Algorithmic challenges of computational geometry motivated by applications in crystallography

Marco Mosca
Materials Innovation Factory and Department of Computer Science, University of Liverpool, UK
m.m.mosca@liverpool.ac.uk

Vitaliy Kurlin
Materials Innovation Factory and Department of Computer Science, University of Liverpool, UK
vitaliy.kurlin@liverpool.ac.uk

Abstract

This paper proposes practically important algorithms in the new area of continuous periodic geometry extending discrete and computational geometry. The main objects are periodic sets defined by periodically translating a finite collection of points along given basis vectors. The resulting periodic sets model any periodic crystals and are naturally equivalent modulo rigid motions, also called isometries, which preserve distances between points. The discovery of new crystals is still based on random searches, because there was no mathematically justified way to compare periodic sets, for example by using a distance satisfying metric axioms. Traditionally crystals were distinguished by discrete invariants such as symmetry groups that are unstable under atomic vibrations. However, large simulated datasets contain too many nearly identical crystals obtained as slightly different approximations to local minima of an energy function. The key results are the first stable-under-noise invariants of finite and periodic sets, which are conjectured to be complete modulo isometries.

2012 ACM Subject Classification Computational Geometry

Keywords and phrases computational geometry, periodic sets, isometry, stability under noise

Digital Object Identifier 10.4230/LIPIcs...

Funding *Vitaliy Kurlin*: EPSRC “Application-driven Topological Data Analysis”, EP/R018472/1.

1 Importance of geometric algorithms for modern crystallography

This introduction motivates new algorithmic challenges in crystallography for experts in computational geometry and related areas. Section 2 will rigorously introduce the underlying mathematical concepts, state the key problem and overview a new algorithmic approach.

A crystal consists of periodically repeated unit cells (boxes or parallelepipeds) containing a finite collection of atoms or molecules, see Fig. 1 and mathematical details in Definition 1. Different unit cells can define physically identical crystals, which is the key problem here.

The state-of-the-art method to discover useful crystals is the Crystal Structure Prediction (CSP), which aims to predict a thermodynamically stable arrangement of given atoms or molecules. The CSP was pioneered in 1960s when ball models of atoms were physically shaken in a box until they settle in a stable configuration [9]. Nowadays this physical shaking is simulated by supercomputers, which start from millions of almost random arrangements and minimize a complicated energy function. This energy has no simple analytic expression and depends on (theoretically infinite) interactions between atoms within a periodic crystal.

A typical CSP software outputs thousands of approximate local minima of this energy, i.e. simulated crystals whose local perturbations are unlikely to produce more stable arrangements. All these simulated crystals are visualized by an *energy landscape* representing each crystal by an isolated point with two coordinates (density,energy), see Fig. 1. Here the *density* is the molecular weight within a periodically repeated cell divided by the volume of this cell.

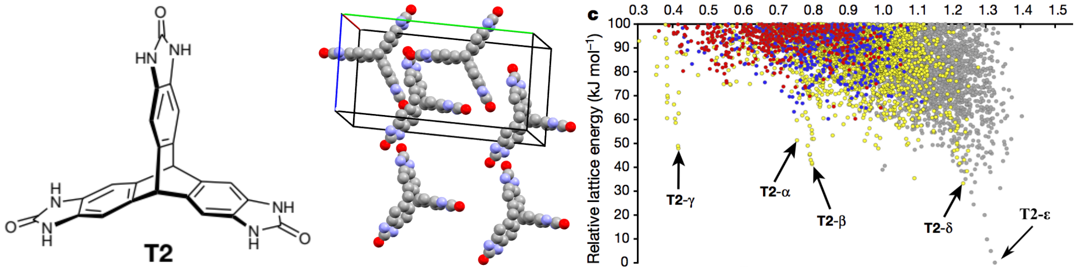

Figure 1 **1st**: the CSP starts from a chemical composition (a molecule or a set of atoms) selected by materials scientists. **2nd**: millions of initial almost random arrangements are iteratively perturbed to minimize an energy. **3rd**: a typical CSP outputs thousands of simulated crystals often visualized as an energy landscape [19, Fig. 2d] , where every crystal has two coordinates (density,energy). This landscape 'hints' at deep minima in downward spikes. The past manual matching of five synthesized crystals to simulated predictions is automated by the new invariant-based algorithm in section 7.

The key challenge of the CSP is the *embarrassment of over-prediction* when the state-of-the-art optimization outputs too many approximate local minima [18]. Materials scientists expect only few different stable crystals (called *metastable polymorphs*) at deep local minima of a complicated energy function on a continuous space of all potential crystals, see Fig. 2.

Any energy landscape should be post-processed to remove numerous near duplicate crystals not to waste even more time on predicting other chemical properties by further often slower simulations; and (2) identify really different crystals separated by high energy barriers. If all local minima are shallow, there is no chance to synthesize a stable crystal and one should change a chemical composition despite optimizing millions of simulated crystals [19].

Since both problems above remain unresolved, even more supercomputer's time (12 weeks [19] in the case of Fig. 1) is spent on predicting target properties of crystals for applications. Since all simulated crystals have the same chemical compositions, they can be distinguished only by their geometry. Geometrically different crystals often have different properties such as solubility, which is vitally important in the pharmaceutical industry. In 1998 manufacturing the HIV drug ritonavir (branded as Kaletra) accidentally produced a more stable but much less soluble polymorph, which has made the drug useless and put many lives at risk [12].

The energy landscape in the last picture of Fig. 1 was lucky due to downward spikes that hinted at deep local minima separated by high energy barriers. This landscape is only a discrete sample from an unknown underlying space of all crystals. To parameterize such a space as a geographic map of an unknown planet, we need to uniquely name each crystal so that different crystals have different names, and similar crystals have similar names.

Here are the key questions: which periodic crystals should be considered equivalent? If not equivalence, how similar are they? Since most solid crystals are rigid, the most natural equivalence is a rigid motion, which is a composition of translations and rotations in $\mathbb{R}^3$. A bit wider concept of an *isometry* also includes reflections with respect to planes in $\mathbb{R}^3$.

Though crystals were studied for centuries, their classification modulo isometries has not been resolved in a satisfactory way that would help to reliably filter out nearly identical crystals from energy landscapes. Most crystal descriptors such as symmetry groups are unstable under atomic vibrations. All atoms vibrate in front of your eyes unless you are reading at the absolute zero temperature. The stability of new invariants in Theorems 12, 14 will help 'join the dots' as in Fig. 2 and automatically identify most stable crystals.

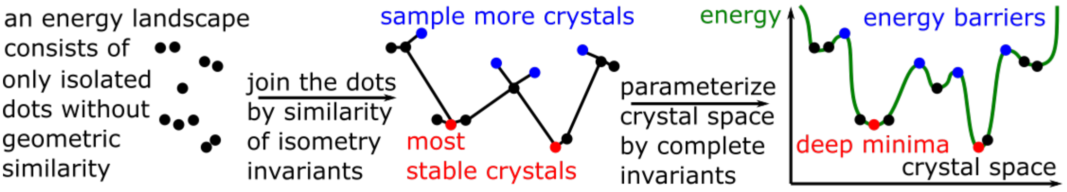
an energy landscape consists of only isolated dots without geometric similarity
join the dots by similarity of isometry invariants
sample more crystals
most stable crystals
parameterize crystal space by complete invariants
energy
energy barriers
deep minima
crystal space

Figure 2 Left: a current energy landscape is a list of simulated crystals. **Middle**: isometry invariants will 'join the dots' and sample a crystal space to find energy barriers. **Right**: a 'mapped' energy landscape (the energy function over a space of crystals) with highlighted deep minima (most stable crystals in red), energy barriers (blue) and other approximations to local minima (black).

2 Algorithmic problems for periodic sets and an overview of results

This section first introduces periodic sets that model all crystals and then states the algorithmic problems for their stable-under-nois classification. In the Euclidean space $\mathbb{R}^n$, any point $p \in \mathbb{R}^n$ can be represented by the vector $\vec{p}$ from the origin of $\mathbb{R}^n$ to the point p. The symbol $\vec{p}$ will also denote the class of all equal vectors that have equal coordinates. The *Euclidean* distance between points $p, q \in \mathbb{R}^n$ is denoted by $|pq| = |\vec{p} - \vec{q}|$. For a standard orthonormal basis $\vec{e}_1, \dots, \vec{e}_n$, the lattice $\mathbb{Z}^n \subset \mathbb{R}^n$ consists of all points with integer coordinates.

Definition 1 below models all atoms in crystals as zero-sized points, which is enough for their isometry classification. To model real atoms of non-zero sizes, one can add combinatorial labels for elements such as C for carbon, O for oxygen etc. Geometrically, atoms can be modeled as weighted points, i.e. balls of different (usually van der Waals) radii. This paper uses the term *periodic set* in any dimension, while *periodic crystals* refer only to dimension 3.

