Appendices for the NeurIPS submission
Resolving the data ambiguity for periodic crystals

The most important contribution is the introduction of the PDD invariants whose speed, continuity, and generic completeness in Problem 1.1 have no analogs among any state-of-the-art comparisons.

Many descriptors and similarities of crystals use only their finite subsets (only molecules or balls of a manually cut-off radius), which can have infinitely many non-isometric classes as shown in Fig. 4 (left). Any minimally useful descriptor should be invariant as in Problem 1.1a because any rigid motion cannot really change a solid crystalline material. A descriptor with false negatives cannot be used for comparison because equivalent objects can have different values of a non-invariant.

False negatives should be avoided first to detect different crystals $S \not\cong Q$ due to $I(S) \neq I(S)$. Only for an invariant I, it makes sense to avoid false positives by proving the completeness of I in condition (1.1b) to justify the converse implication: $I(S) = I(Q) \Rightarrow S \cong Q$. All 'state-of-the-art' comparisons fail to distinguish many periodic crystals [32, 49, 46], hence cannot define a metric satisfying the coincidence axiom: $d(S, Q) = 0$ only for isometric crystals $S \cong Q$. Appendix A gives specific examples when past similarities RMSD [13] and 1-PXRD [54] fail the triangle axiom.

The traditional crystallography tools such as diffraction patterns and pair distribution function cannot distinguish any homometric crystals [48], even in dimension 1 as shown in Fig. 3 (right).

The simpler continuity in conditions (1.1c,d) is needed to quantify the similarity of (near-)duplicates but is failed by most invariants based on (reduced) cells or symmetry groups, see Fig. 2 (right).

Another inconvenience of the tools such as SOAP (Smooth Overlapped Atomic Positions [4]) and ACSF (Atom-Centered Symmetry Functions [5]) is their dependence on parameters such as cut-off radii that require manual choices. If these parameters change, then so do the underlying invariants.

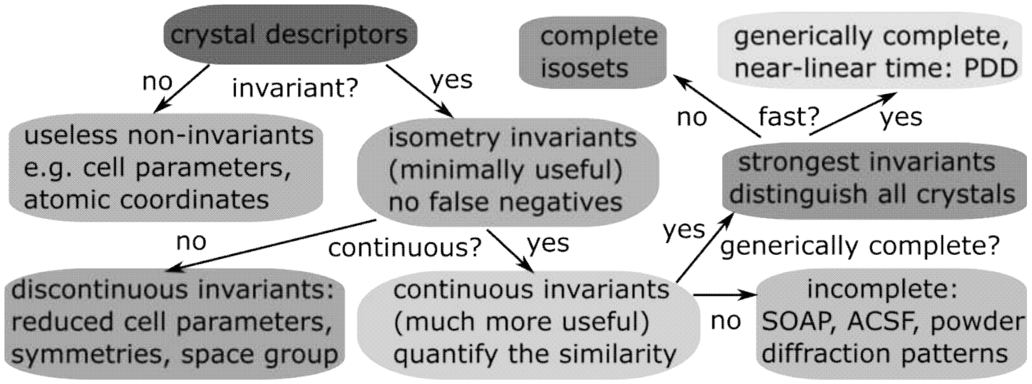
crystal descriptors
no
invariant?
yes
useless non-invariants e.g. cell parameters, atomic coordinates
isometry invariants (minimally useful) no false negatives
no
continuous?
yes
discontinuous invariants: reduced cell parameters, symmetries, space group
continuous invariants (much more useful) quantify the similarity
yes
generically complete?
no
incomplete: SOAP, ACSF, powder diffraction patterns
strongest invariants distinguish all crystals
no
fast?
yes
complete isosets
generically complete, near-linear time: PDD

Figure 7: A hierarchy of crystal descriptors. A minimally useful invariant should have *no false negatives*. Most isometry invariants are discontinuous [60] or incomplete [46], hence fail conditions (1.1bcd) in Problem 1.1. The fastest and generically complete invariants are PDD in this paper.

Only the couple of recently discovered invariants satisfy the practially important conditions (1.1a,b,c,d,f): isosets [3] and PDD in this paper, see Fig. 7. The advantage of PDD is their near-linear time in Theorem 5.1. Pairwise comparisons of millions of real crystals are urgently needed to avoid the paper mills based on duplicates [7], whose new cases are reported in appendix A.

So all 'state-of-the-art' tools were too slow for large experiments such as 200B+ pairwise comparisons of all 660K+ periodic crystals in the Cambridge Structural Database (CSD). Hence it is possible to compare PDD with the past tools only on smaller datasets. The appendices below include

Appendix A: details of extra experimental comparisons with the state-of-the-art tools;

Appendix B: a pseudo-code with examples and instructions for the attached code;

Appendix C: rigorous proofs of all theoretical results in the main paper.

A Appendix A: details of experimental comparisons with the state-of-the-art

This appendix provides extra details for large-scale experimental comparisons of the new invariant PDD with the most widely used past tools: the Root Mean Square Deviation (RMSD) for comparing finite subsets [13] and 1-PXRD similarity based on powder X-ray diffraction patterns. The popularity of these tools does not help to resolve the key question for crystals : 'the same or different?' [54].

The key advantage of Pointwise Distance Distributions (PDD) for searching (near-)duplicates and nearest neighbors in huge datasets is their hierarchical nature. The weighted averages of columns in the PDD matrix form the simpler AMD vectors (Average Minimum Distances), which can be quickly subtracted and compared. Since the Earth Mover's Distance (EMD) between PDD matrices can be only larger than the L_∞ distance between their AMD vectors by Theorem 4.2, it suffices to compute the slower EMD only on pairs of crystals whose AMD vectors are very close to each other.

Any crystal dataset can be further organized into hierarchical families whose levels are parameterized by the number i of atomic neighbors. The i-th level of this hierarchy can consist of disjoint or overlapping clusters of crystals that have close values of the i-th coordinate AMD_i in the vector $\mathrm{AMD}(S;k) = (\mathrm{AMD}_1, \dots, \mathrm{AMD}_k)$. Indeed, $L_\infty(\mathrm{AMD}(S;k), \mathrm{AMD}(Q;k)) \geq |\mathrm{AMD}_i(S) - \mathrm{AMD}_i(Q)|$. Hence, any crystals with distant values of AMD_i cannot have close AMD vectors.

