

The distance $\mathrm{EMD}_\infty^{(2)}[k]$ between crystals S, Q is defined as the maximum of $\mathrm{EMD}_\infty(\mathrm{PDA}^{\{h\}}(S;k), \mathrm{PDA}^{\{h\}}(Q;k))$ for two orders $h = 1, 2$. In most cases, the maximum distance is achieved for order $h = 2$, because the 2nd order invariant $\mathrm{PDA}^{\{2\}}$ collects geometric data for triples of atoms instead of pairs (inter-atomic distances). However, very symmetric crystals can have many equal triangles, so the same number k of smallest inter-atomic distances can be more separating than k smallest perimeters of triangles. For example, the A-lab crystal $\mathrm{KMn_3O_6}$ has the nearest neighbors $\mathrm{K_{1.39}Mn_3O_6}$ (ICSD id 261406) and $\mathrm{KMn_2O_4}$ (mp-2765485 in the Materials Project) with $\mathrm{EMD}_\infty[100]$ distances 0.103Å and 0.51Å, which are larger than the $\mathrm{EMD}_\infty^{\{2\}}[100]$ distances 0.19Å and 0.444Å, respectively.

One reason that it was previously impossible to detect geometric duplicates in each of these databases and find substantial overlaps between different databases

MatterGen ID	ICSD composition	ICSD ID	L_∞[100]	EMD$_\infty$[100]	EMD$^{(2)}_\infty$[100]
Cr_2MoO_6	$LiMgFeF_6$	193630	0.022	0.037	0.074
Cr_2MoO_6	Cr_2WO_6	24793	0.017	0.032	0.075
Cr_2MoO_6	V_2WO_6	2576	0.045	0.064	0.098
$LaMoO_4$	$SmTaO_4$	59218	0.065	0.117	0.146
$LaMoO_4$	$SmTaO_4$	32996	0.056	0.125	0.188
$LaMoO_4$	$NdTaO_4$	79498	0.063	0.104	0.195
Mn_3NiO_6	$MgMnO_3$	690439	0.030	0.052	0.088
Mn_3NiO_6	$MgGeO_3$	171790	0.050	0.071	0.122
Mn_3NiO_6	$MgGeO_3$	171788	0.048	0.070	0.126
$Ta_{0.67}Cr_{1.33}O_4$	MgF_2	9164	0.009	0.012	0.020
$Ta_{0.67}Cr_{1.33}O_4$	MgF_2	117472	0.017	0.022	0.022
$Ta_{0.67}Cr_{1.33}O_4$	MgF_2	8121	0.017	0.022	0.022
$TaCr_2O_6$	$Ti\,Cr\,Sb\,O_6$	81932	0.014	0.025	0.047
$TaCr_2O_6$	MgF_2	8121	0.021	0.031	0.050
$TaCr_2O_6$	MgF_2	117472	0.021	0.031	0.050

Table 1 **Column 1**: IDs of 5 MatterGen crystals [69] in the folder 'experimental' [44]. **Columns 2-3**: compositions and IDs of three nearest neighbors in the ICSD, found by the new invariants, see column 6. **Column 4**: distance L_∞ on vector invariants $\mathrm{ADA}(S;100)$. **Column 5**: distance EMD_∞ on matrix invariants $\mathrm{PDA}(S;100)$. **Column 6**: max distance $\mathrm{EMD}^{(2)}_\infty$ on new invariants $\mathrm{PDA}^{\{h\}}(S;100)$ for orders $h=1,2$. All distances are in Angstroms.

MatterGen ID	MP composition	MP ID	L_∞[100]	EMD$_\infty$[100]	EMD$^{(2)}_\infty$[100]
Cr_2MoO_6	Cr_2VO_6	mp-1101261	0.010	0.018	0.032
Cr_2MoO_6	Cr_2WO_6	mp-19894	0.033	0.042	0.058
Cr_2MoO_6	Ga_2WO_6	mp-770737	0.018	0.028	0.059
$LaMoO_4$	$NdTaO_4$	mp-4718	0.065	0.108	0.207
$LaMoO_4$	$NbCeO_4$	mp-7550	0.083	0.133	0.233
$LaMoO_4$	$SmTaO_4$	mp-3756	0.077	0.131	0.269
Mn_3NiO_6	$MnMgO_3$	mp-770618	0.028	0.055	0.099
Mn_3NiO_6	$MnCoO_3$	mp-20641	0.032	0.062	0.110
Mn_3NiO_6	$FeMgO_3$	mp-754508	0.059	0.093	0.127
$Ta_{0.67}Cr_{1.33}O_4$	MgF_2	mp-1249	0.029	0.037	0.038
$Ta_{0.67}Cr_{1.33}O_4$	NiF_2	mp-559798	0.050	0.063	0.063
$Ta_{0.67}Cr_{1.33}O_4$	TiO_2	mp-2657	0.051	0.067	0.067
$TaCr_2O_6$	$LiNiRhF_6$	mp-1222366	0.027	0.048	0.051
$TaCr_2O_6$	MgF_2	mp-1249	0.033	0.046	0.058
$TaCr_2O_6$	$TiVO_4$	mp-690490	0.019	0.029	0.061

Table 2 **Column 1**: IDs of 5 MatterGen crystals [69] in the folder 'experimental' [44]. **Columns 2-3**: compositions and IDs of three nearest neighbors in the MP, found by the new invariants, see column 6. **Column 4**: distance L_∞ on vector invariants $\mathrm{ADA}(S;100)$. **Column 5**: distance EMD_∞ on matrix invariants $\mathrm{PDA}(S;100)$. **Column 6**: max distance $\mathrm{EMD}^{(2)}_\infty$ on new invariants $\mathrm{PDA}^{\{h\}}(S;100)$ for orders $h=1,2$. All distances are in Angstroms.

is their huge size and the slow speed of traditional comparisons. Our experiments were on a typical desktop (AMD Ryzen 5 5600X 6-core, 32GB RAM).

Another drawback of any distance is very limited information (a single number) per pair of crystals, while invariants such as $\mathrm{PDD}^{\{h\}}$ include many more numerical values per crystal. Detecting near-duplicates by invariants is much faster than by distances due to the hierarchy starting with vectors $\mathrm{ADA}^{\{h\}}(S;100)$, which quickly filter out distant crystals with $L_\infty > 0.01$Å. The stronger invariants $\mathrm{PDD}^{\{h\}}(S;100)$ cannot have smaller distances due to Lemma 4.7(c).

A-lab ID	ICSD composition	ICSD ID	EMD_∞	$\mathrm{EMD}_\infty^{(2)}$
Ba_2ZrSnO_6	$B_2Ho_2Pd_6$	44417	< 0.001	< 0.001
$Ba_6Na_2Ta_2V_2O_{17}$	$Ba_6Na_2Ru_2V_2O_{17}$	97524	0.092	0.132
$Ba_6Na_2Sb_2V_2O_{17}$	$Ba_6Na_2Ru_2V_2O_{17}$	97524	0.081	0.147
$Ba_9Ca_3La_4(Fe_4O_{15})_2$	$Ba_{10}Ca_2Pr_4(Fe_4O_{15})_2$	405911	0.187	0.212
$CaCo(PO_3)_4$	$Cd_{0.5}Co_{1.5}(PO_3)_4$	81574	0.144	0.236
$CaFe_2P_2O_9$	$CaV_2P_2O_9$	79735	0.073	0.093
$CaGd_2Zr(GaO_3)_4$	$Fe_5Tb_3O_{12}$	80550	0.108	0.168
$CaMn(PO_3)_4$	$Cd_2(PO_3)_4$	260975	0.168	0.220
$CaNi(PO_3)_4$	$Cd_{0.5}Co_{1.5}(PO_3)_4$	81574	0.157	0.235
$FeSb_3Pb_4O_{13}$	$Ni_{0.666}Sb_{3.33}Pb_4O_{13}$	88959	0.048	0.116
$Hf_2Sb_2Pb_4O_{13}$	$Ru_4Pb_4O_{13}$	49531	0.095	0.183
$InSb_3(PO_4)_6$	$Sc_4(SeO_4)_6$	1729	0.201	0.240
$InSb_3Pb_4O_{13}$	$Ru_4Pb_4O_{13}$	49531	0.149	0.284
$K_2TiCr(PO_4)_3$	$K_{1.928}Ti_{1.515}Fe_{0.485}(PO_4)_3$	418185	0.037	0.055
$K_4MgFe_3(PO_4)_5$	$K_4MgFe_3(PO_4)_5$	161484	0.075	0.109
$K_4TiSn_3(PO_5)_4$	$K_4Ti_{1.88}Sn_{2.12}(PO_5)_4$	250087	0.091	0.163
$KBaGdWO_6$	$K_2NaF_4NbO_2$	183827	0.004	0.010
$KBaPrWO_6$	$H_8F_6N_2NaV$	246824	0.004	0.009
KMn_3O_6	$K_{1.39}Mn_3O_6$	261406	0.103	0.103
$KNa_2Ga_3(SiO_4)_3$	$Na_3Ga_3(SiO_4)_3$	46861	0.110	0.146
$KNaP_6(PbO_3)_8$	$KNaP_6(PbO_3)_8$	182501	0.005	0.006
$KNaTi_2(PO_5)_2$	$K_{1.04}Na_{0.96}Ti_2(PO_5)_2$	71239	0.062	0.105
$KPr_9(Si_3O_{13})_2$	$Sr_{1.91}Nd_{8.09}(Si_3O_{13})_2$	238283	0.144	0.172
$Mg_3MnNi_3O_8$	$Mg_{1.2}MnNi_{4.8}O_8$	80303	0.020	0.031
Mg_3NiO_4	Mg_4O_4	690939	0.000	0.000
$MgCuP_2O_7$	$Mg_{1.08}Co_{0.92}P_2O_7$	69576	0.218	0.227
$MgNi(PO_3)_4$	$Mg_2(PO_3)_4$	4280	0.082	0.097
$MgTi_2NiO_6$	$Mn_{0.64}Ti_2Ni_{1.36}O_6$	238957	0.045	0.056
$MgTi_4(PO_4)_6$	$FeTi_4(PO_{12})_6$	290966	0.132	0.152
$MgV_4Cu_3O_{14}$	$V_4Cu_4O_{14}$	164189	0.146	0.193
Mn_2VPO_7	$Mn_2V_{0.91}P_{1.09}O_7$	250126	0.219	0.333
$Mn_4Zn_3(NiO_6)_2$	$Mg_6Ti_3O_{12}$	65793	0.128	0.186
$Mn_7(P_2O_7)_4$	$Fe_7(P_2O_7)_4$	67514	0.126	0.155
$MnAgO_2$	$MnAgO_2$	670065	0.097	0.142
$Na_3Ca_{18}Fe(PO_4)_{14}$	$K_2Sr_{18}Mg_2(PO_4)_{14}$	127462	0.173	0.252
$Na_7Mg_7Fe_5(PO_4)_{12}$	$Na_8Ni_8Fe_4(PO_4)_{12}$	169444	0.157	0.157
$NaCaMgFe(SiO_3)_4$	$V_{0.28}Cr_{0.49}Mn_{0.004}Ti_{0.002}$ $Na_{0.792}Ca_{1.208}Mg_{1.17}$ $Fe_{0.016}Si_{3.98}O_{12}$	117172	0.066	0.096
$NaMnFe(PO_4)_2$	$Na_{1.17}Mg_{0.19}Mn_{0.46}Fe_{1.35}$ $(PO_4)_2$	168037	0.232	0.232
$Sn_2Sb_2Pb_4O_{13}$	$Ru_4Pb_4O_{13}$	49531	0.088	0.188
$Y_3In_2Ga_3O_{12}$	$Y_{2.74}Sc_{2.19}Ga_{3.01}O_{12}$	39834	0.018	0.041
$Zn_2Cr_3FeO_8$	$Mg_2Cr_4O_8$	160954	0.022	0.035
$Zn_3Ni_4(SbO_6)_2$	$CoLi_2Ti_{2.8}O_8$	19999	0.173	0.211
$Zr_2Sb_2Pb_4O_{13}$	$Ru_4Pb_4O_{13}$	49531	0.106	0.218