▶ **Definition 1** (a lattice, a periodic set). A *lattice* Λ in $\mathbb{R}^n$ consists of all linear combinations $\sum\limits_{i=1}^{n} \lambda_i \vec{v}_i$ with integer coefficients $\lambda_i \in \mathbb{Z}$. Here the vectors $\vec{v}_1, \dots, \vec{v}_n$ should form a *basis* so that if $\sum\limits_{i=1}^{n} \lambda_i \vec{v}_i = \vec{0}$ for some real λ_i, then all $\lambda_i = 0$. A *periodic set* (or a *crystal*) consists of a lattice with a basis $\vec{v}_1, \dots, \vec{v}_n$ and a *motif* M of finitely many unlabelled points $p_1, \dots, p_m$ (representing molecules, atoms or ions) in the *unit cell* $U(\vec{v}_1, \dots, \vec{v}_n) = \left\{ \sum\limits_{i=1}^{n} \lambda_i \vec{v}_i \, : \, \lambda_i \in [0,1] \right\}$, which is the parallepiped spanned by $\vec{v}_1, \dots, \vec{v}_n$ along its edges. ■

The two pictures in the top left of Fig. 3 show two lattices with a square unit cell and a single black point in a motif. Though the lattices look different at first glance, they are related by a rotation through $\frac{\pi}{4}$, hence are isometric, see details in Definition 17. Any periodic set can be considered as the *Minkowski* sum of a lattice and a motif, i.e. $S = \Lambda + M = \{\vec{u} + \vec{v} \, : \, u \in \Lambda, v \in M\}$. Any periodic set is a finite union of translates of Λ.

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

The two periodic sets in the bottom left of Fig. 3 look even more different than square lattices above. However, they are also isometric and actually represent the same hexagonal lattice, because every black point has exactly 6 nearest neighbors that form a regular hexagon.

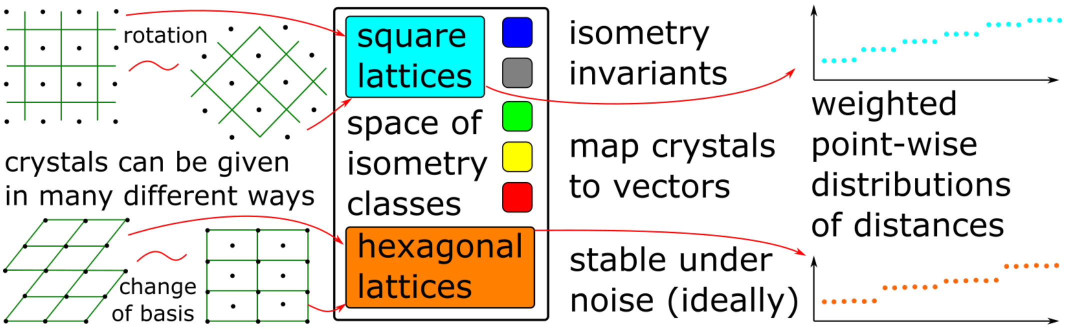
rotation
crystals can be given in many different ways
change of basis
square lattices
space of isometry classes
hexagonal lattices
isometry invariants
map crystals to vectors
stable under noise (ideally)
weighted point-wise distributions of distances

Figure 3 Isometry classes of periodic sets can be distinguished only by isometry invariants.

The key obstacle to compare crystals modulo isometries is the enormous ambiguity or non-uniqueness of a crystal representation illustrated in Fig. 3. A standard Crystallographic Information File (CIF) contains parameters of a unit cell spanned by a linear basis in $\mathbb{R}^3$ and fractional coordinates of atoms from a motif in this basis. If we change a basis as in the bottom left of Fig. 3, the same hexagonal lattice will have a new CIF with a different unit cell possibly containing a different number of points with new fractional coordinates.

Hence cell-dependent descriptors of a crystal can not be justified for comparing crystals modulo isometries. For example, humans should be not be compared or identified by the average color of their clothes, though such colors are easily accessible in photos. Justified comparisons should use only *invariant* features, e.g. biometric data of a human. Any machine learning algorithm can be confused until this representation problem is properly resolved.

The data representation challenge is stated below as the problem to classify periodic sets modulo isometries (or rigid motions) in $\mathbb{R}^n$, see details in Definition 17 in Appendix A.

▶ **Problem 2** (algorithmic classification of periodic sets modulo isometry)**.** Find a function I on periodic crystals or sets in $\mathbb{R}^n$ satisfying the following four conditions:

(2a) *invariance* : I is preserved under any isometry: if S, Q are isometric, then $I(S) = I(Q)$;

(2b) *completeness* : if the invariants coincide: $I(S) = I(Q)$, then S and Q are isometric;

(2c) *continuity* : $I(S)$ continuously changes under any small perturbation of points in S;

(2d) *computability* : $I(S)$ is computable in a polynomial time in the size of a motif of S. ◼

Condition (2b) means that a complete invariant is sufficient to unambiguously identify a periodic crystal in the same way as a DNA code identifies a human. Unfortunately, many claimed 'fingerprints of materials' distinguish only (usually about 90%) crystals in certain datasets. There was no invariant that satisfies all conditions of Problem 2, see section 3.

Though a minimal (by volume) unit cell isn't invariant under a change of basis, the volume is invariant. The second set in Fig. 4 is a slight perturbation of the square lattice, but has a rectangular minimal unit cell, not a square. Hence the volume of a minimal cell is unstable under atomic vibrations. Condition (2c) is needed to continuously quantifying a similarity between crystals. Algorithmic condition (2d) is added to guarantee fast applications in crystal discovery.

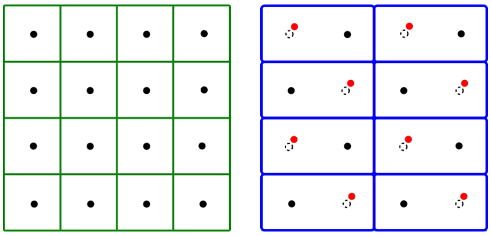

Figure 4 Left: the square lattice. **Right**: a perturbed periodic set has a rectangular minimal cell, so the cell volume is unstable.

Here are the algorithmic contributions.

- Theorem 6 gives a simple algorithmic criterion to check if given crystals are *homometric* (have identical diffraction patterns), which was possible to verify only experimentally.
- For any finite or periodic set $S \subset \mathbb{R}^n$, Definition 8 introduces a weighted point-wise distribution (WPD) of distances, whose isometry invariance is proved in Theorem 10.
- The stability of new invariants under atomic perturbations is justified in Theorems 12, 14.
- Theorem 15 proves completeness of new invariants for finite sets in a general position.

Section 7 discusses how the new invariants helped better compare crystals reported in our colleagues' Nature paper [19] and offers further open problems for algorithmic communities.

3 Related past work on comparisons of point sets and crystals

This section discusses the closest work for finite and periodic sets. The excellent book [11] reviews the wider area of distance geometry. The full distribution of all pairwise Euclidean distances $|ab|$ between points a, b in a finite set $S \subset \mathbb{R}^m$ is a well-known isometry invariant. This invariant is almost complete [2], however Fig. 5 shows non-isometric 4-point sets that are not distinguishable by 6 pairwise distances, i.e. a 4-point set can not be uniquely reconstructed modulo an isometry of $\mathbb{R}^2$ from the distances $\{\sqrt{2}, \sqrt{2}, 2, \sqrt{10}, \sqrt{10}, 4\}$. This example can be extended to any number of points, see Fig. 7. Our methods are similar to the work [10] for finite point clouds.

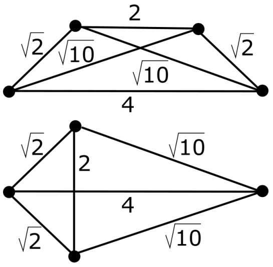

Figure 5 These non-isometric sets have the same pairwise distances.

For periodic sets such as crystals usually given as a CIF file with a unit cell and a motif, it is inevitable to start from a unit cell. However, if output descriptors still depend on a unit cell [7], they are non-invariants modulo isometries. Though the average color can sometimes distinguish all people in a meeting, non-invariants aren't reliable for identifying humans.

Crystallographers have tried to resolve the ambiguity of a unit cell by choosing a canonical unit cell. The best example of such a unique cell is Niggli's reduced cell [6, section 9.3], so Niggli's reduction should be the first step. In 1980 Niggli's cell was shown to be unstable in the sense that a reduced cell of a perturbed lattice can have a basis that substantially differs from that of a non-perturbed lattice [1], see more details of this instability in [13].

More than 40 years since the 1980 paper [1], the main author has highlighted the difficulty of comparing periodic sets at his webpage http://ronininstitute.org/research-scholars/larry-andrews as follows: "...find a measure of the difference between pairs of lattices. Surprisingly, this is not a mathematical problem with a well-defined solution". Our recent work [13] has resolved this problem (only for lattices so far) by introducing two distance functions that satisfy the metric axioms so that the distance between any isometric lattices is exactly 0.

Though there is still no justified distance that satisfies metric axioms for any periodic crystals, the COMPACK algorithm [4] in the Mercury software is widely used for a pairwise comparison of crystals as follows. Within given tolerances (20° for angles and 20% for distances), up to a given number (15 by default) of molecules from two crystals are matched by a rigid motion that minimizes the Root Mean Square deviation of n matched atoms $\text{RMS} = \sqrt{\frac{1}{n}\sum\limits_{i=1}^{n}|p_i - q_i|^2}$. Table 1 shows how this RMS depends on the maximum number of attempted molecules to match. A final number of matched molecules is unpredictable.

matched molecules	5 of 5	8 of 10	10 of 15	11 of 20	16 of 25	18 of 30	21 of 35
RMS in Angstroms	0.603	0.681	0.812	0.825	0.99	1.027	1.079

Table 1 The Root Mean Square (RMS) deviation between the experimental T2-δ crystal and its closest similated version with ID 14 from the dataset reported in the 2017 Nature paper [19]. The irregular dependence of RMS on a number of matched molecules makes this comparison unreliable.