We use the public dataset of 5679 T2 crystals base on the same T2 molecule and predicted by 12-week supercomputer simulations [38]. Only five were synthesized: T2-α, T2-β, T2-γ, T2-δ, T2-ε. The supplementary folders include CIFs of all mentioned crystals, which can be opened by any text editor and visualised by the free software Mercury. Fig. 8 shows 10 PDD curves of T2-ε corresponding to the 10 isometrically unique atomic positions within the T2 molecule. The rotation through 120° around an axis keeps the T2 molecule and the whole infinite crystal T2-ε invariant, while the three oxygen atoms shown in red simply rotate their positions. Hence all oxygen atoms have the same ordered distances to their neighbors and their PDD curves coincide. The same conclusion holds for six nitrogens whose positions are isometrically equivalent in T2. Carbon atoms have five non-isometric positions, which generate five unique PDD curves in Fig. 8. All 10 PDD curves provide a finer classification of atomic types not only by chemical elements, but by their geometric positions within a molecule. For every atom, the first distance for $k = 1$ is the bond distance to its closest neighbor, see the zoomed beginnings of PDD curves. Though chemical elements can be included into the PDD matrix as an extra column, the above shows that the PDD invariants can infer this chemistry.

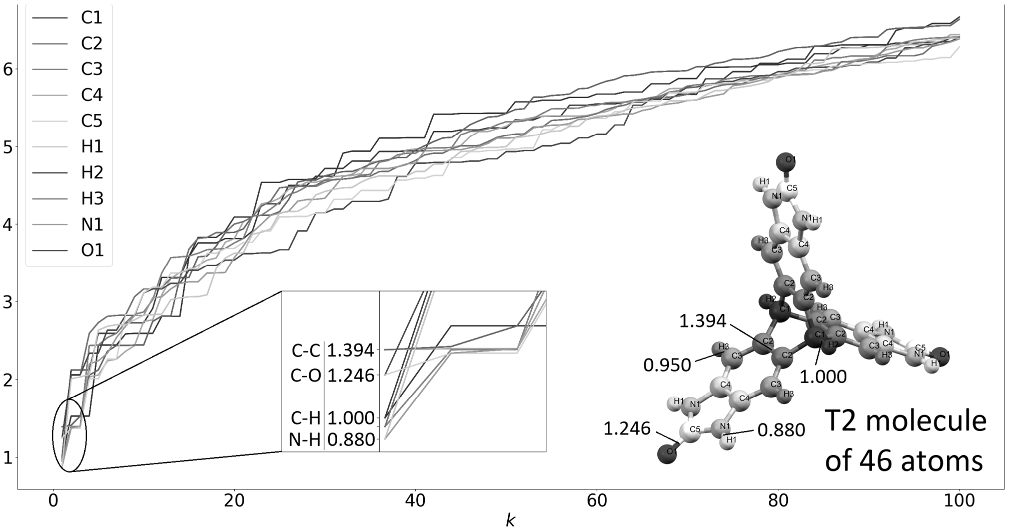

Figure 8: Each curve shows distances d_k from one atom of a T2 molecule to k-th neighbor in the T2-ε crystal. The 46 atoms form 10 groups, one for each atom in the asymmetric unit: three oxygens form one type with the PDD curve in red, 23 carbons have 5 non-isometric positions in T2 with PDD curves in 5 shades of gray. The zoomed part shows the initial bond distances in $\text{Å} = 10^{-10}$ m.

A key challenge in visualization is to justifiably represent invariant data in a simple form. Past algorithms such as t-SNE [59] and UMAP [42] are stochastic and can produce different outputs from repeated runs on the same input. Hence we used the more recent deterministic TreeMap [50] to draw a Minimum Spanning Tree (MST). Informally, any MST can be considered as an optimal road network between cities. Such a network can be drawn in many different ways; the most valuable new data is the distances between PDDs of crystals but the MST is drawn only for visualization. Similarly, positions of cities and distances between them are more important than a sketch of a road network.

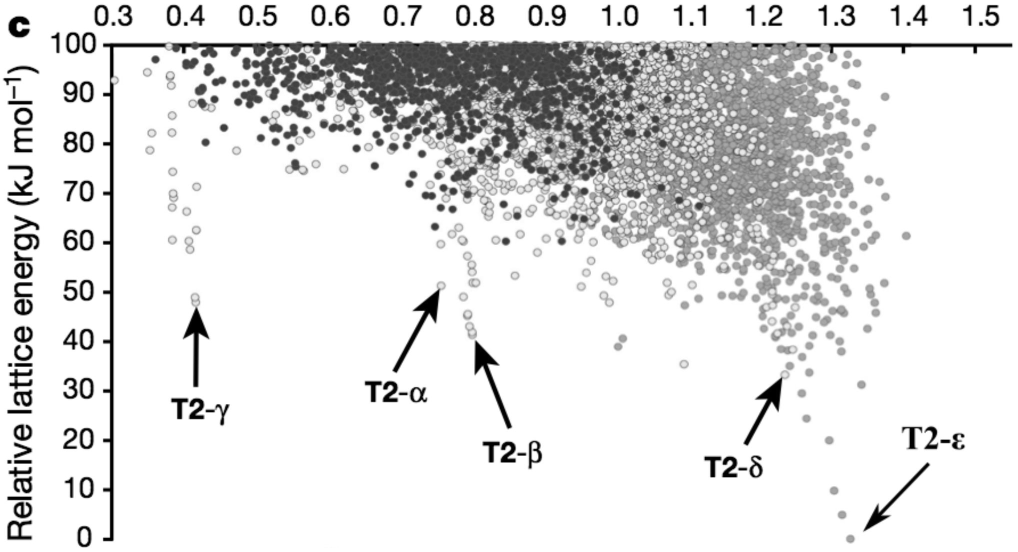

Figure 9: State-of-the-art visualization of a CSP dataset, modified from Fig. 2d in [23]. Every predicted crystal has two coordinates (density, energy). The arrows show the predicted crystals that were manually matched to the five experimental crystals T2-α, T2-β, T2-γ, T2-δ, T2-ε.