Table 3 **Column 1**: IDs of 43 A-lab crystals reported in [59]. **Columns 2-3**: compositions and IDs of the nearest neighbor in the ICSD, found by the new invariants, see column 5. **Column 4**: distance EMD_∞ on matrix invariants $\mathrm{PDA}(S; 100)$. **Column 5**: max distance $\mathrm{EMD}_\infty^{(2)}$ on new invariants $\mathrm{PDA}^{\{h\}}(S; 100)$ for orders $h = 1, 2$. All distances are in Angstroms.

The invariants $\mathrm{PDA}^{(h)}$ obtained by concatenating $\mathrm{PDA}, \mathrm{PDA}^{\{2\}}, \dots, \mathrm{PDA}^{\{h\}}$ form a natural hierarchy so that increasing the order $h = 1, 2, \dots$ adds more invariant information to better distinguish given crystals under isometry.

A-lab ID	MP composition	MP ID	EMD_∞	$\mathrm{EMD}_\infty^{(2)}$
Ba_2ZrSnO_6	Hf_2KPrO_6	mp-1522216	< 0.001	< 0.001
$Ba_6Na_2Ta_2V_2O_{17}$	$Ba_6Na_2Ta_2V_2O_{17}$	mp-1214664	0.029	0.051
$Ba_6Na_2Sb_2V_2O_{17}$	$Ba_6Na_2Sb_2V_2O_{17}$	mp-1214658	0.021	0.030
$Ba_9Ca_3La_4(Fe_4O_{15})_2$	$Ba_9Ca_3La_4(Fe_4O_{15})_2$	mp-1228537	0.136	0.141
$CaCo(PO_3)_4$	$CaCo(PO_3)_4$	mp-1045787	0.090	0.090
$CaFe_2P_2O_9$	$CaV_2P_2O_9$	mp-21541	0.061	0.088
$CaGd_2Zr(GaO_3)_4$	$CaGd_2Zr(GaO_3)_4$	mp-686296	0.069	0.072
$CaMn(PO_3)_4$	$CaTi(PO_3)_4$	mp-1045626	0.071	0.097
$CaNi(PO_3)_4$	$CaCo(PO_3)_4$	mp-1045787	0.105	0.121
$FeSb_3Pb_4O_{13}$	$FeSb_3Pb_4O_{13}$	mp-1224890	0.027	0.034
$Hf_2Sb_2Pb_4O_{13}$	$Hf_2Sb_2Pb_4O_{13}$	mp-1224490	0.012	0.022
$InSb_3(PO_4)_6$	$InSb_3(PO_4)_6$	mp-1224667	0.011	0.018
$InSb_3Pb_4O_{13}$	$InSb_3Pb_4O_{13}$	mp-1223746	0.029	0.036
$K_2TiCr(PO_4)_3$	$K_2TiCr(PO_4)_3$	mp-1224541	0.009	0.019
$K_4MgFe_3(PO_4)_5$	$K_4MgFe_3(PO_4)_5$	mp-532755	0.076	0.088
$K_4TiSn_3(PO_5)_4$	$K_4TiSn_3(PO_5)_4$	mp-1224290	0.014	0.025
$KBaGdWO_6$	$NaSmEuWO_6$	mp-1523299	0.001	0.003
$KBaPrWO_6$	$NaNiRb_2F_6$	mp-556353	0.003	0.007
KMn_3O_6	KMn_2O_4	mp-2765485	0.510	0.510
$KNa_2Ga_3(SiO_4)_3$	$KNa_2Ga_3(SiO_4)_3$	mp-1211711	0.022	0.032
$KNaP_6(PbO_3)_8$	$Na_2P_6(PbO_3)_8$	mp-690977	0.090	0.121
$KNaTi_2(PO_5)_2$	$KNaTi_2(PO_5)_2$	mp-1211611	0.012	0.016
$KPr_9(Si_3O_{13})_2$	$KPr_9(Si_3O_{13})_2$	mp-1223421	0.009	0.021
$Mg_3MnNi_3O_8$	$Mg_3MnNi_3O_8$	mp-1222170	0.029	0.032
Mg_3NiO_4	Mg_3CuO_4	mp-1099249	0.001	0.002
$MgCuP_2O_7$	$MgCuP_2O_7$	mp-1041741	0.093	0.088
$MgNi(PO_3)_4$	$MgNi(PO_3)_4$	mp-1045786	0.018	0.024
$MgTi_2NiO_6$	$MgTi_2NiO_6$	mp-1221952	0.009	0.023
$MgTi_4(PO_4)_6$	$MgTi_4(PO_4)_6$	mp-1222070	0.075	0.076
$MgV_4Cu_3O_{14}$	$MgV_4Cu_3O_{14}$	mp-1222158	0.060	0.070
Mn_2VPO_7	Mn_2VPO_7	mp-1210613	0.125	0.153
$Mn_4Zn_3(NiO_6)_2$	$Mn_4Zn_3(NiO_6)_2$	mp-1222033	0.054	0.063
$Mn_7(P_2O_7)_4$	$Mn_7(P_2O_7)_4$	mp-778008	0.123	0.132
$MnAgO_2$	$MnAgO_2$	mp-996995	0.098	0.112
$Na_3Ca_{18}Fe(PO_4)_{14}$	$Na_3Ca_{18}Fe(PO_4)_{14}$	mp-725491	0.031	0.038
$Na_7Mg_7Fe_5(PO_4)_{12}$	$Na_7Mg_7Fe_5(PO_4)_{12}$	mp-1173791	0.028	0.036
$NaCaMgFe(SiO_3)_4$	$NaCaMgFe(SiO_3)_4$	mp-1221075	0.026	0.032
$NaMnFe(PO_4)_2$	$NaMnFe(PO_4)_2$	mp-1173592	0.032	0.034
$Sn_2Sb_2Pb_4O_{13}$	$Sn_2Sb_2Pb_4O_{13}$	mp-1219056	0.025	0.038
$Y_3In_2Ga_3O_{12}$	$Y_3In_2Ga_3O_{12}$	mp-1207946	0.008	0.028
$Zn_2Cr_3FeO_8$	$Mg_2Ga_4O_8$	mp-4590	0.022	0.040
$Zn_3Ni_4(SbO_6)_2$	$Zn_3Ni_4(SbO_6)_2$	mp-1216023	0.092	0.108
$Zr_2Sb_2Pb_4O_{13}$	$Zr_2Sb_2Pb_4O_{13}$	mp-1215826	0.025	0.042

Table 4 **Column 1**: IDs of 43 A-lab crystals reported in [59]. **Columns 2-3**: compositions and IDs of the nearest neighbor in the MP, found by the new invariants, see column 5. **Column 4**: distance EMD_∞ on matrix invariants $\mathrm{PDA}(S;100)$. **Column 5**: max distance $\mathrm{EMD}_\infty^{(2)}$ on new invariants $\mathrm{PDA}^{\{h\}}(S;100)$ for orders $h=1,2$. All distances are in Angstroms.