The newer COMPSTRU algorithm [5] similarly to COMPACK predicts a similarity between a reference crystal S and other available crystals whose unit cell parameters are close to those of S. The default thresholds are 5° for angles and 0.5Å for distances (1Å $= 10^{10}$m). The COMPSTRU comparison is restricted to crystals that have the same space-group type.

Crystals are compared by diffraction patterns up to a cut off radius [14], which introduces an extra parameter without resolving the underlying instability under perturbations.

4 A fast algorithm to detect sets with identical diffraction patterns

This section discusses homometric crystals that were especially hard to distinguish, because they have identical diffraction patterns depending only on the *difference* set defined below.

▶ **Definition 3** (difference multi-set $\mathrm{Dif}(S)$, distance multi-set $\mathrm{Dist}(S)$)**.** Let $S \subset \mathbb{R}^n$ be a finite or a periodic set. The *difference* multi-set is $\mathrm{Dif}(S) = \{\vec{a} - \vec{b}$ for all points $a, b \in S\}$. The *distance* multi-set is $\mathrm{Dist}(S) = \{|\vec{a} - \vec{b}|$ for all points $a, b \in S\}$. ■

If a set of points $S \subset \mathbb{R}^n$ is finite, then so is the difference set $\mathrm{Dif}(S)$, hence the vector differences $\vec{a} - \vec{b}$ can be counted with multiplicities. For any periodic set S, any vector difference or a distance will be repeated infinitely many times due to periodicity, hence all values in $\mathrm{Dif}(S)$ and $\mathrm{Dist}(S)$ have the same infinite (countable) multiplicity.

▶ **Example 4** (Patterson's homometric 1D periodic sets)**.** Patterson [15, p. 197, Fig. 2] has suggested the 1D periodic sets $S = \{0, 1, 3, 4\} + 8\mathbb{Z}$ and $Q = \{0, 3, 4, 5\} + 8\mathbb{Z}$, see Fig. 6 and Fig. 7 for circular versions. Theorem 10 will justify that the sets S, Q are non-isometric.

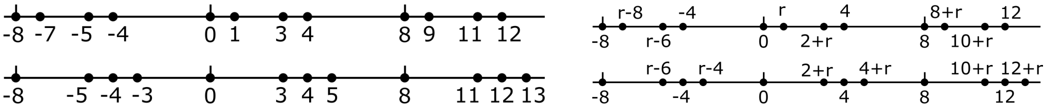

Figure 6 Non-isometric homometric sets, see Definition 5. **Top**: $S(r) = \{0, r, r+2, 4\} + 8\mathbb{Z}$. **Bottom**: $Q(r) = \{0, r+2, 4, r+4\} + 8\mathbb{Z}$, $0 < r \leq 1$ is a parameter. Set $r = 1$ for S, Q on the left.

The vector differences of the 4-point motives of S, Q in the left hand side pictures of Fig. 6 differ: $\begin{array}{c|cccc} S & 0 & 1 & 3 & 4 \\ \hline 0 & 0 & -1 & -3 & -4 \\ 1 & 1 & 0 & -2 & -3 \\ 3 & 3 & 2 & 0 & -1 \\ 4 & 4 & 1 & 3 & 0 \end{array}$ and $\begin{array}{c|cccc} Q & 0 & 3 & 4 & 5 \\ \hline 0 & 0 & -3 & -4 & -5 \\ 3 & 3 & 0 & -1 & -2 \\ 4 & 4 & 1 & 0 & -1 \\ 5 & 5 & 2 & 1 & 0 \end{array}$, but they coincide modulo the period (with infinite multiplicities): $\mathrm{Dif}(S) \equiv \{0, 1, 2, 3, 4, 5, 6, 7\} \equiv \mathrm{Dif}(Q) \bmod 8$.

The equivalence modulo 8 gives rise to a bijection between all 16 elements of the distance matrices above, hence to a bijection between the sets of vector differences with multiplicities $D(S) \to D(Q)$, e.g. the difference $(8i+1) - (8j+4) = 8(i-j) - 3 \equiv 5 \pmod 8$ in S can be bijectively mapped to $(8i+5) - 8j = 8(i-j) + 5$ in Q. Fig. 7 shows a generic pair from the 1-parameter family of homometric sets $S(r), Q(r)$, where $r = 1$ is for the sets S, Q on the left.

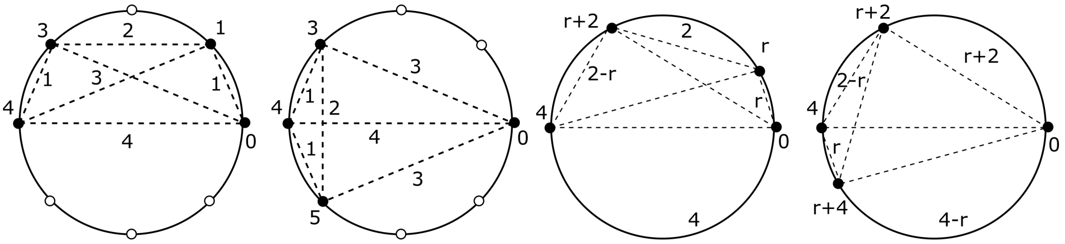

Figure 7 First two: circular versions of the homometric sets S, Q in Fig. 6. Each circle splits into 8 equal arcs. The distances between points (shown outside the disk) are arc lengths (shown inside the disk). **Last two**: homometric sets $S(r) = \{0, r, r+2, 4\} + 8\mathbb{Z}$, $Q(r) = \{0, r+2, 4, r+4\} + 8\mathbb{Z}$, $0 < r < 2$. The distances between points (shown outside the disk) are arc lengths (inside the disk).

The mirror image of $S(r) = \{0, r, r+2, 4\} + 8\mathbb{Z}$ under the reflection $t \mapsto 4 - t$ coincides with $S(2-r) = \{0, 2-r, 4-r, 4\} + 8\mathbb{Z}$, so they are equivalent modulo all isometries including reflections. Similarly, $Q(r)$ and $Q(2-r)$ are isometric by the reflection $t \mapsto -t$. To distinguish all these sets modulo an isometry in section 5, we can assume that $0 < r \leq 1$.

▶ **Definition 5** (homometric sets)**.** Finite or periodic sets $S, Q \subset \mathbb{R}^n$ are called *homometric* if there is a bijection between their multi-sets of vector differences $\mathrm{Dif}(S) \to \mathrm{Dif}(Q)$ from Definition 3. So if $S, Q \subset \mathbb{R}^3$ are crystals, they have identical diffraction patterns. ■

The following result makes the experimental concept of a homometric crystal completely verifiable in an algorithmic way. Theorem 6 and all others are proved in Appendix B.

▶ **Theorem 6** (a fast algorithmic criterion of homometric sets)**.** *Any periodic sets $S, Q \subset \mathbb{R}^n$ are homometric in the sense of Definition 5 if and only if S, Q have a common lattice Λ such that their sets of vector differences are equal modulo this lattice: $\mathrm{Dif}(S) \equiv \mathrm{Dif}(Q) \pmod{\Lambda}$. Given a common unit cell containing m points of periodic set $S, Q \subset \mathbb{R}^n$, there is an algorithm of complexity $O(m^2)$ to determine whether S, Q are homometric.* ■

5 Distributions of distances in finite and periodic sets

This section introduces new isometry invariants of periodic crystals: point-wise distributions of distances in Definition 7 and graphs of their simplified averages (AMDs) in Definition 9.

▶ **Definition 7** (point-wise distribution of distances PDD)**.** Let a periodic set $C = M + \Lambda$ have a motif M of m points $p_1, \dots, p_m$ within a unit cell U of a lattice Λ. For a fixed integer $k \geq 1$, the *point-wise distribution of distances* is the $m \times k$ matrix $\mathrm{PDD}(C; k)$, whose i-th row corresponds to the point p_i, $i = 1, \dots, m$. The i-th row consists of the ordered distances $d_{i1} \leq \dots \leq d_{ik}$ measured from p_i to its first k nearest neighbors within C. ■

The sets S, Q in Fig. 6 have these point-wise distribution of distances (PDD) for $k = 3$.

S	1st distance	2nd distance	3rd	Q	1st distance	2nd distance	3rd
$p_1 = 0$	$\lvert 0-1 \rvert = 1$	$\lvert 0-3 \rvert = 3$	4	$p_1 = 0$	$\lvert 0-3 \rvert = 3$	$\lvert 0-(-3) \rvert = 3$	4
$p_2 = 1$	$\lvert 1-0 \rvert = 1$	$\lvert 1-3 \rvert = 2$	3	$p_2 = 3$	$\lvert 3-4 \rvert = 1$	$\lvert 3-5 \rvert = 2$	3
$p_3 = 3$	$\lvert 3-4 \rvert = 1$	$\lvert 3-1 \rvert = 2$	3	$p_3 = 4$	$\lvert 4-3 \rvert = 1$	$\lvert 4-5 \rvert = 1$	4
$p_4 = 4$	$\lvert 4-3 \rvert = 1$	$\lvert 4-1 \rvert = 3$	4	$p_4 = 5$	$\lvert 5-4 \rvert = 1$	$\lvert 5-3 \rvert = 2$	3

The rows of $\mathrm{PDD}(C; k)$ correspond to an arbitrary order of given points $p_1, \dots, p_m \in M$. There is a suitable convention to order rows by using columns. The columns of $\mathrm{PDD}(C)$ are naturally ordered by increasing distances to neighbors. Then the rows (hence, the points

$p_1, \dots, p_m$) can *lexicographically* ordered as follows. A row $(d_{i1}, \dots, d_{ik})$ is smaller than $(d_{j1}, \dots, d_{jk})$ if the first (possibly none) distances coincide: $d_{i1} = d_{j1}, \dots, d_{il} = d_{jl}$ for some $l \in \{1, \dots, k-1\}$ and the next distances satisfy $d_{i,l+1} < d_{j,l+1}$. In this lexicographic order, the sets S, Q in Fig. 6 have $\mathrm{PDD}(S;3) = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \\ 1 & 3 & 4 \\ 1 & 3 & 4 \end{pmatrix}$ and $\mathrm{PDD}(Q;3) = \begin{pmatrix} 1 & 1 & 4 \\ 1 & 2 & 3 \\ 1 & 2 & 3 \\ 3 & 3 & 4 \end{pmatrix}$.