Fig. 10 illustrates key advantages of the invariant-based visualization in the right picture compared with the energy landscape in the left picture extracted from the state-of-the-art [23] in Fig. 9. Briefly, the MST based on stronger isometry invariants $\mathrm{PDD}(S; 100)$ provides more information about crystal similarity than the cloud of points (density, energy). In the past, simulated versions were manually searched to match experimental crystals only by density. Measuring the energy of an experimental crystal needs an explosion disintegrating this crystal. For the experimental crystal T2-δ, dozens of simulated crystals were searched in the vertical strip in the left picture of Fig. 10. Simulated crystal 14 was visually chosen in [23] as the best match for T2-δ.

The new PDD invariant automates the search for closest crystals, because PDD can be computed for all types of simulated and experimental crystals. The MST in the right picture of Fig. 10 includes T2-δ as a red dot close to the near duplicate crystals 14 and 15, which were not filtered out by past tools because of very different unit cells and motifs in Fig. 6. There were many other pairs of crystals with almost identical values of (density, energy), for example crystals 5920 and 0049, whose structures were found away from each other in the MST. When comparing this pair by the COMPACK algorithm, only one molecule is matched in Table 4. The resulting zero value of RMSD only means that crystals 5920 and 0049 are different (no similarity found), while EMD provides a proper distance. Fig. 10 (right) shows crystals 5920, 0049 in different branches of the MST.

The 4950 comparisons of 100 crystals by COMPACK took 3 hours 53 min (2.825 sec per comparison), which is 3 orders of magnitude slower than EMD on PDD invariants. Table 3 says that over half the time (56%) only one molecule is matched, which means that no similarity was found between crystals based on identical molecules. In fact, in 95% of cases three or fewer molecules are matched, which is also considered a non-match for the default number of 15 maximum matched molecules. The number of matched molecules rarely exceed 10 (0.5%), and only 6 comparisons matched 15/15.

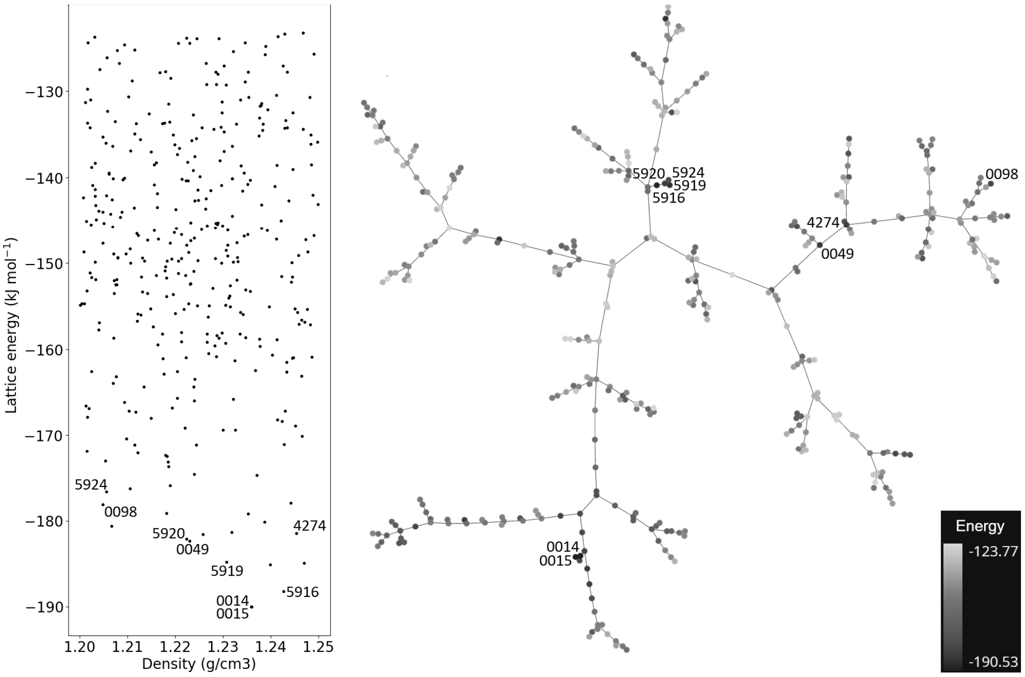

Figure 10: **Left**: a vertical 'strip' of the energy landscape from Fig. 9 in the density range $[1.2, 1.25]$ around the density of the experimental crystal T2-δ. Some of the lowest energy crystals are highlighted by their IDs in the T2 dataset reported in [23]. **Right**: MST of the simulated crystals on the left plus the experimental crystal T2-δ in red, based on $\mathrm{PDD}(S; 100)$. All simulated crystals are colored by lattice energy according to the key in the bottom right corner.

Table 4: The traditional COMPACK algorithm [13] measures similarity by attempting to match up to 15 (by default) molecules from two crystals. The low values 1, 2, 3 (of 15) mean that only a few molecules could be aligned in the two crystals with a small enough Root Mean Square Deviation (RMSD), so this is not considered a match.

	0014	0015	0049	0098	4274	5916	5919	5920	5924
0014	15	15	3	1	1	1	2	2	2
0015	15	15	3	1	1	1	2	2	2
0049	3	3	15	1	10	2	1	1	1
0098	1	1	1	15	1	1	1	1	1
4274	1	1	10	1	15	2	2	2	2
5916	1	1	2	1	2	15	10	10	10
5919	2	2	1	1	2	10	15	15	15
5920	2	2	1	1	2	10	15	15	15
5924	2	2	1	1	2	10	15	15	15

We also compare the new invariant to powder X-ray diffraction patterns (PXRD), a popular way of measuring similarity between structures. We note that unlike Earth mover's distance between PDDs, the 'distance' between PXRDs is not a proper metric on the space of crystals, and so is not appropriate to use it to continuously explore the space. Our experiments show that $1-$PXRD (since a PXRD score of 1 means identical) fails to satisfy the triangle inequality on the T2 dataset over 8% of the time, and the RMSD values given by COMPACK similarly fail to satisfy this axiom, see Table 5.