In addition to L_∞-based distances in Tables 1- 4, below we also use metrics based on RMS (Root Mean Square) between vectors or rows of PDA matrices, so the resulting EMD on $\mathrm{PDA}^{\{h\}}$ is written without a subscript for simplicity. The RMS-based metrics have Lipschitz constant 2 (or 4 for $h > 1$) by Corollary 5.9.

Since any computations accumulate arithmetic errors, we start by finding geometric near-duplicates (under isometry including reflections) with the threshold $10^{-10}\text{Å} = 10^{-18}m$ for all distances. Then we gradually increase the threshold to 0.01Å, which is about 1% of the smallest interatomic distance and considered experimental noise. Tables 5-10 summarize all-vs-all comparisons across the databases ICSD, MP, and GNoME by using two distances on the new PDA$^{(2)}$.

Table 5 Count and percentage of pure periodic crystals in each database (left) found to have a near-duplicate in other databases (top) by EMD and EMD$_\infty$ under 10^{-6}Å on PDA$^{(2)}(S; 100)$.

Data	ICSD				MP				GNoME			
	L_∞		RMS		L_∞		RMS		L_∞		RMS	
	count	%	count	%	count	%	count	%	count	%	count	%
ICSD	**9454**	**8.05**	**9462**	**8.05**	53	0.05	154	0.13	1	0.00	8	0.01
MP	26	0.02	87	0.06	**80**	**0.05**	**293**	**0.19**	10	0.01	21	0.01
GNoME	1	0.00	8	0.00	10	0.00	20	0.01	**4351**	**1.13**	**4392**	**1.14**

Table 6 Count and percentage of pure periodic crystals in each database (left) found to have a near-duplicate in other databases (top) by EMD and EMD$_\infty$ under 10^{-5}Å on PDA$^{(2)}(S; 100)$.

Data	ICSD				MP				GNoME			
	L_∞		RMS		L_∞		RMS		L_∞		RMS	
	count	%	count	%	count	%	count	%	count	%	count	%
ICSD	**9509**	**8.09**	**9779**	**8.32**	273	0.23	1021	0.87	18	0.02	84	0.07
MP	176	0.11	764	0.50	**545**	**0.36**	**2067**	**1.35**	41	0.03	161	0.11
GNoME	14	0.00	55	0.01	38	0.01	138	0.04	**4432**	**1.15**	**4590**	**1.19**

Table 7 Count and percentage of pure periodic crystals in each database (left) found to have a near-duplicate in other databases (top) by EMD and EMD$_\infty$ under 10^{-4}Å on PDA$^{(2)}(S; 100)$.

Data	ICSD				MP				GNoME			
	L_∞		RMS		L_∞		RMS		L_∞		RMS	
	count	%	count	%	count	%	count	%	count	%	count	%
ICSD	**10411**	**8.86**	**12845**	**10.93**	1910	1.63	4708	4.01	170	0.14	636	0.54
MP	1595	1.04	5182	3.38	**3709**	**2.42**	**7018**	**4.58**	343	0.22	1289	0.84
GNoME	122	0.03	393	0.10	268	0.07	507	0.13	**4808**	**1.25**	**5070**	**1.32**

Tables 5-9 count near-duplicates (under isometry not distinguishing mirror images) within each database, which can be filtered out for any analysis or training, else the data becomes skewed. The ultra-fast ADA$(S; 100)$ finds nearest neighbors within and between all databases using KD-trees [25]. All pairs within a given threshold by ADA$(S; 100)$ were re-compared by the stronger ADA$^{(2)}(S; 100)$, followed by PDA$(S; 100)$ and finally PDA$^{(2)}(S; 100)$, each time keeping only the pairs with distances within the threshold. To avoid repeated calculations, all invariants were computed separately before making comparisons, see Fig. 9 and Table 11.

Some experimental materials of different compositions may have very close geometries because their structures were determined under different conditions, such as temperature and pressure, which will be discussed in future work.

Table 8 Count and percentage of pure periodic crystals in each database (left) found to have a near-duplicate in other databases (top) by EMD and EMD$_\infty$ under 10^{-3}Å on PDA$^{(2)}(S;100)$.

Data	ICSD				MP				GNoME			
	L_∞		RMS		L_∞		RMS		L_∞		RMS	
	count	%	count	%	count	%	count	%	count	%	count	%
ICSD	**16052**	**13.66**	**21073**	**17.94**	6722	5.72	9975	8.49	1263	1.08	3637	3.10
MP	7228	4.72	9275	6.05	**8301**	**5.42**	**10511**	**6.86**	2460	1.61	5821	3.80
GNoME	589	0.15	793	0.21	625	0.16	906	0.24	**5581**	**1.45**	**8049**	**2.09**

Table 9 Count and percentage of pure periodic crystals in each database (left) found to have a near-duplicate in other databases (top) by EMD and EMD$_\infty$ under 0.01Å on PDA$^{(2)}(S;100)$.

Data	ICSD				MP				GNoME			
	L_∞		RMS		L_∞		RMS		L_∞		RMS	
	count	%	count	%	count	%	count	%	count	%	count	%
ICSD	**30855**	**25.9**	**27898**	**23.4**	12540	10.5	12004	10.1	6120	5.14	5853	4.92
MP	12607	5.99	12588	5.98	**18466**	**8.77**	**18047**	**8.57**	11283	5.35	11296	5.36
GNoME	1379	0.36	1230	0.32	4645	1.21	4998	1.30	**35314**	**9.17**	**49403**	**12.8**

Table 10 Each database has thousands of (near-)duplicates (by L_∞) whose all atomic positions can be matched by tiny perturbations. Duplication with different compositions is unexpected for very low thresholds as replacing an atom with a different one should affect geometry.

near-duplicates	database	10^{-10}Å	10^{-6}Å	10^{-5}Å	10^{-4}Å	10^{-3}Å	10^{-2}Å
pairs of entries	ICSD	8994	8995	9036	10353	33314	259169
within a threshold	MP	5	40	283	2718	26703	278739
by EMD on PDA$^{(2)}$	GNoME	1852	2482	2524	2719	3284	39487
percentage of all entries	ICSD	8.05	8.05	8.09	8.86	13.66	26.24
in close pairs vs	MP	0.01	0.05	0.36	2.42	5.42	9.42
the full database	GNoME	0.84	1.13	1.15	1.25	1.45	9.17
percentage of close	ICSD	46.91	46.90	46.70	45.36	51.96	71.19
pairs with different	MP	60.00	85.00	97.53	99.01	99.88	99.93
chemical compositions	GNoME	33.86	47.38	46.71	44.28	47.05	90.96

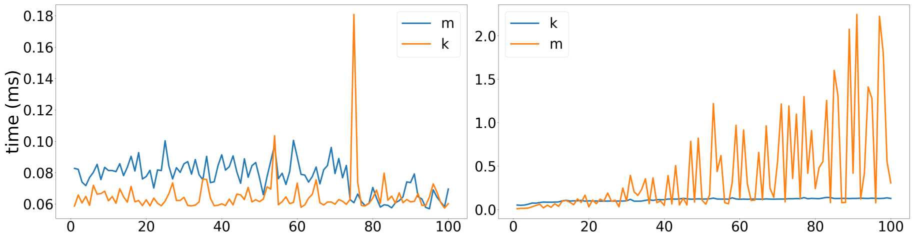


Fig. 9 Time in milliseconds to compare invariants ADA (left) and PDA (right). Blue: average over pairwise comparisons of 10 random crystals from 3 databases for $k = 100$ and a fixed size m of an asymmetric unit. Orange: average per atom over 500 random crystals for $k = 1, \ldots, 100$.

Table 12 shows five pairs that were found in the Materials Project within 10^{-10}Å for L_∞ on ADA$(S;100)$. Three pairs with different compositions have identical numerical data and likely need updating because changing chemical elements should perturb geometry. Two pairs with identical compositions have identical cells and atomic coordinates that can be matched by reflection, see the appendix.

Table 11 Times (hours:minutes:seconds) to calculate PDA and PDA$^{\{2\}}$ for each database, and times to compare each pair of databases by the metric EMD$_\infty$ to produce Table 9. The vectors ADA and ADA$^{\{2\}}$ are near-instantly computable from PDA and PDA$^{\{2\}}$, respectively.

data	Invariants		Comparisons			Sum of
	PDA	PDA$^{\{2\}}$	ICSD	MP	GNoME	times
ICSD	0:01:07	5:57:26	0:04:02	0:04:47	0:00:50	6:08:12
MP	0:04:35	25:44:33	0:04:10	0:05:32	0:01:44	26:00:34
GNoME	0:03:47	9:54:48	0:00:43	0:01:51	0:14:53	10:16:02

Table 12 Geometrically identical entries in MP, three of which have different compositions.

MP id 1	MP id 2	composition 1	composition 2	compositional distance [29]
mp-1100417	mp-631388	VSbRh	CdIrRu	8
mp-1013559	mp-1013733	Sr_3As_2	Ca_3BiSb	1.2
mp-1013536	mp-1013552	Sr_3PN	Sr_3P_2	0.2
mp-771976	mp-1345479	$Rb_2Be_2O_3$	$Rb_2Be_2O_3$	0
mp-29783	mp-1338697	B_5H_9	B_5H_9	0

7 Discussion of limitations, scientific integrity, and growing significance

Diffraction patterns helped predict cell-based representations of crystals for 100+ years. Recently, [57] showed how to convert any crystal into many different homometric structures that have identical diffraction. Fig. 1 (right) illustrated how any known crystal can be easily disguised by changing or expanding its cell, perturbing atoms to make the new cell primitive, and changing chemical elements.