Notice that $\mathrm{PDD}(S;3)$ contains two pairs of identical rows, because $S = \{0,1,3,4\} + 8\mathbb{Z}$ is symmetric with respect to the reflection $t \mapsto 4 - t \pmod 8$. The similar reflection $t \mapsto -t \pmod 8$ explains two identical rows in $\mathrm{PDD}(Q;3)$ of $Q = \{0,3,4,5\} + 8\mathbb{Z}$. Below we discuss other cases when PDD can contain many identical rows and how to reduce them.

The point-wise distribution of distances (PDD) in Definition 7 depends on a number m of points in a given unit cell U of a crystal C. If we make one of its edges twice longer, the resulting non-primitive unit cell will contain $2m$ points and PDD will be twice larger. However, a translated copy of any point $p_i \in U$ will have exactly the same ordered distances to its neighbors as p_i due to periodicity. After doubling U as above, every row appears twice in PDD. These repetitions can be taken into account by using the weights of rows below.

▶ **Definition 8** (weighted point-wise distribution WPD). The *weight* of a row in $\mathrm{PDD}(S;k)$ of a periodic set $C = M + \Lambda$ with m points in a motif M is defined as the number of times that the row appears in $\mathrm{PDD}(S;k)$ divided by m so that all weights sum up to 1. The *weighted point-wise distribution* $\mathrm{WPD}(S;k)$ is obtained from $\mathrm{PDD}(S;k)$ by keeping only one of identical rows and putting the weight of this row into the extra $(k+1)$-st column. ◼

The rows of $\mathrm{WPD}(C;k)$ are lexicographically ordered as in $\mathrm{PDD}(C;k)$. Then S, Q in Fig. 6 have $\mathrm{WPD}(S;3) = \left(\begin{array}{ccc|c} 1 & 2 & 3 & 1/2 \\ 1 & 3 & 4 & 1/2 \end{array}\right)$ and $\mathrm{WPD}(Q;3) = \left(\begin{array}{ccc|c} 1 & 1 & 4 & 1/4 \\ 1 & 2 & 3 & 1/2 \\ 3 & 3 & 4 & 1/4 \end{array}\right)$.

Any isometric crystals have the same lattice, the same Niggli's reduced cell, the same number m of points in a motif, hence the same number of rows in PDD and WPD. However, this isometry invariant (the number of points within a primitive cell) is unstable under perturbations by the following reasons. If we slightly perturb one point within a motif, hence all its periodic copies of this point in a crystal, the new perturbed crystal might have a different number of rows in WPD. The weights of rows will allow us to continuously quantify such a small perturbation using a suitable distance between distributions. Another stable way to remove the dependence on a cell is to average distances over all points in a motif.

▶ **Definition 9** (average minimum distance AMD_k). For $k \geq 1$, the *average minimum distance* of a periodic set $S = M + \Lambda$ with m points $p_1, \dots, p_m$ in a unit cell U of a lattice Λ is $\mathrm{AMD}_k(S) = \frac{1}{m}\sum\limits_{i=1}^{m} \mathrm{PDD}_{ik}(S;k)$, the average of the last k-th column in the matrix $\mathrm{PDD}(S;k)$. Alternatively, if $\mathrm{WPD}(S;k)$ has l rows with weights $w_1, \dots, w_l$ such that $\sum\limits_{i=1}^{l} w_i = 1$, then $\mathrm{AMD}_k(S) = \sum\limits_{i=1}^{l} w_i \mathrm{WPD}_{ik}$ is the weighted average of the k-th column in $\mathrm{WPD}(S;k)$. ◼

Any lattice $L \subset \mathbb{R}^n$ has a unit cell with only one point in a motif. After a suitable translation, this point can be assumed to be at the origin $0 \in \mathbb{R}^n$. Then $\mathrm{AMD}_k(L)$ is the k-th minimum distance from 0 to another point of L, see Fig. 11 in appendix A.

6 Invariance of distance distributions and their stability under noise

This section proves the main results justifying that WPDs and AMDs from section 5 are stable isometry invariants (Theorems 10, 12, 14), complete for generic sets (Theorem 15).

Since the periodic sets $S(r) = \{0, r, r+2, 4\} + 8\mathbb{Z}$ and $S(2-r)$ are isometric by the reflection $t \mapsto 4-t$, their WPDs are identical. The similar conclusion holds for $Q(r) = \{0, r+2, 4, r+4\} + 8\mathbb{Z}$ and $Q(2-r)$ isometric by the reflection $t \mapsto -t$. Theorem 10 allows us to justify that all $S(r)$ and $Q(r)$ are not isometric to each other for any $0 < r \leq 1$.

▶ **Theorem 10** (isometry invariance of WPDs and AMDs)**.** For any finite or periodic set $S \subset \mathbb{R}^n$, the weighted point-wise distribution $\mathrm{WPD}(S;k)$ and average minimum distance $\mathrm{AMD}_k(S)$ are isometry invariants for any number $k \geq 1$ of neighbors. ■

Theorem 10 and all others are proved in section 9. Now the power of WPDs is illustrated by classifying the homometric sets that are impossible to distinguish by diffraction patterns. Table 2 in appendix A shows the detailed computations of $\mathrm{PDD}(S(r);5)$ and $\mathrm{PDD}(Q(r);5)$ for the homometric sets $S(r), Q(r)$ in Fig. 7, where the rows are ordered by the given points. The lexicographic re-ordering of the rows of PDDs gives the weighted point-wise distributions:

$$\mathrm{WPD}(S(r);5) = \left(\begin{array}{ccccc|c} r & 2 & 4-r & 4+r & 6 & 1/4 \\ r & 2+r & 4 & 4 & 6-r & 1/4 \\ 2-r & 2 & 2+r & 6-r & 6 & 1/4 \\ 2-r & 4-r & 4 & 4 & 4+r & 1/4 \end{array}\right),$$

$$\mathrm{WPD}(Q(r);5) = \left(\begin{array}{ccccc|c} r & 2-r & 4 & 4 & 6-r & 1/4 \\ r & 2 & 4-r & 4+r & 6 & 1/4 \\ 2-r & 2 & 2+r & 6-r & 6 & 1/4 \\ 2+r & 4-r & 4 & 4 & 4+r & 1/4 \end{array}\right).$$

Already the first columns of WPDs for $k = 1$ distinguish all $S(r), Q(r)$ for any $0 < r \leq 1$. $\mathrm{AMD}_k(S(r))$ are independent of r, hence don't distinguish these periodic sets modulo isometries. However, for $0 < r \leq 1$, the first minimum distance between any points in $S(r)$ equals r and implies that $S(r)$ are not isometric to each other for different parameters r.

Since all atoms thermodynamically vibrate above the absolute zero temperature, the bottleneck distance between crystals in Definition 11 naturally quantifies crystal similarities.

▶ **Definition 11** (bottleneck distance BND)**.** For a fixed bijection $g : S \to Q$ between finite or periodic sets $S, Q \subset \mathbb{R}^n$, the *maximum deviation* of points is the supremum $\sup\limits_{a \in S} |a - g(a)|$ of Euclidean distances over all points in S. The *bottleneck distance* is $\mathrm{BND}(S,Q) = \inf\limits_{g:S\to Q} \sup\limits_{a\in S} |a - g(a)|$ is the infimum of maximum deviations over all bijections $g : S \to Q$. ■

The key obstacle in applying the bottleneck distance to real crystals is the required optimization over bijections between infinite sets. Any reduction to a finite set is hard to justify because of the inevitable instability of a unit cell under perturbations [1].

▶ **Theorem 12** (stability of $\mathrm{AMD}_k(S)$ under perturbations of S)**.** For any number $k \geq 1$ of neighbors, any finite or periodic sets S, Q satisfy $|\mathrm{AMD}_k(S) - \mathrm{AMD}_k(Q)| \leq 2\mathrm{BND}(S,Q)$. ■

Weighted point-wide distributions are matrices that can have different sizes, hence are harder to compare than AMD vectors of a fixed length k. Definition 13 [20] introduces a suitable distance measure between such weighted point-wise distributions of different sizes.