Our second dataset has all 12576 structures in the Drug Subset [12] of the Cambridge Structural Database (CSD). Though the CSD is the world's largest collection of more than 1.1M crystals, it is effectively a huge folder of Crystallographic Information Files (CIF) with limited search functions, for example by a chemical composition.

Crystal 1	Crystal 2	RMSD
T2-γ	T2-α	2.805
T2-γ	T2-ϵ	0.184
T2-α	T2-ϵ	0.231

Crystal 1	Crystal 2	$1-$PXRD
T2-0001	T2-0029	0.187
T2-0001	T2-5333	0.110
T2-0029	T2-5333	0.0683

Table 5: Triples of crystals where the commonly used tools RMSD and PXRD fail the traingle inequality axiom required of a proper metric. The CIFs are in the supplementary materials.

The CSD Drug subset only includes small-molecule crystal structures containing a drug molecule, typically smaller than the T2 molecule. Since larger than necessary values of k don't affect distances to closer neighbors, we tried $k = 100$ and $k = 200$ for the CSD Drug Subset.

After computing $\mathrm{PDD}(S; k)$, we use a standard EMD algorithm for distances between the distributions. A Minimum Spanning Tree can be computed from a smaller graph on PDD invariants not requiring all pairwise distances. Since distances asymptotically approach $\mathrm{PPC}(S)\sqrt[3]{k}$, we used the instantly computable Point Packing Coefficient $\mathrm{PPC}(S)$ to find 3000 neighbors for each of 12576 crystals. For $k = 200$, the computations of 12576×3000 EMD distances (excluding all repetitions) took 7 hours 39 min (1.238 ms per comparison, 2.19 sec per crystal for all 3000 neighbors).

The key contribution is the new invariant data for edge-lengths of the MST, which can be visualized by TreeMap [50] or any other graph drawing algorithm. The continuity and invariance of Pointwise Distance Distributions in Theorems 3.2 and 4.3 guarantee that any crystal data analysis based on these invariants is theoretically justified. Since a Minimum Spanning Tree has no cycles, some close crystals might appear at terminal vertices of distant branches, which could be connected to form a cycle. Drawing a graph with cycles will require more sophisticated algorithms. However, the new invariants can be used for a justifiable search of similar crystals without visualization.

We also have interactive versions of full Minimum Spanning Trees of the T2 dataset and CSD drug subset. In any Javascript-enabled browser one can zoom these TreeMaps, color the vertices by properties, hover over any vertex to get more information. Though aspirin and paracetamol are chemically different, their crystal structures are found in close branches confirming their similar properties.

The 'structure' of a crystal was previously measured by only one continuous isometry invariant: the physical density equal to the weight of atoms within a unit cell, divided by the cell volume. All other descriptors are either non-invariants, e.g. powder diffraction patterns depend on various thresholds and are not always reliable, or are discontinuous under perturbations, e.g. symmetry groups.

The main contribution is replacing the single-value density by much stronger more isometry invariants PDD. Comparing crystals by only their densities was the latest 'state-of-the-art' in [23], where 5 forms of 9 experimental crystals were initially matched to dozens of simulated crystals with close densities. Then hundreds of crystals were visually inspected to select the five matches in Fig. 9.

Now a 'structure' of a crystal can be much better represented by many more numerical invariants in PDD. Fig. 11 is the first justified map of (a discretely sampled) space of T2 crystals, because the underlying distance between PDD invariants satisfies all metric axioms and continuity under perturbations by [52] and Theorem 4.3.

The map in Fig. 11 shows a Minimum Spanning Tree built on 5679+9 crystals. Each crystal is a vertex. The distance between any crystals is the Earth Mover's distance between the first order PDD invariants. The most interesting vertices in the T2 map are experimental crystals shown in red, because their energies are impractical to measure. The map shows experimental crystals with simulated predictions for the first time, because past plots such as in Fig. 9 need all energy values.

Fig. 11 is a substantial advancement in visualizing crystal datasets, because the underlying invariants $\mathrm{PDD}(S; k)$ are much more informative than a single-value density. The nine experimental crystals (red dots) clearly split into five groups marked by T2-α, T2-β, T2-γ, T2-δ, T2-ε. These nine experimental crystals have no known energies and cannot appear in the past energy landscapes. The four red dots at the very top represent slightly different forms of a T2-γ crystal synthesized at different temperatures.