As a result, artificially generated structures threaten the integrity of experimental databases [13], which are already skewed by previously undetectable near-duplicates in other databases [1]. These practical challenges motivated us to formalize the fundamental questions *Same or different, and by how much?* [54]. Problem 1.2 asked for a complete, Lipschitz continuous, and polynomial-time invariant of all periodic point sets with up to m points in a unit cell, and is being addressed for other real objects in the emerging area of Geometric Data Science [38].

While diffraction patterns and PDDs cannot distinguish infinitely many homometric crystals, PDD$^{\{2\}}$ distinguished all known (infinitely many) counter-examples to the completeness of the PDD under isometry in Examples 3.4 and 4.2. For practical dimensions and orders $n, h \leq 3$, the time of PDD$^{\{h\}}$ is near-linear in both key input sizes k, m by Theorem 5.10. The new hierarchy of ADA$^{\{h\}}$ and PDA$^{\{h\}}$ for $h \geq 1$ allows us to use the stronger invariants PDA$^{\{2\}}$ only in rare cases to confirm exact duplicates after much faster filtering by ADA, PDA.

The limitations are Conjectures 3.8 (completeness of PDD$^{(h)}$ in $\mathbb{R}^h$) and 5.2 (exact asymptotic of PDD$^{\{h\}}$ for $h > 1$), which we plan to tackle in future work.

Before Theorem 3.7, there was no complete, continuous, and polynomial-time invariant of periodic sets even in dimension $n = 1$. The developed hierarchy quickly detects near-duplicates of any newly claimed materials in existing databases and hence becomes an efficient barrier for noisy disguises of known crystals.

This research was supported by the Royal Society APEF fellowship “New geometric methods for mapping the space of periodic crystals’ ’(APX/R1/231152) and EPSRC New Horizons grant “ Inverse design of periodic crystals” (EP/X018474/1).

References

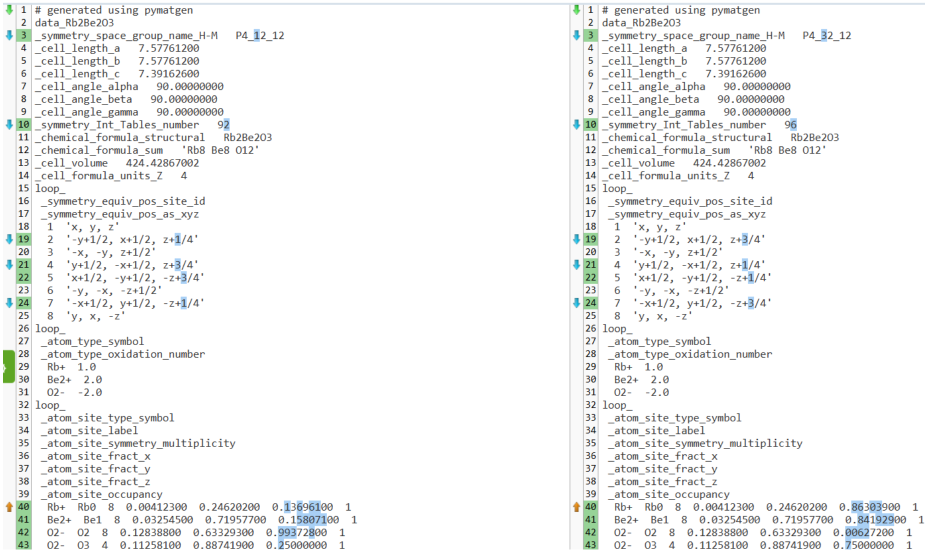
1. Anosova, O., Gorelov, A., Jeffcott, W., Jiang, Z., Kurlin, V.: A complete and bi-continuous invariant of protein backbones under rigid motion. MATCH Comm. Math. Comp. Chemistry **94**, 97–134 (2025)
2. Anosova, O., Kurlin, V.: An isometry classification of periodic point sets. In: Proceedings of Discrete Geometry and Mathematical Morphology, pp. 229–241 (2021)
3. Anosova, O., Kurlin, V.: Density functions of periodic sequences. In: LNCS Proceedings of Discrete Geometry and Mathematical Morphology, vol. 13493, pp. 395–408 (2022)
4. Anosova, O., Kurlin, V.: Density functions of periodic sequences of continuous events. Journal of Mathematical Imaging and Vision **65**, 689–701 (2023)
5. Anosova, O., Kurlin, V., Senechal, M.: The importance of definitions in crystallography. IUCrJ **11**, 453–463 (2024)
6. Anosova, O., Widdowson, D., Kurlin, V.: Recognition of near-duplicate periodic patterns by continuous metrics with approximation guarantees. Pattern Recognition **171**(112108)
7. Astrov, D., Kryukova, N., Zorin, R., Makarov, V., Ozerov, R., Rozhdestvenskij, F., Smirnov, V., Turchaninov, A., Fadeeva, N.: Atomic and magnetic ordering in tantalates of transitional metals $MeTaO_4$ (Me=Ti,V,C,Cr,Te) with a rutile-like structure (1972)
8. Batatia, I., Kovacs, D.P., Simm, G., Ortner, C., Csányi, G.: Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. Advances in Neural Information Processing Systems **35**, 11423–11436 (2022)
9. Boutin, M., Kemper, G.: On reconstructing n-point configurations from the distribution of distances or areas. Advances in Applied Mathematics **32**(4), 709–735 (2004)
10. Bright, M., Cooper, A., Kurlin, V.: Continuous chiral distances for 2-dimensional lattices. Chirality **35**, 920–936 (2023)
11. Bright, M., Cooper, A., Kurlin, V.: Geographic-style maps for 2-dimensional lattices. Acta Cryst A **79**, 1–13 (2023)
12. Caelli, T.: On generating spatial configurations with identical interpoint distance distributions. In: Combinatorial Mathematics VII (1979), pp. 69–75. Springer (2006)
13. Chawla, D.S.: Crystallography databases hunt for fraudulent structures. ACS Central Science **9**, 1853–1855 (2023)
14. Cheetham, A.K., Seshadri, R.: Artificial intelligence driving materials discovery? Chemistry of Materials **36**(8), 3490–3495 (2024)
15. Chisholm, J., Motherwell, S.: Compack: a program for identifying crystal structure similarity using distances. J. Applied Crystal. **38**, 228–231 (2005)
16. Cohen, S., Guibas, L.: The earth mover's distance: Lower bounds and invariance under translation. Tech. rep., Stanford University (1997)
17. Deza, M.M., Deza, E.: Encyclopedia of distances. Springer (2009)
18. Dusson, G., Bachmayr, M., Csányi, G., Drautz, R., Etter, S., van Der Oord, C., Ortner, C.: Atomic cluster expansion: Completeness, efficiency and stability. J Computational Physics **454**, 110946 (2022)
19. Edelsbrunner, H., Heiss, T.: Merge trees of periodic filtrations. arXiv:2408.16575 (2024)
20. Edelsbrunner, H., Heiss, T., Kurlin, V., Smith, P., Wintraecken, M.: The density fingerprint of a periodic point set. In: Proceedings of SoCG, vol. 189, pp. 32:1–32:16 (2021)
21. Efrat, A., Itai, A., Katz, M.J.: Geometry helps in bottleneck matching and related problems. Algorithmica **31**(1), 1–28 (2001)
22. Elkin, Y., Kurlin, V.: Counterexamples expose gaps in the proof of time complexity for cover trees introduced in 2006. In: Top. Data Analysis and Visualization, pp. 9–17 (2022)
23. Elkin, Y., Kurlin, V.: A new near-linear time algorithm for k-nearest neighbor search using a compressed cover tree. In: Int.l Conf. Machine Learning, pp. 9267–9311 (2023)
24. Farhi, B.: Nontrivial lower bounds for the least common multiple of some finite sequences of integers. Journal of Number Theory **125**(2), 393–411 (2007)
25. Gieseke, F., Heinermann, J., Oancea, C., Igel, C.: Buffer kd trees: processing massive nearest neighbor queries on gpus. In: Int. Conf. Machine Learning, pp. 172–180 (2014)
26. Google: (2023). URL `https://deepmind.google/discover/blog/millions-of-new-materials-discovered-with-deep-learning`
27. Grohe, M., Schweitzer, P.: The graph isomorphism problem. Communications of the ACM **63**(11), 128–134 (2020)
28. Grünbaum, Moore: The use of higher-order invariants in the determination of generalized Patterson cyclotomic sets. Acta Cryst A **51**, 310–323 (1995)