▶ **Definition 13** (earth mover's distance EMD)**.** Fix two finite or periodic sets $S, Q \subset \mathbb{R}^n$ and a number k of nearest neighbors for each point. Let the weighted point-wise distribution $\mathrm{WPD}(S;k)$ consist of m_S rows $R_i(S) \in \mathbb{R}^k$. Each row $R_i(S)$ has a weight $w_i(S)$, $i = 1,\dots,m(S)$ so that $\sum\limits_{i=1}^{m(S)} w_i(S) = 1$. Using the similar notations for the set Q, we quantify by a parameter $0 \le f_{ij} \le 1$ a move from each row $R_i(S) \in \mathbb{R}^k$ to another row $R_j(Q) \in \mathbb{R}^k$ of a weight $w_j(Q)$, where $j = 1,\dots,m(Q)$. The distance between such rows (vectors in $\mathbb{R}^k$) is Euclidean. The *earth mover's distance* is defined as the minimum value of the cost flow $\mathrm{EMD}(S,Q) = \sum\limits_{i=1}^{m(S)} \sum\limits_{j=1}^{m(Q)} f_{ij}|R_i(S) - R_j(Q)|$ over all $0 \le f_{ij} \le 1$ subject to the constraints $\sum\limits_{j=1}^{m(Q)} f_{ij} = w_i(S)$ for $i = 1,\dots,m(S)$, and $\sum\limits_{i=1}^{m(S)} f_{ij} = w_j(Q)$ for $j = 1,\dots,m(Q)$. ■

The first condition $\sum\limits_{j=1}^{m(Q)} f_{ij} = w_i(S)$ means that the full weight $w_i(S)$ of the row $R_i(S)$ 'flows' into the rows $R_j(Q)$, each via a 'flow' f_{ij}, $j = 1,\dots,m(Q)$. Similarly, the second condition $\sum\limits_{i=1}^{m(S)} f_{ij} = w_j(Q)$ means that all 'flows' f_{ij} from rows $R_i(S)$ for $i = 1,\dots,m(S)$ 'flow' into the row $R_j(Q)$ and sum up to the full weight $w_j(Q)$.

The earth mover's distance (EMD) has more than one advantage over the bottleneck distance (BND) for periodic sets. First, the EMD uses the isometry invariant WPD, whose stability in the EMD is proved in Theorem 14. The BND between infinite sets can be computed only on finite subsets, e.g. on points in an extended cell, which is unstable [1].

Second, even for finite subsets, the fastest algorithm by Kerber et al. [8, Theorem 3.1] computes the BND (for 2D set of m points) in time $O(m^{1.5}\log m)$. The EMD can be approximated [17] in a time linear in the size of any k-dimensional distributions.

▶ **Theorem 14** (stability of weighted point-wise distribution WPD)**.** For any number $k \ge 1$ of neighbors, any finite or periodic sets S, Q satisfy $\mathrm{EMD}(S,Q) \le 2\sqrt{k}\,\mathrm{BND}(S,Q)$. So any small perturbation of atomic positions in the bottleneck distance (BND) leads to a small change of the weighted point-wise distribution in the earth mover's distance (EMD). ■

After satisfying the invariance and stability conditions in Problem 2, Theorem 15 proves a generic completeness of the weighted point-wise distributions (WPDs) for finite sets.

▶ **Theorem 15** (unique reconstruction of a finite set from WPD)**.** Let a finite set $S \subset \mathbb{R}^n$ have m points such that all pairwise distances between points of S are distinct. Then S can be uniquely reconstructed modulo an isometry of $\mathbb{R}^n$ from the distribution $\mathrm{WPD}(S; m-1)$. ■

We conjecture that $\mathrm{WPD}(S;k)$ are complete isometry invariants for sufficiently large k depending on a complexity of $S \subset \mathbb{R}^n$. If S is a finite set of m points, then $k = m-1$ should be enough. For example, the 4-point sets $A, B \subset \mathbb{R}^2$ in Fig. 5 have $\mathrm{WPD}(A;3) = \left(\begin{array}{ccc|c} \sqrt{2} & 2 & \sqrt{10} & 1/2 \\ \sqrt{2} & \sqrt{10} & 4 & 1/2 \end{array}\right)$ and $\mathrm{WPD}(B;3) = \left(\begin{array}{ccc|c} \sqrt{2} & \sqrt{2} & 4 & 1/4 \\ \sqrt{2} & 2 & \sqrt{10} & 1/2 \\ \sqrt{10} & \sqrt{10} & 4 & 1/4 \end{array}\right)$, which distinguish $A, B \subset \mathbb{R}^2$ modulo isometries. Actually, $k = 1$ is already enough in this case.

If $m = 1$, then any set $S \subset \mathbb{R}^n$ is a lattice and k needs to be at least $n(n+1)$. For example, any lattice in $\mathbb{R}^2$ can be reconstructed from the distribution of 6 minimum distances from (say, the origin) to 3 pairs of 6 neighbors symmetric with respect to the origin.

7 Computations, applications to crystal comparisons and a discussion

Theorem 16 covers final computability condition in Isometry Classification Problem 2.

▶ **Theorem 16** (algorithm for weighted point-wise distributions WPDs and average minimum distances AMDs)**.** Let a periodic crystal $S \subset \mathbb{R}^n$ have m points in a unit cell whose extension by a factor μ covers all k of neighbors of the given points. Then the matrix $\mathrm{WPD}(S;k)$ and all $\mathrm{AMD}_i(S)$ for $i = 1,\dots,k$ can be computed in time $O(m(n\mu^n + k)\log(\mu^n m))$. ■

Though we have no exact value of the factor μ, our experiments show that $\mu = O(n)$. So in the practical case of $m = 3$ the time is near linear in the number m of points in a motif.

Now we discuss how the new invariants help to accelerate the Crystal Structure Prediction (CSP) using the real crystals and simulated data reported by our colleagues [19].

The only input for computing the new isometry invariants is a crystal itself (a unit cell with a finite set of points) without any extra parameters. The number of neighbors k is independent of a crystal and reflects our desire to extra more distance information. For example, vectors of 1000 AMDs will better differentiate crystals than vectors of 100 AMDs.

The Nature paper [19] has reported 4 experimental crystals T2-α, T2-β, T2-γ, T2-δ (one more T2-ε was synthesized after the publication), see Fig. 12 in Appendix A. The synthesis in a lab started only after an energy landscape in Fig. 1 of 5679 simulated crystals produced by 12-week simulations on a supercomputer hinted at potential stable crystals in downward spikes (imaginable deep minima). To validate this approach, the synthesized crystals should be matched with closest crystals from the simulated dataset of 5679. If there was no close match, the expensive simulations missed a real crystal, which is always possible, because the continuous space of all potential crystals in Fig. 2 is randomly and discretely sampled.

Until now the density was practically used as a stable isometry invariant of crystals. The density values on the horizontal axis in Fig. 1 can separate nano-porous organic crystals, while inorganic crystals are much denser and can not be well-separated by densities.

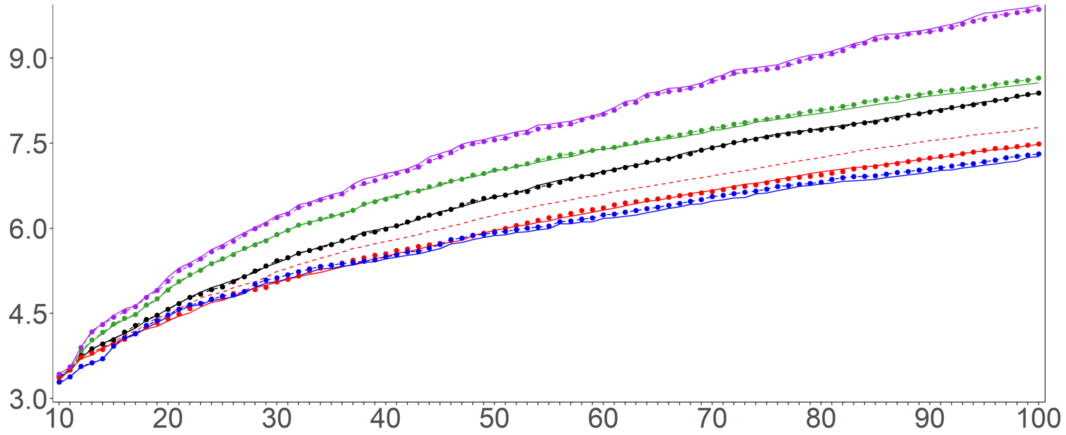

Figure 8 The average minimum distances AMD_k in Angstroms with $k = 10,\dots,100$. Solid curves are for experimental crystals. Dashed curves are for the past simulated matches reported in [19]. Dotted curves are for new matches found by smallest Euclidean distances between AMD vectors with $k = 100$. From top to bottom: purple T2-γ, green T2-α, black T2-β, red T2-δ (new match 15 is much closer by AMDs to the experimental crystal than the old match 14), blue T2-ε.

Using only the density Δ of an experimental crystal, chemists look for a corresponding simulated crystal in a vertical strip of the energy landscape in Fig. 1 over a small interval around Δ to allow for errors. From this strip one takes the crystal with the lowest energy as the best guess, which depends on the width of the strip. A final match is confirmed by the RMS deviation between finite portions of crystals, which is also uncertain, see Table 1.

For the experimental crystal T2-δ, the past method above found crystal 14 in the simulated dataset of 5679. However, another crystal 15 has a much closer AMD curve to the experimental one in Fig. 8. Though both crystals have almost identical energy and density in Fig. 1, they are now separated by their AMD curves (red: dotted vs dashed) in Fig. 8.