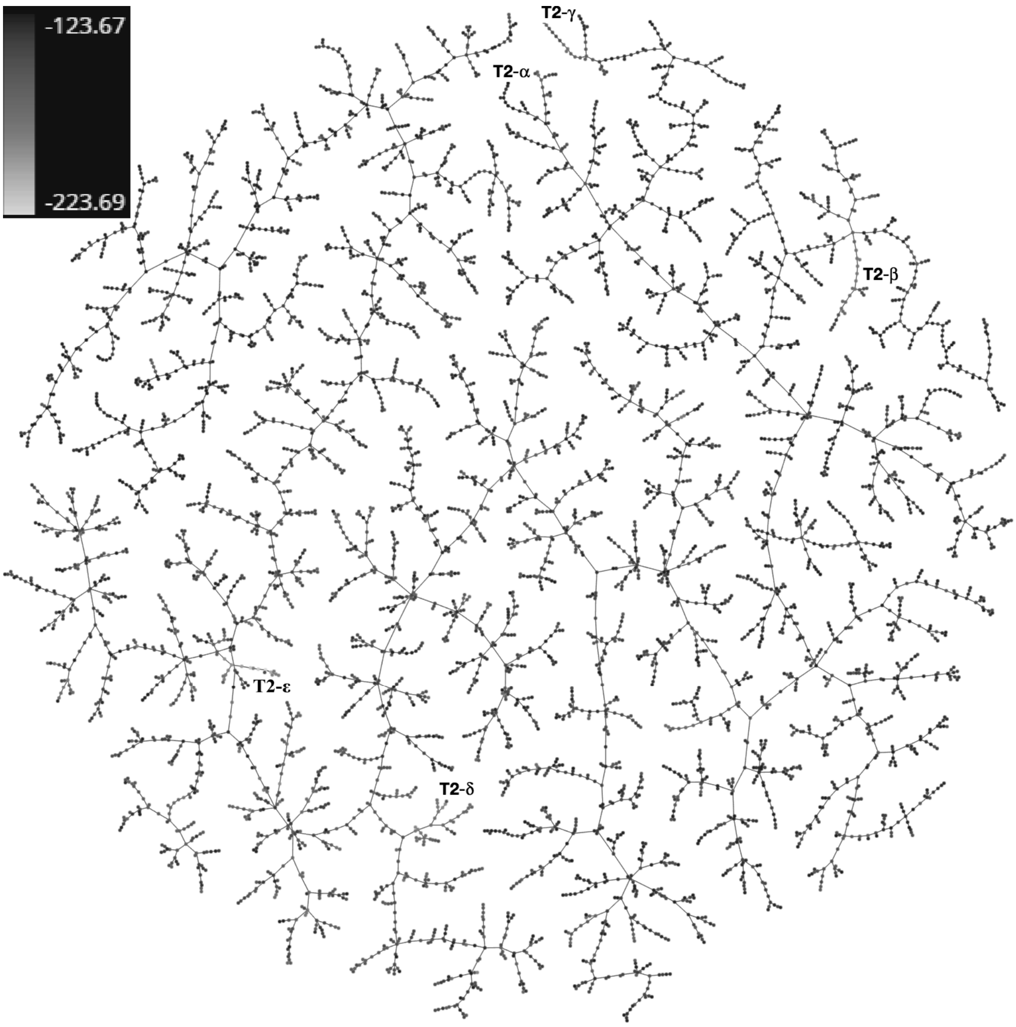

Figure 11: A Minimum Spanning Tree of 9+5679 T2 crystals drawn by TreeMap [50]. The distance between any crystals S, Q is the $\mathrm{EMD}(\mathrm{PDD}(S; 125), \mathrm{PDD}(Q; 125))$. The color bar in the top left corner shows energy values. The minimum of -223.69 kJ/mol corresponds to 0 in Figure 9. Nine experimental crystals, which form the families T2-α, T2-β, T2-γ, T2-δ, T2-ε, are shown in red, because their energies are not experimentally measured (a crystal is disintegrated only by a blow up).

The map can be interactively explored by opening the file T2L_PDD125_TMap.html from the supplementary materials in any browser in the same folder with the corresponding JavaScript files. In the interactive version one can color vertices by a property, which can be changed in the drop-down menu in the bottom right corner. For example, all crystals with low energies appear in yellow colors.

In Fig. 11 six of nine experimental T2 crystals have the codes DEBXIT01. . . 06, though they split into two groups (unique polymorphs): four nearly identical crystals T2-γ and two nearly identical crystals T2-β. Our final experiment on all 229K molecular organic crystals found thousands of such pairs in the Cambridge Structural Database.

In addition to the five pairs of isometrically identical crystals discussed in section 6, a search on the CSD using PDD on molecular centers instead of atomic centers revealed nine pairs of structures which were almost identical, listed in table 6. The supplementary code includes a script to compare the previously mentioned five pairs by PDD, however the code to compute molecular centers is

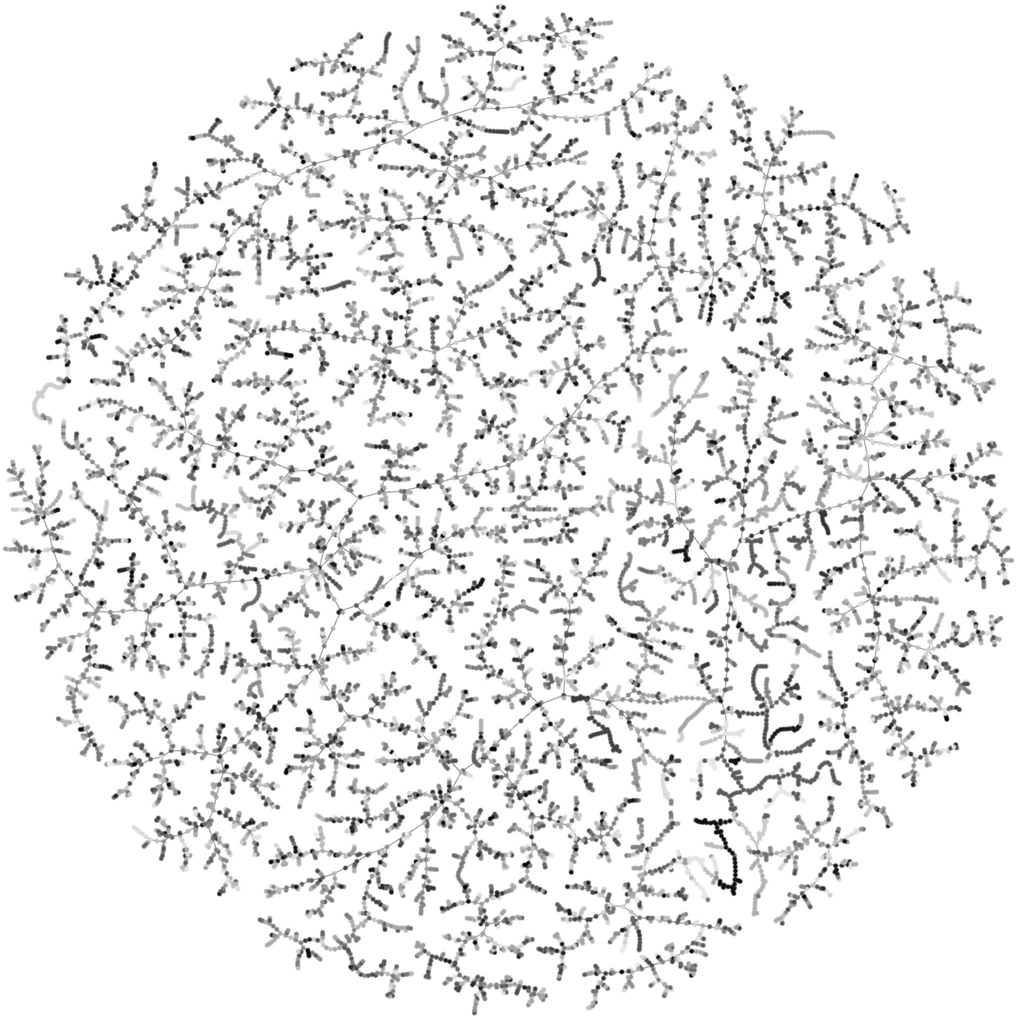

Figure 12: A Minimum Spanning Tree of 12576 structures in the Drug Subset of the Cambridge Structural Database. The distance between any crystals S, Q is $\mathrm{EMD}(\mathrm{PDD}(S; 100), \mathrm{PDD}(Q; 100))$. All structures from the same family are shown in one random color.

private, so it was not possible to include a script reproducing the results below. Instead, the .CIFs for each of these crystals is included in the folder `identical_by_mol_centers` and can be manually inspected (or opened in visualization software) to see the similarity.