29. Hargreaves, C., et al.: The earth mover's distance as a metric for the space of inorganic compositions. Chemistry of Materials **32**, 10610–10620 (2020)
30. Hyde, D.: The sorites paradox. In: Vagueness: A guide, pp. 1–17. Springer (2011)
31. Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., et al.: Commentary: The materials project: A materials genome approach to accelerating materials innovation. APL materials **1**(1) (2013)
32. Juelsholt, M.: Continued challenges in high-throughput materials predictions: Mattergen predicts compounds from the training dataset. chemrxiv-2025-mkls8
33. Kantorovich, L.V.: Mathematical methods of organizing and planning production. Management science **6**(4), 366–422 (1960)
34. Keeping, E.S.: Introduction to statistical inference. Courier Corporation (1995)
35. Kruskal, J.B., Wish, M.: Multidimensional scaling. 11. Sage (1978)
36. Kurlin, V.: A complete isometry classification of 3D lattices. arxiv:2201.10543 (2022)
37. Kurlin, V.: Mathematics of 2D lattices. Found. Comp. Mathematics **24**, 805–863 (2024)
38. Kurlin, V.: Complete and continuous invariants of 1-periodic sequences in polynomial time. SIAM J Mathematics of Data Science (2025, to appear)
39. Lawton, S., Jacobson, R.: The reduced cell and its crystallographic applications. Tech. rep., Ames Lab, Iowa State University (1965)
40. Macdonald, I.G.: Symmetric functions and Hall polynomials. Oxford Univ. Press (1998)
41. McManus, J., Kurlin, V.: Computing the bridge length: the key ingredient in a continuous isometry classification of periodic point sets. Acta Cryst A (2025). DOI 10.1107/S2053273325008253
42. Mémoli, F.: Gromov–Wasserstein distances and the metric approach to object matching. Foundations of Computational Mathematics **11**(4), 417–487 (2011)
43. Merchant, A., Batzner, S., Schoenholz, S.S., Aykol, M., Cheon, G., Cubuk, E.D.: Scaling deep learning for materials discovery. Nature pp. 80–85 (2023)
44. Microsoft: Mattergen repository contains cifs.zip, including the folder 'experimental'. `https://github.com/microsoft/mattergen/tree/main/data-release` (2025)
45. Niggli, P.: Krystallographische und strukturtheoretische Grundbegriffe, vol. 1. Akademische verlagsgesellschaft mbh (1928)
46. Orlin, J.B.: A faster strongly polynomial minimum cost flow algorithm. Operations research **41**(2), 338–350 (1993)
47. Parsaeifard, B., Goedecker, S.: Manifolds of quasi-constant soap and acsf fingerprints and the resulting failure to machine learn four-body interactions. The Journal of Chemical Physics **156**(3), 034302 (2022)
48. Patterson, A.: Homometric structures. Nature **143**, 939–940 (1939)
49. Pauling, L., Shappell, M.D.: The crystal structure of bixbyite and the c-modification of the sesquioxides. Zeitschrift für Kristallographie-Crystalline Materials **75**(1), 128–142 (1930)
50. Pozdnyakov, S.N., Ceriotti, M.: Incompleteness of graph neural networks for points clouds in three dimensions. Machine Learning: Science and Technology **3**(4), 045020 (2022)
51. Pozdnyakov, S.N., Willatt, M.J., Bartók, A.P., Ortner, C., Csányi, G., Ceriotti, M.: Comment on "Manifolds of quasi-constant soap and acsf fingerprints and the resulting failure to machine learn four-body interactions". Journal of Chemical Physics **157**(17) (2022)
52. Rass, S., König, S., Ahmad, S., Goman, M.: Metricizing the euclidean space towards desired distance relations in point clouds. IEEE Trans. Inf. Forensics and Security (2024)
53. Rubner, Y., Tomasi, C., Guibas, L.: The earth mover's distance as a metric for image retrieval. Int. J Computer Vision **40**(2), 99–121 (2000)
54. Sacchi, P., et al.: Same or different – that is the question: identification of crystal forms from crystal structure data. Cryst Eng Comm **22**(43), 7170–7185 (2020)
55. Schoenberg, I.: Remarks to Maurice Frechet's article "Sur la definition axiomatique d'une classe d'espace distances vectoriellement applicable sur l'espace de Hilbert". Annals of Mathematics pp. 724–732 (1935)
56. Seeley, R.T.: Spherical harmonics. Amer. Math. Monthly **73**(4P2), 115–121 (1966)
57. Shen, Y., Jiang, Y., Lin, J., Wang, C., Sun, J.: A general method for searching for homometric structures. Acta Cryst B **78**(1), 14–19 (2022)
58. Shervashidze, N., Schweitzer, P., Van Leeuwen, E.J., Mehlhorn, K., Borgwardt, K.M.: Weisfeiler-Lehman graph kernels. Journal of Machine Learning Research **12**(9) (2011)
59. Szymanski, N., et al.: An autonomous laboratory for the accelerated synthesis of novel materials. Nature pp. 86–91 (2023)
60. Vaserstein, L.N.: Markov processes over denumerable products of spaces, describing large systems of automata. Problemy Peredachi Informatsii **5**(3), 64–72 (1969)

61. Villar, S., et al.: Scalars are universal: equivariant machine learning, structured like classical physics. Advances in Neural Information Processing Systems **34**, 28848–28863 (2021)
62. Weyl, H.: The classical groups. Princeton Univ. Press (1946)
63. Widdowson, D., Kurlin, V.: Resolving the data ambiguity for periodic crystals. Advances in Neural Information Processing Systems **35**, 24625–24638 (2022)
64. Widdowson, D., Kurlin, V.: Geographic-style maps with a local novelty distance help navigate in the material space. Scientific Reports **15**(27588) (2025)
65. Widdowson, D., Kurlin, V.: Pointwise distance distributions for detecting near-duplicates in large materials databases. SIAM J Applied Mathematics, arxiv:2108.04798 (2025)
66. Widdowson, D., Mosca, M.M., Pulido, A., Cooper, A.I., Kurlin, V.: Average minimum distances of periodic point sets - foundational invariants for mapping all periodic crystals. MATCH Commun. Math. Comput. Chem. **87**, 529–559 (2022)
67. Widdowson, D.E., Kurlin, V.A.: Recognizing rigid patterns of unlabeled point clouds by complete and continuous isometry invariants with no false negatives and no false positives. In: Computer Vision and Pattern Recognition, pp. 1275–1284 (2023)
68. Zagorac, D., et al.: Recent developments in the inorganic crystal structure database: theoretical crystal structure data and related features. J Appl Cryst **52**, 918–925 (2019)
69. Zeni, C., et al.: A generative model for inorganic materials design. Nature **639**(8055), 624–632 (2025)
70. Zwart, P., Grosse-Kunstleve, R., Lebedev, A., Murshudov, G., Adams, P.: Surprises and pitfalls arising from (pseudo) symmetry. Acta Cryst. D **64**, 99–107 (2008)

A Appendix: details of hierarchical comparisons across three databases

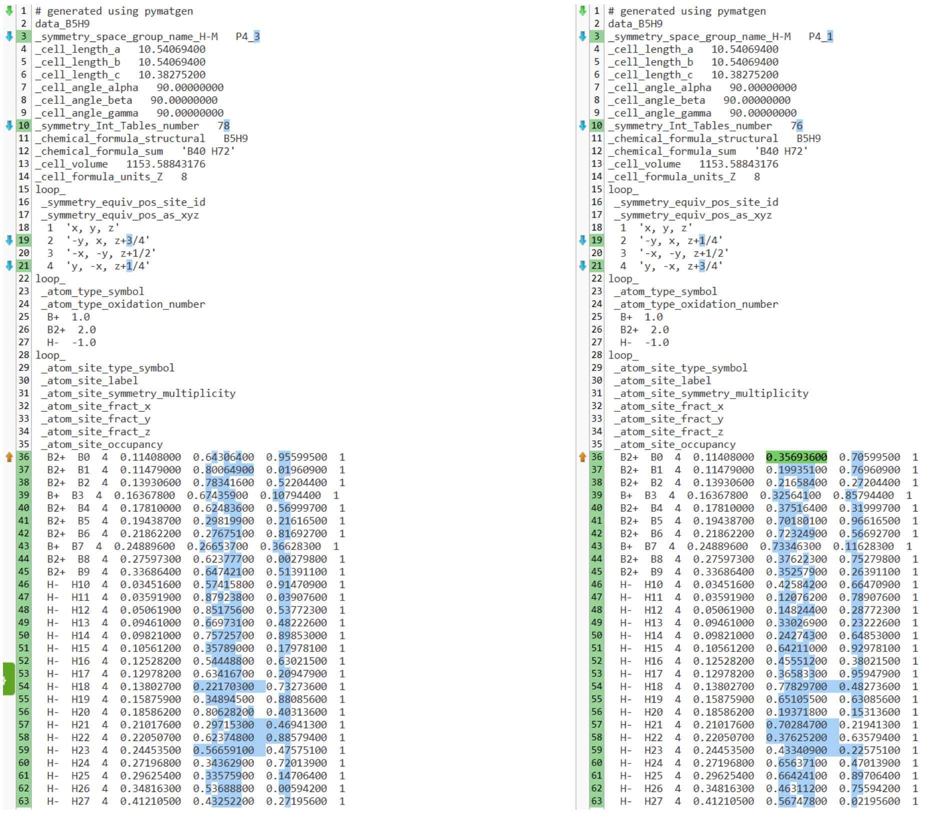
Fig. 10-11 include screenshots (from https://text-compare.com) of different CIFs for the pairs from the last two rows of Table 12. Though distance-based invariants, such as PDA$^{\{h\}}$, cannot distinguish mirror images, our slower metric on isosets [6], which are complete under rigid motion, has approximate values 0.468Å and 0.33552Å, so these mirror images are not related by translations and rotations.



```
# generated using pymatgen
data_Rb2Be2O3
_symmetry_space_group_name_H-M   P4_12_12
_cell_length_a   7.57761200
_cell_length_b   7.57761200
_cell_length_c   7.39162600
_cell_angle_alpha   90.00000000
_cell_angle_beta   90.00000000
_cell_angle_gamma   90.00000000
_symmetry_Int_Tables_number   92
_chemical_formula_structural   Rb2Be2O3
_chemical_formula_sum   'Rb8 Be8 O12'
_cell_volume   424.42867002
_cell_formula_units_Z   4
loop_
 _symmetry_equiv_pos_site_id
 _symmetry_equiv_pos_as_xyz
  1  'x, y, z'
  2  '-y+1/2, x+1/2, z+1/4'
  3  '-x, -y, z+1/2'
  4  'y+1/2, -x+1/2, z+3/4'
  5  'x+1/2, -y+1/2, -z+3/4'
  6  '-y, -x, -z+1/2'
  7  '-x+1/2, y+1/2, -z+1/4'
  8  'y, x, -z'
loop_
 _atom_type_symbol
 _atom_type_oxidation_number
  Rb+  1.0
  Be2+  2.0
  O2-  -2.0
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
  Rb+  Rb0  8  0.00412300  0.24620200  0.13696100  1
  Be2+  Be1  8  0.03254500  0.71957700  0.15807100  1
  O2-  O2  8  0.12838800  0.63329300  0.99372800  1
  O2-  O3  4  0.11258100  0.88741900  0.25000000  1
```

```
# generated using pymatgen
data_Rb2Be2O3
_symmetry_space_group_name_H-M   P4_32_12
_cell_length_a   7.57761200
_cell_length_b   7.57761200
_cell_length_c   7.39162600
_cell_angle_alpha   90.00000000
_cell_angle_beta   90.00000000
_cell_angle_gamma   90.00000000
_symmetry_Int_Tables_number   96
_chemical_formula_structural   Rb2Be2O3
_chemical_formula_sum   'Rb8 Be8 O12'
_cell_volume   424.42867002
_cell_formula_units_Z   4
loop_
 _symmetry_equiv_pos_site_id
 _symmetry_equiv_pos_as_xyz
  1  'x, y, z'
  2  '-y+1/2, x+1/2, z+3/4'
  3  '-x, -y, z+1/2'
  4  'y+1/2, -x+1/2, z+1/4'
  5  'x+1/2, -y+1/2, -z+1/4'
  6  '-y, -x, -z+1/2'
  7  '-x+1/2, y+1/2, -z+3/4'
  8  'y, x, -z'
loop_
 _atom_type_symbol
 _atom_type_oxidation_number
  Rb+  1.0
  Be2+  2.0
  O2-  -2.0
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
  Rb+  Rb0  8  0.00412300  0.24620200  0.86303900  1
  Be2+  Be1  8  0.03254500  0.71957700  0.84192900  1
  O2-  O2  8  0.12838800  0.63329300  0.00627200  1
  O2-  O3  4  0.11258100  0.88741900  0.75000000  1
```