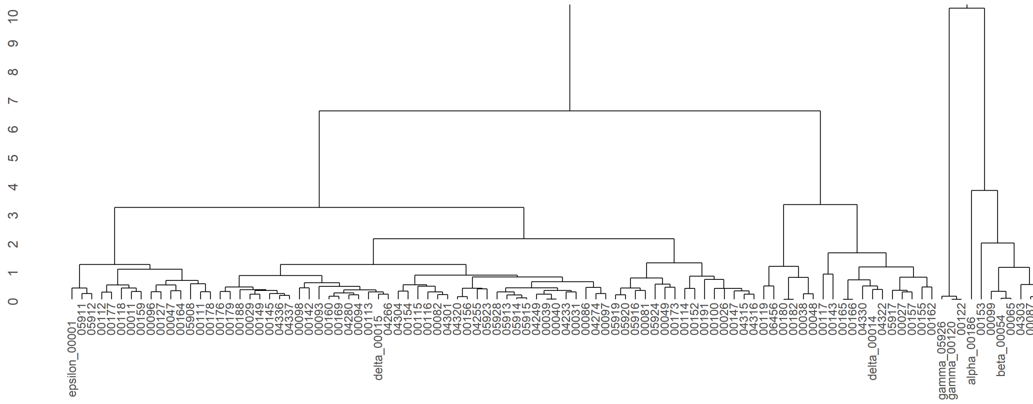

Figure 9 Complete-linkage clustering of 100 crystals with lowest energies [19] by AMD100 curves.

Fig. 9 shows that that most stable 100 simulated crystals split into two clusters, which merge at a high threshold of about 17Å. The 1st large cluster on the left contains two simulated matches of the experimental crystals T2-ε and T2-δ, though the past match 14 is in another subcluster than the new match 15. The 2nd small cluster has all matches of T2-α, T2-β, T2-γ. The thresholds between the largest subclusters of 5 crystals are about 3Å.

Fig. 8 and 9 show that the new invariants from Definitions 8 and 9 continuously quantify similarities between periodic crystals, which was impossible by past non-invariant descriptors or unstable discrete invariants such as symmetry groups, or using the single-value density.

We have resolved the following algorithmic challenges in the geometry for periodic sets.

- The algorithmic criterion in Theorem 6 detects crystals with identical diffraction patterns.
- Theorems 10, 12, 14 have proved that the weighted point-wise distributions $\mathrm{WPD}(S;k)$ and average minimum distances $\mathrm{AMD}_k(S)$ are stable isometry invariants of a finite or a periodic set $S\subset\mathbb{R}^k$ and are computable fast enough by algorithmic Theorem 16.
- Completeness Theorem 15 proves that any set $S\subset\mathbb{R}^n$ of m points with distinct pairwise distances can be uniquely reconstructed from its $\mathrm{WPD}(S;m-1)$ modulo isometries of $\mathbb{R}^n$.

The paper has opened the new horizons for the algorithmic community by stating the important problems from crystallography in the language of computational geometry.

- Prove completeness of weighted point-wise distributions $\mathrm{WPD}(S;k)$ for periodic sets S.
- Given a number k of neighbors, find an extension factor μ of any initial cell in Theorem 16 to guarantee that the extended reduced cell covers k nearest neighbors of all motif points.

The C++ code for the new invariants can be made available in September 2020. We thank all reviewers in advance for their valuable time and helpful suggestions.

References

1 LC Andrews, HJ Bernstein, and GA Pelletier. A perturbation stable cell comparison technique. *Acta Crystallographica A*, 36(2):248–252, 1980.

2 M. Boutin and G. Kemper. On reconstructing n-point configurations from the distribution of distances or areas. *Advances in Applied Mathematics*, 32(4):709–735, 2004.

3 Russell A. Brown. Building a balanced k-d tree in $o(kn \log n)$ time. *J. Computer Graphics Techniques)*, 4(1):50–68, 2015.

4 J. Chisholm and S. Motherwell. Compack: a program for identifying crystal structure similarity using distances. *J. Applied Crystallography*, 38(1):228–231, 2005.

5 G Flor, D. Orobengoa, E. Tasci, J. Perez-Mato, and M. Aroyo. Comparison of structures applying the tools available at the bilbao crystallographic server. *J. Applied Crystallography*, 49(2):653–664, 2016.

6 Theo Hahn, Uri Shmueli, and JC Wilson Arthur. *International tables for crystallography*, volume 1. 1983.

7 Lauri Himanen, Marc OJ Jäger, Eiaki V Morooka, Filippo Federici Canova, Yashasvi S Ranawat, David Z Gao, Patrick Rinke, and Adam S Foster. Dscribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247:106949, 2020.

8 M Kerber, D Morozov, and A Nigmetov. Geometry helps to compare persistence diagrams. *J Experimental Algorithmics*, 22:1–20, 2017.

9 A. Kitaigorodsky. *Molecular crystals and molecules*. Elsevier, 2012 (reprint of 1973).

10 Rongjie Lai and Hongkai Zhao. Multi-scale non-rigid point cloud registration using robust sliced-wasserstein distance via laplace-beltrami eigenmap. *arXiv:1406.3758*, 2014.

11 Leo Liberti and Carlile Lavor. *Euclidean distance geometry: an introduction*. Springer, 2017.

12 Sherry L Morissette, Stephen Soukasene, Douglas Levinson, Michael J Cima, and Örn Almarsson. Elucidation of crystal form diversity of the hiv protease inhibitor ritonavir by high-throughput crystallization. *Proceedings of the National Academy of Sciences*, 100(5):2180–2184, 2003.

13 Marco Mosca and Vitaliy Kurlin. Voronoi-based similarity distances between arbitrary crystal lattices. *Crystal Research and Technology*, 55(5):1900197, 2020.

14 Anton O Oliynyk, Erin Antono, Taylor D Sparks, Leila Ghadbeigi, Michael W Gaultois, Bryce Meredig, and Arthur Mar. High-throughput machine-learning-driven synthesis of full-heusler compounds. *Chemistry of Materials*, 28(20):7324–7331, 2016.

15 A Patterson. Ambiguities in the x-ray analysis of crystal structures. *Physical Review*, 65:195, 1944.

16 AL Patterson. Homometric structures. *Nature*, 143:939–940, 1939.

17 O. Pele and M. Werman. A linear time histogram metric for improved sift matching. In *European Conference on Computer Vision*, pages 495–508, 2008.

18 Sarah L Price. Is zeroth order crystal structure prediction (csp_0) coming to maturity? what should we aim for in an ideal crystal structure prediction code? *Faraday Discussions*, 211:9–30, 2018.

19 Angeles Pulido, Linjiang Chen, Tomasz Kaczorowski, Daniel Holden, M Little, Samantha Chong, Benjamin Slater, D McMahon, Baltasar Bonillo, C Stackhouse, A Stephenson, C Kane, R Clowes, Tom Hasell, Andrew Cooper, and Graeme Day. Functional materials discovery using energy–structure–function maps. *Nature*, 543:657–664, 2017.

20 Y. Rubner, C. Tomasi, and L. Guibas. The earth mover's distance as a metric for image retrieval. *Intern. Journal of Computer Vision*, 40(2):99–121, 2000.

8 Appendix A: background on isometries and isometry invariants

This appendix starts from reminding key concepts of isometries in $\mathbb{R}^n$ and then gives the proofs of all results from sections 4 and 6. The strongest possible equivalence on rigid materials is defined by isometries (or rigid motions) that preserve interpoint distances.

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

Any isometry of $\mathbb{R}^n$ can be decomposed into at most $n+1$ reflections over hyperspaces, hence is bijective and can be inverted. A composition of isometries is also an isometry, hence a well-operation in the group $\text{Iso}(\mathbb{R}^n)$ of all isometries in $\mathbb{R}^n$. Rigid motions are orientation-preserving isometries and form the smaller subgroup $\text{Iso}^+(\mathbb{R}^n) \subset \text{Iso}(\mathbb{R}^n)$. Materials are compared modulo rigid motions or the isometries from $\text{Iso}(\mathbb{R}^n)$, because mirror images of materials can have different properties. Examples in $\mathbb{R}^3$ are translations by vectors and rotations around straight lines.

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

All orthogonal matrices A with $\det A = 1$ form the special orthogonal group $\text{SO}(\mathbb{R}^n)$, where the operation is the matrix multiplication. The group $\text{SO}(\mathbb{R}^2)$ consists of rotations about the origin in the plane. The group $\text{SO}(\mathbb{R}^3)$ consists of rotations about axes passing through the origin in $\mathbb{R}^3$. In general, $\text{SO}(\mathbb{R}^n)$ consists of all isometries from $\text{Iso}^+(\mathbb{R}^n)$ that preserve the origin. Any objects should be classified by invariants that are independent of a given representation of an object. Many machine learning algorithms struggle when features or descriptors include non-invariants of crystals, e.g. parameters of an ambiguous unit cell or atomic coordinates in an arbitrary linear basis.