B Appendix B: examples and instructions for the attached code and data files

B.1 Pseudocode for computing Pointwise Distance Distributions (PDD)

The algorithm accepts any periodic point set $S \subset \mathbb{R}^n$ in the form of a unit cell U and a motif $M \subset S$. The cell is given as a square $n \times n$ matrix with basis vectors in the columns, and the motif points in Cartesian form lying inside the unit cell. For dimension 3, the typical Crystallographic Information File (CIF) with six unit cell parameters and motif points in terms of the cell basis is easily converted to this format. Otherwise, the unit cell and motif points can be given directly, in any dimension.

Specifically, the PDD function's interface is as follows:

Input:

CSD ID 1	CSD ID 2	EMD ($k = 100$)
ADESAG	REWPOB	1.45×10^{-5}
ROWBIQ	TUHMOA	9.00×10^{-4}
BESNOC01	XILXIA	3.93×10^{-3}
GIMHOA	XORCOX	5.63×10^{-3}
FUQMAG	GAVFEP	6.73×10^{-3}
COXJIJ	DUPZIY	6.79×10^{-3}
HAPNOE	HAPNUK	7.15×10^{-3}
LETYAM	UKUWUT	8.66×10^{-3}
DUCPUN	XUWMEI01	9.88×10^{-3}

Table 6: Selection of pairs of structures from the CSD with exactly zero EMD when compared by molecular centers. The closest pair ADESAG and REWPOB have identical fractional coordinates except for one difference at the fifth decimal place, a difference 100 times smaller than the typical threshold for considering two sites identical (0.001Å).

- `motif`: array shape (m, n). Coordinates of motif points in Cartesian form.
- `cell`: array shape (n, n). Represents the unit cell in Cartesian form.
- `k`: `int` > 0. Number of columns to return in PDD$(S; k)$.

Output:

- pdd: array with $k + 1$ columns.

Before giving the pseudocode, we outline some of the key objects and functions in use:

- A generator g, which creates points from the periodic set S to find distances to,
- KDTrees (canonically k is the dimension here, in our case it's denoted n), data structures designed for fast nearest-neighbour lookup in n-dimensional space.

Once g is constructed, `next(g)` is called to get new points from the infinite set S. The first call returns all points in the given unit cell (i.e. the motif), and successive calls returns points from unit cells further from the origin in a spherical fashion.

A KDTree is constructed with a point set T, then queried with another Q, returning a matrix with distances from all points in Q to their nearest neighbors (up to some given number, k below) in T, as well as the indices of these neighbors in T.

The functions `collapse_equal_rows` and `lexsort_rows`, which perform the collapsing and lexicographical sorting steps of computing PDD (Definition 3.1) respectively, are assumed to be implemented elsewhere.

The following pseudocode finds PDD$(S; k)$ for a periodic set S described by `motif` and `cell`:

```
def PDD(motif, cell, k):

    cloud = [] # contains points from S
    g = point_generator(motif, cell)

    # at least k points will be needed
    while len(cloud) < k:
        points = next(g)
        cloud.extend(points)

    # first distance query
    tree = KDTree(cloud)
    D_, inds = tree.query(motif, k)
    D = zeros_like(D_)

    # repeat until distances don't change,
    # then all nearest neighbors are found
    while not D == D_:
        D = D_
        cloud.extend(next(g))
        tree = KDTree(cloud)
        D_, inds = tree.query(motif, k)

    pdd = collapse_equal_rows(D_)
    pdd = lexsort_rows(pdd)
    return pdd
```

B.2 Instructions for the attached PDD code and specific examples

A Python script implementing Pointwise Distance Distributions along with examples can be found in the zip archive included in this submission. Python 3.7 or greater is required. The dependency packages are NumPy (< 1.22), SciPy ($\geq 1.6.1$), numba ($\geq 0.55.0$) and ase ($\geq 3.22.0$); if you do not wish to affect any currently installed versions on your machine, create and activate a virtual environment before the following.

Unzip the archive and in a terminal navigate to the unzipped folder. Install the requirements by running `pip install -r requirements.txt`. Then run `python` followed by the example script of choice, and then any arguments (outlined below), e.g.

```
$ python kite_trapezium_example.py

trapezium: [(0, 0), (1, 1), (3, 1), (4, 0)]
PDD:
[[0.5        1.41421356 2.         3.16227766]
 [0.5        1.41421356 3.16227766 4.        ]]

kite: [(0, 0), (1, 1), (1, -1), (4, 0)]
PDD:
[[0.25       1.41421356 1.41421356 4.        ]
 [0.5        1.41421356 2.         3.16227766]
 [0.25       3.16227766 3.16227766 4.        ]]

EMD between trapezium and kite: 0.874032
```

List of included example scripts and their parameters:

- `kite_trapezium_example.py` prints the PDDs of the finite 4-point sets K (kite) and T (trapezium) in Fig. 3, along with their Earth mover's distance.
- `1D_sets_example.py` shows that the 1D periodic sets in Fig. 3 are distinguished by their PDDs for any parameter $0 < r \leq 1$. This script requires the parameter r to be passed after the file name, e.g. '`python 1D_sets_example.py 0.5`'.
- `T2_14_15_example.py` compares the crystals shown in Fig. 6, whose original .CIFs are included. This optionally accepts the parameter k controlling the number of columns in the computed PDD, e.g. '`python T2_14_15_example.py --k 50`' compares by PDD with $k = 50$. If not included, $k = 100$ is used as the default.
- `CSD_duplicates_example.py` computes and compares the PDDs of the 5 pairs of isometric crystals from the CSD discussed in section 6, giving distances of exactly zero. This optionally accepts the parameter k controlling the number of columns in the computed PDD, in the same way as `T2_14_15_example.py` above.