Fig. 10 The CIFs of the MP entries mp-771976 (left) and mp-1345479 (right) have identical cells and different coordinates, which can be matched under reflection $(x, y, z) \mapsto (x, y, 1 - z)$.



```
# generated using pymatgen
data_B5H9
_symmetry_space_group_name_H-M   P4_3
_cell_length_a   10.54069400
_cell_length_b   10.54069400
_cell_length_c   10.38275200
_cell_angle_alpha   90.00000000
_cell_angle_beta   90.00000000
_cell_angle_gamma   90.00000000
_symmetry_Int_Tables_number   78
_chemical_formula_structural   B5H9
_chemical_formula_sum   'B40 H72'
_cell_volume   1153.58843176
_cell_formula_units_Z   8
loop_
 _symmetry_equiv_pos_site_id
 _symmetry_equiv_pos_as_xyz
  1  'x, y, z'
  2  '-y, x, z+3/4'
  3  '-x, -y, z+1/2'
  4  'y, -x, z+1/4'
loop_
 _atom_type_symbol
 _atom_type_oxidation_number
  B+  1.0
  B2+  2.0
  H-  -1.0
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
  B2+  B0  4  0.11408000  0.64306400  0.95599500  1
  B2+  B1  4  0.11479000  0.80064900  0.01960900  1
  B2+  B2  4  0.13930600  0.78341600  0.52204400  1
  B+  B3  4  0.16367800  0.67435900  0.10794400  1
  B2+  B4  4  0.17810000  0.62483600  0.56999700  1
  B2+  B5  4  0.19438700  0.29819900  0.21616500  1
  B2+  B6  4  0.21862200  0.27675100  0.81692700  1
  B+  B7  4  0.24889600  0.26653700  0.36628300  1
  B2+  B8  4  0.27597300  0.62377700  0.00279800  1
  B2+  B9  4  0.33686400  0.64742100  0.51391100  1
  H-  H10  4  0.03451600  0.57415800  0.91470900  1
  H-  H11  4  0.03591900  0.87923800  0.03907600  1
  H-  H12  4  0.05061900  0.85175600  0.53772300  1
  H-  H13  4  0.09461000  0.66973100  0.48222600  1
  H-  H14  4  0.09821000  0.75725700  0.89853000  1
  H-  H15  4  0.10561200  0.35789000  0.17978100  1
  H-  H16  4  0.12528200  0.54448800  0.63021500  1
  H-  H17  4  0.12978200  0.63416700  0.20947900  1
  H-  H18  4  0.13802700  0.22170300  0.73273600  1
  H-  H19  4  0.15875900  0.34894500  0.88085600  1
  H-  H20  4  0.18586200  0.80628200  0.40313600  1
  H-  H21  4  0.21017600  0.29715300  0.46941300  1
  H-  H22  4  0.22050700  0.62374800  0.88579400  1
  H-  H23  4  0.24453500  0.56659100  0.47575100  1
  H-  H24  4  0.27196800  0.34362900  0.72013900  1
  H-  H25  4  0.29625400  0.33575900  0.14706400  1
  H-  H26  4  0.34816300  0.53688800  0.00594200  1
  H-  H27  4  0.41210500  0.43252200  0.27195600  1
```

```
# generated using pymatgen
data_B5H9
_symmetry_space_group_name_H-M   P4_1
_cell_length_a   10.54069400
_cell_length_b   10.54069400
_cell_length_c   10.38275200
_cell_angle_alpha   90.00000000
_cell_angle_beta   90.00000000
_cell_angle_gamma   90.00000000
_symmetry_Int_Tables_number   76
_chemical_formula_structural   B5H9
_chemical_formula_sum   'B40 H72'
_cell_volume   1153.58843176
_cell_formula_units_Z   8
loop_
 _symmetry_equiv_pos_site_id
 _symmetry_equiv_pos_as_xyz
  1  'x, y, z'
  2  '-y, x, z+1/4'
  3  '-x, -y, z+1/2'
  4  'y, -x, z+3/4'
loop_
 _atom_type_symbol
 _atom_type_oxidation_number
  B+  1.0
  B2+  2.0
  H-  -1.0
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
  B2+  B0  4  0.11408000  0.35693600  0.70599500  1
  B2+  B1  4  0.11479000  0.19935100  0.76960900  1
  B2+  B2  4  0.13930600  0.21658400  0.27204400  1
  B+  B3  4  0.16367800  0.32564100  0.85794400  1
  B2+  B4  4  0.17810000  0.37516400  0.31999700  1
  B2+  B5  4  0.19438700  0.70180100  0.96616500  1
  B2+  B6  4  0.21862200  0.72324900  0.56692700  1
  B+  B7  4  0.24889600  0.73346300  0.11628300  1
  B2+  B8  4  0.27597300  0.37622300  0.75279800  1
  B2+  B9  4  0.33686400  0.35257900  0.26391100  1
  H-  H10  4  0.03451600  0.42584200  0.66470900  1
  H-  H11  4  0.03591900  0.12076200  0.78907600  1
  H-  H12  4  0.05061900  0.14824400  0.28772300  1
  H-  H13  4  0.09461000  0.33026900  0.23222600  1
  H-  H14  4  0.09821000  0.24274300  0.64853000  1
  H-  H15  4  0.10561200  0.64211000  0.92978100  1
  H-  H16  4  0.12528200  0.45551200  0.38021500  1
  H-  H17  4  0.12978200  0.36583300  0.95947900  1
  H-  H18  4  0.13802700  0.77829700  0.48273600  1
  H-  H19  4  0.15875900  0.65105500  0.63085600  1
  H-  H20  4  0.18586200  0.19371800  0.15313600  1
  H-  H21  4  0.21017600  0.70284700  0.21941300  1
  H-  H22  4  0.22050700  0.37625200  0.63579400  1
  H-  H23  4  0.24453500  0.43340900  0.22575100  1
  H-  H24  4  0.27196800  0.65637100  0.47013900  1
  H-  H25  4  0.29625400  0.66424100  0.89706400  1
  H-  H26  4  0.34816300  0.46311200  0.75594200  1
  H-  H27  4  0.41210500  0.56747800  0.02195600  1
```

Fig. 11 The CIFs of the MP entries mp-29783 (left) and mp-1338697 (right) have identical compositions and unit cells, but different fractional coordinates of atoms, which can be exactly matched under reflection $(x, y, z) \mapsto (x, 1-y, z-0.25)$, where $z-0.25$ is taken modulo 1.

Tables 13–18 include the running times and numbers of compared pairs and resulting unique entries for two versions of (L_∞ and RMS-based) distances between the invariants $\mathrm{ADA}(S; 100)$, $\mathrm{PDA}(S; 100)$, $\mathrm{ADA}^{(2)}(S; 100)$, $\mathrm{PDA}^{(2)}(S; 100)$.

The smallest threshold 10^{-10}Å in Table 13 is considered a floating-point error, and the resulting pairs of geometric duplicates are available by request. At the higher threshold 10^{-6}Å in Table 14 only a few extra duplicates appear. However, the further tables show that the numbers of near-duplicates substantially grow for larger thresholds up to 0.01Å, which is still considered experimental noise.

In the bottom section of Table 13, the number 3248 of geometric duplicates in the GNoME was previously found in [5, Table 1] by comparisons of CIFs by numerical data (unit cell parameters and atomic coordinates) than by invariants.

Table 13 Number of pairs, unique entries (and as a percentage of the database size), and running time (in seconds) taken at each stage of the duplicate finding process, using the threshold 10^{-10}Å and $k = 100$ atomic neighbors with sequentially stronger invariants.