▶ **Definition 18** (isometry invariant)**.** An *isometry class* is a set of all materials that are isometric to each other, i.e. any materials S, Q from the same class are related by an isometry $S \to Q$. An *isometry invariant* is a function I that maps all materials from a certain class, e.g. all periodic crystals, to a simpler set (e.g. numbers, matrices) so that $I(S) = I(Q)$ for any isometric materials S, Q. An invariant I is called *complete* if the converse is also true: if $I(S) = I(Q)$, then the materials S, Q are isometric. ■

The original definition of homometric crystals [16] said that they should be non-isometric contrary to the discovered isometry above. Hence Definition 5 does not have this restriction and defines an equivalence relation on sets satisfying the three axioms:

reflexivity: any set S is equivalent to itself;

symmetry: if a set S is equivalent to Q, then Q is equivalent to S;

transitivity: if a set S is equivalent to Q, which is equivalent to T, then S is equivalent to T.

The three axioms above guarantee that all sets can split (or classified) into disjoint equivalence classes (consisting of all sets equivalent to each other) and a classification modulo an equivalence relation makes sense. The even better equivalence relation is the isometry combined with the homometry saying that one set S is equivalent to a set Q if $\text{Dif}(f(S)) = \text{Dif}(Q)$ for a suitable isometry f.

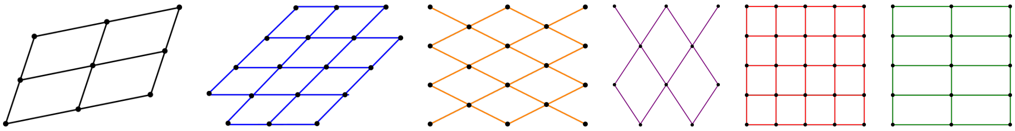

Figure 10 2D lattices whose AMD curves are in Fig. **??** and **??**. **1st**: a generic black lattice with the basis $(1.25, 0.25), (0.25, 0.75)$. **2nd**: the blue hexagonal lattice with the basis $(1, 0), (1/2, \sqrt{3}/2)$. **3rd**: the orange rhombic lattice with the basis $(1, 0.5), (1, -0.5)$. **4th**: the purple rhombic lattice with the basis $(1, 1.5), (1, -1.5)$. **5th**: the red square lattice with the standard basis $(1, 0), (0, 1)$. **6th**: the green rectangular lattice with the basis $(2, 0), (0, 1)$.

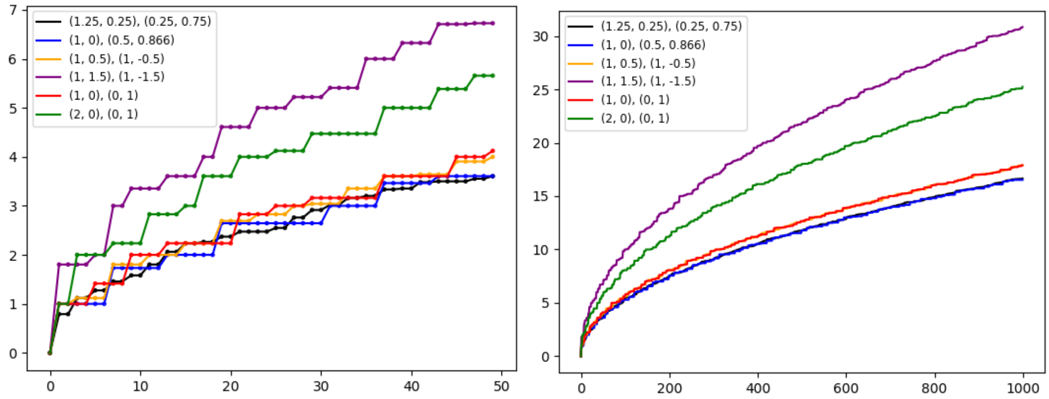

Figure 11 **Left**: AMD$_k$, $k \in [0, 50]$, for the 2D lattices with given bases. **Right**: extended AMD$_k$ up to $k = 1000$. The orange and red graphs are very close as well as the blue and black graphs, see clearer differences for smaller k on the left.

Fig. 11 shows the AMD graphs for six 2D lattices specified by their basis vectors and also shown in Fig. 11 in the same order. Though the orange and red AMD graphs clearly differ Fig. 11 up to $k = 50$, their asymptotic behaviors are very similar for k close to 1000, also for the black and blue graphs, which is interesting to study further.

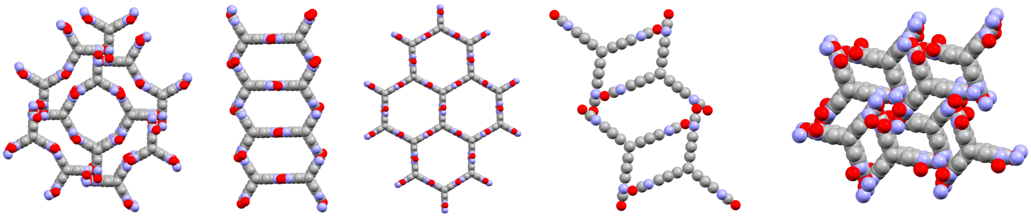

Figure 12 The crystals T2-α, T2-β, T2-γ, T2-δ, T2-ε based on the T2 molecule were synthesized for methane capture following the crystal structure prediction in Fig. 1, see the Nature paper [19].

9 Appendix B: proofs of all theorems from the main paper

Proof of Theorem 6. For the 'only if' part, assume that $\mathrm{Dif}(S) = \mathrm{Dif}(Q)$ as infinite sets. Let $S = M + \Lambda$ be any representation of the crystal S in terms of its arbitrarily unit cell. If

$\mathrm{PDD}(S(r);5)$	1st distance	2nd	3rd	4th	5th
$p_1 = 0$	$\vert 0-r\vert = r$	$2+r$	4	4	$6-r$
$p_2 = r$	$\vert r$	2	$4-r$	$4+r$	6
$p_3 = 2+r$	$2-r$	2	$2+r$	$6-r$	6
$p_4 = 4$	$2-r$	$4-r$	4	4	$4+r$
$\mathrm{AMD}_k(S(r))$	$\mathrm{AMD}_1 = 1$	$\mathrm{AMD}_2 = 2.5$	$\mathrm{AMD}_3 = 3.5$	$\mathrm{AMD}_4 = 4.5$	$\mathrm{AMD}_5 = 5.5$

$\mathrm{PDD}(Q(r);5)$	1st distance	2nd	3rd	4th	5th
$p_1 = 0$	$2+r$	$4-r$	4	4	$r+4$
$p_2 = 2+r$	$2-r$	2	$2+r$	$6-r$	6
$p_3 = 4$	r	$2-r$	4	4	$6+r$
$p_4 = 4+r$	r	2	$4-r$	$4+r$	6
$\mathrm{AMD}_k(Q(r))$	$\mathrm{AMD}_1 = 1+0.5r$	$2.5-0.5r$	$\mathrm{AMD}_3 = 3.5$	$\mathrm{AMD}_4 = 4.5$	$5+0.5r$

Table 2 The point-wise distributions of distances (PDDs) and average minimum distances (AMDs) from Definitions 7, 9 for $S(r) = \{0, r, r+2, 4\} + 8\mathbb{Z}$, $Q(r) = \{0, r+2, 4, r+4\} + 8\mathbb{Z}$, $0 < r \leq 1$.

the motif M consists of m points $p_1, \ldots, p_m$ within the unit cell with a basis $\vec{v}_1, \ldots, \vec{v}_n$, then

$$\mathrm{Dif}(S) = \{p_i - p_j + \sum_{k=1}^{n} \lambda_k \vec{v}_k \ : \ 1 \leq i, j \leq m, \ \lambda_1, \ldots, \lambda_n \in \mathbb{Z}\}.$$

Hence $\mathrm{Dif}(S)$ is also a periodic crystal with the same unit cell. The similar conclusion for $\mathrm{Dif}(Q)$ implies that the crystals $\mathrm{Dif}(Q) = \mathrm{Dif}(S)$, hence S and Q, have the same lattice Λ. Then the difference sets should be equal modulo this lattice: $\mathrm{Dif}(S) \equiv \mathrm{Dif}(Q) \pmod{\Lambda}$.

For the 'if' part, we start from a common lattice Λ of S, Q. Any lattice has a unique Niggli's reduced cell, so we assume that given crystals S, Q has a common unit cell $U$with a basis $\vec{v}_1, \ldots, \vec{v}_n$. The equality $\mathrm{Dif}(S) \equiv \mathrm{Dif}(Q) \pmod{\Lambda}$ means that for any pair of points $q_i, q_j \in Q$, there is a unique pair $p_i, p_j \in S$ and unique coefficients $\lambda_1, \ldots, \lambda_n \in \mathbb{Z}$ such that $q_i - q_j = p_i - p_j + \sum\limits_{k=1}^{n} \lambda_k \vec{v}_k$ and vice versa. We extend this 1-1 correspondence to the infinite set $\mathrm{Dif}(Q)$. For any $q_i - q_j + \sum\limits_{k=1}^{n} \mu_k \vec{v}_k \in \mathrm{Dif}(Q)$, the corresponding difference in $\mathrm{Dif}(S)$ is $p_i - p_j + \sum\limits_{k=1}^{n} (\lambda_k + \mu_k) \vec{v}_k$, which extends the bijection $\mathrm{Dif}(S) \to \mathrm{Dif}(Q)$ to full crystals.