If you wish to run the code on your own sets or CIF files, you can use the functions exposed in the main script `pdd.py`. Use `pdd.read_cif()` to parse a cif and return a crystal, or define one manually as a tuple `(motif, cell)` with NumPy arrays. Pass this as the first argument to `pdd.pdd()` with an integer `k` as the second to compute the PDD. Pass two PDDs to `pdd.emd()` to calculate the Earth mover's distance between them. For finite sets, the function `pdd.pdd_finite()` accepts just one argument, an array containing the points, and returns the PDD.

C Appendix C: rigorous proofs of Theorems 3.2, 4.2, 4.3, 4.4, 5.1.

Proof of Theorem 3.2. For any periodic point set $S \subset \mathbb{R}^n$, we first show scaling up a unit cell U to a non-primitive cell keeps $\mathrm{PDD}(S;k)$ invariant. It suffices to scale up a cell U by a factor of l, say along the first basis vector $\vec{v}_1$ of U, then the number m of motif points of S is multiplied by l.

Then $D(S;k)$ from Definition 3.1 has the larger size $lm \times k$ but (due to periodicity) consists of l-tuples of identical rows of distances from points $p + i\vec{v}_1$, $i = 0, \dots, l-1$, to their k neighbors within S. Hence $\mathrm{PDD}(S;k)$ remains invariant, under all isometries due to the result below.

Now we show that the matrix $D(S;k)$ from Definition 3.1, hence $\mathrm{PDD}(S;k)$, is independent of a primitive unit cell. Let U, U' be primitive cells of a periodic set $S \subset \mathbb{R}^n$ with a lattice Λ. Any point $q \in S \cap U'$ can be translated by some $\vec{v} \in \Lambda$ to a point $p \in S \cap U$ and vice versa. These translations establish a bijection between the motifs $S \cap U \leftrightarrow S \cap U'$ and preserve distances. So $\mathrm{PDD}(S;k)$ is the same for both U, U'.

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

For any finite or periodic sets S, Q, since f maintains distances, every list of ordered distances from $p_i \in S \cap U$ to its first k nearest neighbors in S, coincides with the list of the ordered distances from $f(p_i)$ to its first k neighbors in Q. These coincidences of distance lists give $\mathrm{PDD}(S;k) = \mathrm{PDD}(Q;k)$ after collapsing identical rows. ☐

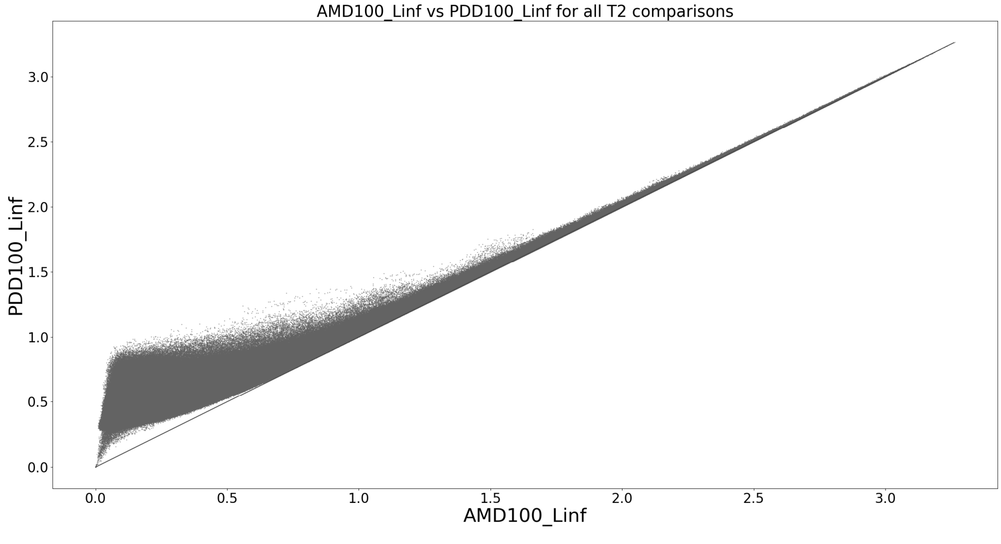

Figure 13: Illustration of Theorem 4.2. The coordinates of each point are the L_∞-based distances between $\mathrm{AMD}(S;100), \mathrm{AMD}(Q;100)$ and $\mathrm{PDD}(S;100), \mathrm{PDD}(Q;100)$ for all pairs S, Q of 5679 T2 crystals reported in [23].

Proof of Theorem 4.2. Considering $\mathrm{PDD}(S;k)$ as a weighted distribution of rows, $\mathrm{AMD}(S;k)$ is its centroid from [14, section 3]. The lower bound follows from [14, Theorem 1]. ☐

Lemma C.1. *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*-th 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 $b_1, \dots, b_i \in Q$ of a belong to the open ball with the center a and the radius $|a - a_i| - 2\varepsilon$.