		L_∞				RMS			
		ADA	PDA	ADA$^{(2)}$	PDA$^{(2)}$	ADA	PDA	ADA$^{(2)}$	PDA$^{(2)}$
ICSD vs ICSD	Pairs	8994	8994	8994	8994	8994	8994	8994	8994
	Entries	9452	9452	9452	9452	9452	9452	9452	9452
	%	8.05	8.05	8.05	8.05	8.05	8.05	8.05	8.05
	Time (s)	3.1	2.6	0.0	2.2	6.5	1.8	0.0	1.9
ICSD vs MP	Pairs	33	33	33	33	33	33	33	33
	Entries	33	33	33	33	33	33	33	33
	%	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
	Time (s)	4.3	0.0	0.0	0.0	4.9	0.0	0.0	0.0
ICSD vs GNoME	Pairs	0	0	0	0	0	0	0	0
	Entries	0	0	0	0	0	0	0	0
	%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Time (s)	11.2	0.0	0.0	0.0	10.5	0.0	0.0	0.0
MP vs ICSD	Pairs	33	33	33	33	33	33	33	33
	Entries	15	15	15	15	15	15	15	15
	%	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	Time (s)	3.6	0.0	0.0	0.0	4.7	0.0	0.0	0.0
MP vs MP	Pairs	5	5	5	5	5	5	5	5
	Entries	10	10	10	10	10	10	10	10
	%	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	Time (s)	5.1	0.0	0.0	0.0	7.6	0.0	0.0	0.0
MP vs GNoME	Pairs	4	4	4	4	4	4	4	4
	Entries	4	4	4	4	4	4	4	4
	%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Time (s)	8.7	0.0	0.0	0.0	11.2	0.0	0.0	0.0
GNoME vs ICSD	Pairs	0	0	0	0	0	0	0	0
	Entries	0	0	0	0	0	0	0	0
	%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Time (s)	4.3	0.0	0.0	0.0	8.5	0.0	0.0	0.0
GNoME vs MP	Pairs	4	4	4	4	4	4	4	4
	Entries	4	4	4	4	4	4	4	4
	%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Time (s)	5.2	0.0	0.0	0.0	9.7	0.0	0.0	0.0
GNoME vs GNoME	Pairs	1852	1852	1852	1852	1852	1852	1852	1852
	Entries	3248	3248	3248	3248	3248	3248	3248	3248
	%	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84
	Time (s)	14.5	0.5	0.0	0.5	21.4	0.5	0.0	0.5

Table 14 Number of pairs, unique entries (and as a percentage of the database size), and running time (in seconds) taken at each stage of the duplicate finding process, using the threshold 10^{-6}Å and $k = 100$ atomic neighbors with sequentially stronger invariants.

		L_∞				RMS			
		ADA	PDA	ADA$^{(2)}$	PDA$^{(2)}$	ADA	PDA	ADA$^{(2)}$	PDA$^{(2)}$
ICSD vs ICSD	Pairs	8999	8998	8996	8995	9036	9027	9003	8999
	Entries	9462	9460	9456	9454	9510	9493	9470	9462
	%	8.05	8.05	8.05	8.05	8.09	8.08	8.06	8.05
	Time (s)	3.1	1.9	0.0	1.9	8.3	5.2	0.1	4.7
ICSD vs MP	Pairs	102	101	53	53	310	283	168	163
	Entries	102	101	53	53	292	265	159	154
	%	0.09	0.09	0.05	0.05	0.25	0.23	0.14	0.13
	Time (s)	3.3	0.1	0.0	0.0	6.2	0.2	0.0	0.1
ICSD vs GNoME	Pairs	3	3	1	1	15	15	8	8
	Entries	3	3	1	1	15	15	8	8
	%	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01
	Time (s)	8.8	0.0	0.0	0.0	13.5	0.0	0.0	0.0
MP vs ICSD	Pairs	102	101	53	53	310	283	168	163
	Entries	57	56	26	26	186	169	91	87
	%	0.04	0.04	0.02	0.02	0.12	0.11	0.06	0.06
	Time (s)	2.8	0.0	0.0	0.0	4.9	0.1	0.0	0.0
MP vs MP	Pairs	91	90	40	40	290	279	148	148
	Entries	182	180	80	80	558	537	293	293
	%	0.12	0.12	0.05	0.05	0.36	0.35	0.19	0.19
	Time (s)	7.1	0.1	0.0	0.0	8.7	0.3	0.0	0.1
MP vs GNoME	Pairs	12	12	10	10	44	42	22	22
	Entries	12	12	10	10	43	41	21	21
	%	0.01	0.01	0.01	0.01	0.03	0.03	0.01	0.01
	Time (s)	12.3	0.0	0.0	0.0	19.1	0.2	0.0	0.1
GNoME vs ICSD	Pairs	3	3	1	1	15	15	8	8
	Entries	3	3	1	1	13	13	8	8
	%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Time (s)	4.7	0.0	0.0	0.0	13.1	0.1	0.0	0.0
GNoME vs MP	Pairs	12	12	10	10	44	42	22	22
	Entries	12	12	10	10	40	38	20	20
	%	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01
	Time (s)	8.0	0.0	0.0	0.0	10.5	0.1	0.0	0.0
GNoME vs GNoME	Pairs	2511	2490	2489	2482	2547	2516	2513	2504
	Entries	4406	4367	4365	4351	4477	4416	4410	4392
	%	1.14	1.13	1.13	1.13	1.16	1.15	1.15	1.14
	Time (s)	18.4	3.3	0.0	2.5	30.2	3.5	0.0	2.6

Table 15 Number of pairs, unique entries (and as a percentage of the database size), and running time (in seconds) taken at each stage of the duplicate finding process, using the threshold 10^{-5}Å and $k = 100$ atomic neighbors with sequentially stronger invariants.

		L_∞				RMS			
		ADA	PDA	ADA$^{(2)}$	PDA$^{(2)}$	ADA	PDA	ADA$^{(2)}$	PDA$^{(2)}$
ICSD vs ICSD	Pairs	9173	9143	9041	9036	10400	10190	9397	9339
	Entries	9661	9620	9511	9509	10465	10294	9843	9779
	%	8.22	8.19	8.10	8.09	8.91	8.76	8.38	8.32
	Time (s)	4.9	5.4	0.1	4.7	9.5	5.7	0.1	5.0
ICSD vs MP	Pairs	788	774	291	291	2641	2486	1245	1199
	Entries	702	690	273	273	1946	1848	1066	1021
	%	0.60	0.59	0.23	0.23	1.66	1.57	0.91	0.87
	Time (s)	5.0	0.6	0.0	0.2	7.3	1.6	0.0	0.7
ICSD vs GNoME	Pairs	67	61	18	18	191	171	89	85
	Entries	67	61	18	18	173	159	88	84
	%	0.06	0.05	0.02	0.02	0.15	0.14	0.07	0.07
	Time (s)	14.5	0.1	0.0	0.0	16.6	0.1	0.0	0.1
MP vs ICSD	Pairs	788	774	291	291	2641	2486	1245	1199
	Entries	490	477	176	176	1628	1537	788	764
	%	0.32	0.31	0.11	0.11	1.06	1.00	0.51	0.50
	Time (s)	4.4	0.3	0.0	0.1	7.0	0.8	0.0	0.4
MP vs MP	Pairs	821	792	289	283	2746	2659	1309	1285
	Entries	1430	1378	557	545	3740	3620	2104	2067
	%	0.93	0.90	0.36	0.36	2.44	2.36	1.37	1.35
	Time (s)	6.3	0.7	0.0	0.2	9.8	1.5	0.0	0.7
MP vs GNoME	Pairs	116	111	44	42	381	368	170	169
	Entries	110	106	43	41	346	333	162	161
	%	0.07	0.07	0.03	0.03	0.23	0.22	0.11	0.11
	Time (s)	14.9	0.1	0.0	0.0	17.6	0.3	0.0	0.1
GNoME vs ICSD	Pairs	67	61	18	18	191	171	89	85
	Entries	41	36	14	14	123	115	57	55
	%	0.01	0.01	0.00	0.00	0.03	0.03	0.01	0.01
	Time (s)	6.6	0.1	0.0	0.0	12.7	0.2	0.0	0.1
GNoME vs MP	Pairs	116	111	44	42	381	368	170	169
	Entries	99	94	40	38	271	262	139	138
	%	0.03	0.02	0.01	0.01	0.07	0.07	0.04	0.04
	Time (s)	8.1	0.1	0.0	0.0	12.7	0.3	0.0	0.1
GNoME vs GNoME	Pairs	2640	2564	2549	2524	2721	2678	2648	2606
	Entries	4658	4506	4479	4432	4809	4726	4674	4590
	%	1.21	1.17	1.16	1.15	1.25	1.23	1.21	1.19
	Time (s)	22.1	3.7	0.0	2.7	28.5	3.4	0.0	2.6

Table 16 Number of pairs, unique entries (and as a percentage of the database size), and running time (in seconds) taken at each stage of the duplicate finding process, using the threshold 10^{-4}Å and $k = 100$ atomic neighbors with sequentially stronger invariants.