To determine if S, Q are homometric, one can start with a common cell U containing m points of S, Q, which is needed by the above criterion. First compute all $O(m^2)$ pairwise differences (translated to U if necessary) for both S, Q. To check if these vector sets coincide, we could lexicographically order them using coordinates in the basis of the cell U. Then a single pass over $O(m^2)$ vector differences is enough to decide if $\mathrm{Dif}(S) \equiv \mathrm{Dif}(Q) \pmod{\Lambda}$. ◀

Proof of Theorem 10. Any isometry $f : S \to Q$ between sets $S, Q \subset \mathbb{R}^n$ establishes a 1-1 correspondence between points of S and Q. If S, Q are periodic, f bijectively maps a unit cell U of S to a unit cell $U(Q)$ of Q. Hence f allows us to order points $p_1, \ldots, p_m \in U(S)$ according to the order of their images $f(p_1), \ldots f(p_m) \in U(Q)$.

Since the isometry f preserves distances between points, every i-th row of $\mathrm{PDD}(S;k)$, which contains the ordered distances from p_i to its first k nearest neighbors, coincides the

i-th row of $\mathrm{PDD}(Q;k)$, $i=1,\dots,m$. These coincidence of rows gives rise to the equality of the matrices $\mathrm{WPD}(S;k)=\mathrm{WPD}(Q;k)$ of Weighted Point-wise Distributions, which are independent of of point ordering. ◀

Lemma 19 is needed to prove Stability Theorem 12.

▶ **Lemma 19** (perturbed distances). For some $\varepsilon>0$, let $g:S\to Q$ be a bijection between finite or periodic sets such that $|a-g(a)|\leq\varepsilon$ for all $a\in S$. Then, for any $i\geq 1$, let $a_i\in S$ and $b_i\in Q$ be the i-nearest neighbors of points $a\in S$ and $b=g(a)\in Q$, respectively. Then the Euclidean distances from the points a,b to their i-th neighbors a_i,b_i are 2ε-close to each other, i.e. $||a-a_i|-|b-b_i||\leq 2\varepsilon$.

Proof. Shifting the point $g(a)$ back to a, assume that $a=g(a)$ is fixed and all other points change their positions by at most 2ε.

Assume by contradiction that the distance from a to its new i-th neighbor b_i is less than $|a-a_i|-2\varepsilon$. Then all first new i neighbors $b_1,\dots,b_i$ of a within Q belong to the open ball with the center a and the radius $|a-a_i|-2\varepsilon$. Since the bijection g shifted every point $b_1,\dots,b_i$ by at most 2ε, their preimages $g^{-1}(b_1),\dots,g^{-1}(b_i)$ belong to the open ball with the center a and the radius $|a-a_i|$.

Then the i-th neighbor of a within S should be among these i preimages, i.e. the distance from a to its i-th nearest neighbor should be strictly less than the assumed value $|a-a_i|$. A similar contradiction is obtained from the assumption that the distance from a to its new i-th neighbor b_i is more than $|a-a_i|+2\varepsilon$. ◀

Proof of Theorem 12. By Lemma 19 each element of $\mathrm{PDD}(S;k)$ changes by at most 2ε. The average of the k-th column changes by at most 2ε, i.e. $|\mathrm{AMD}_k(S)-\mathrm{AMD}_k(Q)\ \leq 2\varepsilon$. ◀

Lemma 20 is needed to prove Stability Theorem 14.

▶ **Lemma 20** (perturbed rows). For some $\varepsilon>0$, let $g:S\to Q$ be a bijection between finite or periodic sets such that $|a-g(a)|\leq\varepsilon$ for all $a\in S$. Then, for any $k\geq 1$, the bijection g changes the vector $\vec{R}_a(S)=(|a-a_1|,\dots,|a-a_k|)\in\mathbb{R}^k$ of the first k minimum distances from any point $a\in S$ to its k nearest neighbors $a_1,\dots,a_k\in S$ by a Euclidean distance at most $2\varepsilon\sqrt{k}$. So if $b_1,\dots,b_k\in Q$ are the k nearest neighbors of $b=g(a)$ within Q and $\vec{R}_b(S)=(|b-b_1|,\dots,|b-b_k|)\in\mathbb{R}^k$ is the vector of the first k minimum distances from the point $b=g(a)$, then $|\vec{R}_a(S)-\vec{R}_b(Q)|\leq 2\varepsilon\sqrt{k}$.

Proof. By Lemma 19 every coordinate of the row $\vec{R}_a(S)\in\mathbb{R}^k$ changes by at most 2ε. Hence the Euclidean distance from $\vec{R}_a(S)$ to the perturbed row $\vec{R}_b(Q)$ is at most $2\varepsilon\sqrt{k}$. ◀

Proof of Theorem 14. $\mathrm{BND}(S,Q)=\inf\limits_{g:S\to Q}\sup\limits_{a\in S}|a-g(a)|$ by Definition 11 means, for any $\delta>0$, there is a bijection $g:S\to Q$ such that $\sup\limits_{a\in S}|a-g(a)|\leq\mathrm{BND}(S,Q)+\delta$. If the given sets S,Q are finite, one can set $\delta=0$. Indeed, there are only finitely many bijections $S\to Q$, hence the infimum in Definition 11 will be achieved for one of them. If the sets S,Q are periodic, then the chosen bijection g restricts to a bijection between all points in corresponding unit cells of S,Q, so $m(S)=m(Q)$, say both are equal to m.

For any fixed $k \geq 1$, we will design a specific flow from the rows of $\mathrm{WPD}(S;k)$ to the rows of $\mathrm{WPD}(Q;k)$ with f_{ij} satisfying Definition 13. We first start from a 1-1 flow with $f_{ij} = 0$ for $i \neq j$.

If not all rows $R_i(S)$ in $\mathrm{PDD}(S;k)$ are distinct, we make them symbolically distinct so that $\mathrm{WPD}(S;k)$ is obtained from $\mathrm{PDD}(S;k)$ by adding the column of equal weights $\frac{1}{m}$, similarly for $\mathrm{WPD}(Q;k)$. Identifying equal rows later will mean that flows to (or from) equal (symbolically different) rows are combined into a many-to-one (or one-to-many, respectively) flow. Since we have the same number m of rows in both matrices $\mathrm{WPD}(S;k)$ and $\mathrm{WPD}(Q;k)$, we set $f_{ij} = 0$ for $i \neq j$ and $f_{ii} = \frac{1}{m}$, $i = 1, \dots, m$.

Then $\mathrm{EMD}(S,Q) \leq \frac{1}{m}\sum\limits_{i=1}^{m} |\vec{R}_i(S) - \vec{R}_i(Q)|$, because EMD minimizes the cost over all flows in Definition 13. Since each $|\vec{R}_i(S) - \vec{R}_i(Q)| \leq 2\sqrt{k}(\mathrm{BND}(S,Q) + \delta)$ by Lemma 20, we conclude that $\mathrm{EMD}(S,Q) \leq \frac{1}{m}\sum\limits_{i=1}^{m} 2\sqrt{k}(\mathrm{BND}(S,Q)+\delta) = 2\sqrt{k}(\mathrm{BND}(S,Q)+\delta)$. Since the last inequality holds for any small $\delta > 0$, we get the Lipschitz continuity $\mathrm{EMD}(S,Q) \leq 2\sqrt{k}\mathrm{BND}(S,Q)$. ◀

The key idea of the following proof is to convert a weighted point-wise distribution (in a generic case when all pairwise distances are distinct) into a distance matrix on ordered points.

Proof of Theorem 15. Since all pairwise distances between m points of S are distinct, every distance appears in the matrix $\mathrm{WPD}(S;m-1)$ exactly twice, once as the distance from a point p_i to its neighbor p_j, and once more as the distance from p_j to p_i, though these equal entries are not symmetric.

We will convert $\mathrm{WPD}(S;m-1)$ into the distance matrix $D(S)$ as follows. Let $d_1 < d_2 < \cdots < d_{m-1}$ be all strictly increasing distances from a (say) first point p_1 of S to the $m-1$ others.

Each distance d_i from the first row appears exactly once more in another (say, i'-th) row of $\mathrm{WPD}(S;m-1)$. Then d_i is the distance between the points p_1 and $p_{i'}$ numbered as the i'-th row. The map of indices $i \mapsto i'$ is a permutation of $\{2,\dots,m\}$. We set $D_{11} = 0$ and $D_{1,i'} = d_i$ for each $i = 2,\dots,m$. Then we similarly permute indices in the 2nd row of $\mathrm{WPD}(S;m-1)$, starting from the 3rd index due to the symmetry of $D(S)$, and so on.

The full distance matrix $D(S)$ uniquely determines a set with ordered points $S \subset \mathbb{R}^n$ modulo isometries by the classical multi-dimensional scaling [11, Section 8.5.1]. ◀

Proof of Theorem 16. We build a k-d tree [3] on $\mu^n m$ points in the extended unit cell in time $O(n(\mu^n m)\log(\mu^n m))$. Then all k neighbors of m initial points can be found in time for each $O(km\log(\mu^n m))$. After completing the $m \times k$ matrix PDD, we lexicographically sort its m rows. Since each row comparison needs $O(k)$ time, the matrix $\mathrm{WPD}(S;k)$ can be obtained in time $O(km\log m)$. Finally, all $\mathrm{AMD}_i(S)$ are computed as column averages in time $O(km)$. The total time is $O(m(n\mu^n + k)\log(\mu^n m))$. ◀

To guarantee that all k neighbors are correctly found, we incrementally increase μ and check if a new layer of cells leads to any updates in the current $m \times k$ matrix containing distances from m initial points to their k nearest neighbors. If no updates happen for a point p, all k minimum distances from p are correctly found.