		L_∞				RMS			
		ADA	PDA	ADA$^{(2)}$	PDA$^{(2)}$	ADA	PDA	ADA$^{(2)}$	PDA$^{(2)}$
ICSD vs ICSD	Pairs	14009	13669	10360	10353	33936	31231	18669	18021
	Entries	12045	11857	10425	10411	15732	15004	13061	12845
	%	10.25	10.09	8.87	8.86	13.39	12.77	11.12	10.93
	Time (s)	4.5	6.7	0.1	5.0	5.5	10.8	0.1	6.7
ICSD vs MP	Pairs	7304	7087	2586	2586	26146	24318	11831	11481
	Entries	3795	3729	1910	1910	6852	6645	4833	4708
	%	3.23	3.17	1.63	1.63	5.83	5.66	4.11	4.01
	Time (s)	4.6	3.5	0.0	1.3	9.3	14.2	0.2	7.3
ICSD vs GNoME	Pairs	504	481	184	184	1892	1725	844	820
	Entries	434	416	170	170	1285	1173	656	636
	%	0.37	0.35	0.14	0.14	1.09	1.00	0.56	0.54
	Time (s)	16.0	0.9	0.0	0.3	18.0	2.3	0.5	1.0
MP vs ICSD	Pairs	7304	7087	2586	2586	26146	24318	11831	11481
	Entries	3881	3828	1595	1595	7275	7115	5266	5182
	%	2.53	2.50	1.04	1.04	4.75	4.64	3.44	3.38
	Time (s)	3.1	4.2	0.0	1.7	6.3	11.1	0.1	4.9
MP vs MP	Pairs	7843	7710	2718	2718	27091	26019	12634	12426
	Entries	6196	6121	3709	3709	8406	8176	7107	7018
	%	4.04	3.99	2.42	2.42	5.49	5.34	4.64	4.58
	Time (s)	4.2	3.5	0.0	1.4	8.9	9.4	0.1	5.3
MP vs GNoME	Pairs	1083	1067	378	377	3855	3689	1743	1713
	Entries	883	873	344	343	2470	2368	1312	1289
	%	0.58	0.57	0.22	0.22	1.61	1.55	0.86	0.84
	Time (s)	12.7	1.6	0.0	0.6	13.5	2.6	0.0	1.3
GNoME vs ICSD	Pairs	504	481	184	184	1892	1725	844	820
	Entries	292	280	122	122	589	565	403	393
	%	0.08	0.07	0.03	0.03	0.15	0.15	0.10	0.10
	Time (s)	6.1	0.7	0.0	0.3	10.9	1.5	0.0	0.7
GNoME vs MP	Pairs	1083	1067	378	377	3855	3689	1743	1713
	Entries	456	453	269	268	629	601	516	507
	%	0.12	0.12	0.07	0.07	0.16	0.16	0.13	0.13
	Time (s)	7.1	1.4	0.0	0.6	12.6	2.5	0.0	1.0
GNoME vs GNoME	Pairs	2859	2804	2741	2719	3461	3144	2955	2901
	Entries	5038	4941	4845	4808	5830	5366	5159	5070
	%	1.31	1.28	1.26	1.25	1.51	1.39	1.34	1.32
	Time (s)	14.6	4.7	0.1	3.6	28.5	4.1	0.0	3.8

Table 17 Number of pairs, unique entries (and as a percentage of the database size), and running time (in seconds) taken at each stage of the duplicate finding process, using the threshold 10^{-3}Å and $k = 100$ atomic neighbours with sequentially stronger invariants.

		L_∞				RMS			
		ADA	PDA	ADA$^{(2)}$	PDA$^{(2)}$	ADA	PDA	ADA$^{(2)}$	PDA$^{(2)}$
ICSD vs ICSD	Pairs	82033	78770	33672	33314	258044	227421	119734	115102
	Entries	20066	18893	16352	16052	27894	24479	22104	21073
	%	17.08	16.08	13.92	13.66	23.74	20.84	18.81	17.94
	Time (s)	4.7	27.0	0.5	14.2	9.8	75.9	1.0	36.7
ICSD vs MP	Pairs	75607	73226	25476	25416	275005	251022	121371	117451
	Entries	9455	9320	6759	6722	11921	11406	10130	9975
	%	8.05	7.93	5.75	5.72	10.15	9.71	8.62	8.49
	Time (s)	4.0	23.6	0.4	9.6	9.0	88.7	3.3	38.4
ICSD vs GNoME	Pairs	6255	5751	1881	1853	27803	23438	10315	9543
	Entries	3036	2755	1279	1263	5814	5359	3952	3637
	%	2.58	2.35	1.09	1.08	4.95	4.56	3.36	3.10
	Time (s)	12.5	3.4	0.0	1.3	15.7	10.5	0.8	5.1
MP vs ICSD	Pairs	75607	73226	25476	25416	275005	251022	121371	117451
	Entries	9014	8884	7240	7228	11124	10415	9414	9275
	%	5.88	5.80	4.72	4.72	7.26	6.80	6.14	6.05
	Time (s)	4.4	25.0	0.4	10.2	10.3	75.1	2.0	36.1
MP vs MP	Pairs	79298	77364	26760	26703	284625	267243	127150	124642
	Entries	10482	9973	8369	8301	14386	12798	10962	10511
	%	6.84	6.51	5.46	5.42	9.39	8.35	7.15	6.86
	Time (s)	6.0	25.1	0.4	9.5	13.2	82.5	1.3	37.6
MP vs GNoME	Pairs	12057	11622	3890	3864	44774	41267	19263	18717
	Entries	5011	4823	2475	2460	7959	7345	6016	5821
	%	3.27	3.15	1.62	1.61	5.19	4.79	3.93	3.80
	Time (s)	13.8	5.7	0.1	2.5	21.6	16.0	0.3	8.2
GNoME vs ICSD	Pairs	6255	5751	1881	1853	27803	23438	10315	9543
	Entries	802	760	603	589	1224	1059	838	793
	%	0.21	0.20	0.16	0.15	0.32	0.28	0.22	0.21
	Time (s)	6.1	3.9	0.0	1.5	16.5	10.6	0.1	4.1
GNoME vs MP	Pairs	12057	11622	3890	3864	44774	41267	19263	18717
	Entries	930	848	638	625	1655	1317	981	906
	%	0.24	0.22	0.17	0.16	0.43	0.34	0.25	0.24
	Time (s)	7.3	5.8	0.1	2.4	18.9	15.3	0.2	6.5
GNoME vs GNoME	Pairs	9932	4542	3595	3284	74039	14086	9894	5781
	Entries	13720	6889	6016	5581	49640	14993	12992	8049
	%	3.56	1.79	1.56	1.45	12.90	3.89	3.38	2.09
	Time (s)	19.9	11.1	0.0	4.5	44.7	51.0	0.1	8.7

Table 18 Number of pairs, unique entries (and as a percentage of the database size), and running time (in seconds) taken at each stage of the duplicate finding process, using the threshold 10^{-2}Å and $k = 100$ atomic neighbors with sequentially stronger invariants.

		L_∞				RMS			
		ADA	PDA	ADA$^{(2)}$	PDA$^{(2)}$	ADA	PDA	ADA$^{(2)}$	PDA$^{(2)}$
ICSD vs ICSD	Pairs	687635	631318	268187	259169	2342691	1976021	1047622	966117
	Entries	46723	37356	33746	30831	68312	53203	50481	43508
	%	39.77	31.80	28.72	26.24	58.15	45.29	42.97	37.03
	Time (s)	9.1	193.4	3.5	81.6	54.5	581.1	6.7	251.7
ICSD vs MP	Pairs	851607	792015	273619	269722	3467416	2888095	1359712	1274212
	Entries	19875	15834	12968	12398	47467	28057	23532	18785
	%	16.92	13.48	11.04	10.55	40.40	23.88	20.03	15.99
	Time (s)	10.0	209.3	4.8	60.6	60.6	819.0	10.7	316.2
ICSD vs GNoME	Pairs	164605	125416	35516	30837	1307587	779813	299214	223812
	Entries	10637	9094	6460	6079	25425	16358	12792	10833
	%	9.05	7.74	5.50	5.17	21.64	13.92	10.89	9.22
	Time (s)	15.3	49.8	1.4	12.8	49.6	323.2	2.6	71.3
MP vs ICSD	Pairs	851607	792015	273619	269722	3467416	2888095	1359712	1274212
	Entries	17575	14364	11759	11156	38866	23146	19735	16263
	%	11.47	9.37	7.67	7.28	25.36	15.10	12.88	10.61
	Time (s)	10.0	227.8	6.1	76.4	76.0	794.3	10.3	310.4
MP vs MP	Pairs	903434	828727	285041	278739	3906101	3071761	1404598	1324840
	Entries	28806	19177	16067	14430	66908	34277	30425	23452
	%	18.80	12.51	10.49	9.42	43.66	22.37	19.86	15.30
	Time (s)	13.4	259.7	5.1	84.1	110.1	1417.2	11.5	335.9
MP vs GNoME	Pairs	202503	156999	51103	47040	1646545	928066	364659	291505
	Entries	13362	10681	8411	7894	29364	18680	15365	12422
	%	8.72	6.97	5.49	5.15	19.16	12.19	10.03	8.11
	Time (s)	16.7	62.7	1.0	14.4	61.9	441.8	3.4	87.8
GNoME vs ICSD	Pairs	164605	125416	35516	30837	1307587	779813	299214	223812
	Entries	4702	2515	1624	1374	60377	11310	6923	3631
	%	1.22	0.65	0.42	0.36	15.68	2.94	1.80	0.94
	Time (s)	8.6	47.6	0.8	13.1	103.2	326.8	2.7	70.2
GNoME vs MP	Pairs	202503	156999	51103	47040	1646545	928066	364659	291505
	Entries	11124	3401	2282	1733	97553	19661	14347	5792
	%	2.89	0.88	0.59	0.45	25.34	5.11	3.73	1.50
	Time (s)	9.8	61.5	0.8	14.7	116.5	439.3	3.6	91.3
GNoME vs GNoME	Pairs	1815980	174478	123171	39487	30732727	2059788	1726547	421833
	Entries	197340	82859	73733	35315	326550	216030	208265	127820
	%	51.27	21.53	19.15	9.17	84.83	56.12	54.10	33.21
	Time (s)	33.5	880.8	0.9	67.8	549.5	21659.9	12.7	1061.5