Since the bijection g shifted every $b_1, \dots, b_i$ 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 is among these i preimages, i.e. the distance from a to its i-th nearest neighbor should be strictly less than the assumed value $|a - a_i|$. We get a contradiction assuming that the distance from a to its new i-th neighbor b_i is more than $|a - a_i| + 2\varepsilon$. □

Lemma C.2. *For* $\varepsilon > 0$*, let* $g : S \to Q$ *be a bijection between finite or periodic sets so that* $|a - g(a)| \leq \varepsilon$ *for all* $a \in S$*. Then* g *changes the vector* $\vec{R}_a(S) = (|a - a_1|, \dots, |a - a_k|)$ *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 at most* 2ε *in the* L_∞*-distance. So if* $b_1, \dots, b_k \in Q$ *are the* k *nearest neighbors of* $b = g(a)$ *within* Q *and* $\vec{R}_b(S) = (|b - b_1|, \dots, |b - b_k|)$ *is the vector of the first* k *minimum distances from* $b = g(a)$*, then the* L_∞*-distance* $|\vec{R}_a(S) - \vec{R}_b(Q)|_\infty \leq 2\varepsilon$*.* ■

Proof. By Lemma C.1 every coordinate of the vector $\vec{R}_a(S)$ changes by at most 2ε. Hence the L_∞-distance from $\vec{R}_a(S)$ to the perturbed vector $\vec{R}_b(Q)$ is at most 2ε. □

Proof of Theorem 4.3. The bottleneck distance is $d_B(S, Q) = \inf\limits_{g : S \to Q} \sup\limits_{a \in S} |a - g(a)|$ between point sets S, Q. Then for any $\delta > 0$ there is a bijection $g : S \to Q$ such that $\sup\limits_{a \in S} |a - g(a)| \leq d_B(S, Q) + \delta$. If the given sets S, Q are finite, one can set $\delta = 0$. Indeed, there are only finitely many bijections $S \to Q$, hence the infimum in the definition above is achieved for one of them.

By [21, Lemma 4.1], if the sets S, Q are periodic, they have a common lattice Λ. Any primitive cell U of Λ is a unit cell of S, Q, i.e. $S = \Lambda + (S \cap U)$ and $Q = \Lambda + (Q \cap U)$. Since the bottleneck distance $\varepsilon = d_B(S, Q) < \dfrac{r(S)}{4}$, we can define a bijection g from every point $a \in S$ to a closest point $g(a) \in Q$. If U is a non-primitive unit cell of S, the distance matrix $D(S; k)$ can be constructed as in Definition 3.1, but each row will be repeated $n(S)$ times, where $n(S)$ is $\mathrm{Vol}[U]$ divided by the volume of a primitive unit cell of S. So we can assume that S, Q share a unit cell U and have in U the same number $m(S) = m(Q)$, say both are equal to m. For any $k \geq 1$, we first define the simple 1-1 flow from the rows of $D(S; k)$ to the rows of $D(Q; k)$ by setting $f_{ii} = \frac{1}{m}$ and $f_{ij} = 0$ for $i \neq j$, where $i, j = 1, \dots, m$. Recall that Definition 3.1 collapses all rows of $D(S; k)$ that are identical to each other to a single row, similar for $D(Q; k)$. By summing up weights of collapsed rows, the above flow induces a flow from all distance vectors in $\mathrm{PDD}(S; k)$, e.g. $R_i(S)$ in the i-th row of $\mathrm{PDD}(S; k)$, to all distance vectors in $\mathrm{PDD}(Q; k)$.

Then $\mathrm{EMD}(\mathrm{PDD}(S; k), \mathrm{PDD}(Q; k)) \leq \frac{1}{m} \sum\limits_{i=1}^{m} |\vec{R}_i(S) - \vec{R}_i(Q)|_\infty$, because EMD minimizes the cost over all flows in Definition 4.1. Since $|\vec{R}_i(S) - \vec{R}_i(Q)| \leq 2(d_B(S, Q) + \delta)$ by Lemma C.2, we get $\mathrm{EMD}(\mathrm{PDD}(S; k), \mathrm{PDD}(Q; k)) \leq \frac{1}{m} \sum\limits_{i=1}^{m} 2(d_B(S, Q) + \delta) = 2(d_B(S, Q) + \delta)$. Since the last inequality holds for any small $\delta > 0$, we get $\mathrm{EMD}(\mathrm{PDD}(S; k), \mathrm{PDD}(Q; k)) \leq 2d_B(S, Q)$. □

For any point p in a lattice $\Lambda \subset \mathbb{R}^n$, the open *Voronoi domain* $V(\Lambda; p) = \{q \in \mathbb{R}^n \mid |q - p| < |q - p'| \text{ for any } p' \in \Lambda - p\}$ is the neighborhood of all points $q \in \mathbb{R}^n$ that are strictly close to p than to all other points p' of the lattice Λ [22].

The Voronoi domains $V(\Lambda;p)$ of different points $p \in \Lambda$ are disjoint translation copies of each other and their closures tile $\mathbb{R}^n$, so $\cup_{p\in\Lambda}\bar{V}(\Lambda;p) = \mathbb{R}^n$.

For a generic lattice $\Lambda \subset \mathbb{R}^2$, $V(\Lambda;p)$ is a centrally symmetric hexagon. Points $p, p' \in \Lambda$ are *Voronoi neighbors* if their Voronoi domains share a boundary point, so $\bar{V}(\Lambda;p) \cap \bar{V}(\Lambda,p') \neq \emptyset$.

Below we always assume that any lattice Λ is shifted to contain the origin 0, also any periodic point set $S = \Lambda + M$ has a point at 0.

Definition C.3 (neighbor set $N(\Lambda)$ and basis distances)**.** *For any lattice $\Lambda \subset \mathbb{R}^n$, the* neighbor set *(of the origin 0) is $N(\Lambda) = \Lambda \cap \bar{B}(0;r) - \{0\}$ for a minimum radius r such that $N(\Lambda)$ is not contained in any affine $(n-1)$-dimensional subspace of $\mathbb{R}^n$ and $N(\Lambda)$ includes all $n+1$ nearest neighbors (within Λ) of any point $q \in V(\Lambda;p)$.*

For any point $q \in V(\Lambda;0)$, consider all n-tuples $(p_1,\dots,p_n)$ of points $p_i \in N(\Lambda)$ such that the vectors $\vec{p}_1,\dots,\vec{p}_n$, form a linear basis of $\mathbb{R}^n$. Order $p_1,\dots,p_n$ by their distances to q. Choose a lexicographically smallest list of basis distances *$d_1(q) \leq \dots \leq d_n(q)$ from the point q over all n-tuples $(p_1,\dots,p_n)$ described above.* ■

The lattice $\mathbb{Z}^2$ has the neighbor set $N(\mathbb{Z}^2) = \{(\pm1,0),(0,\pm1)\}$. If Λ is generated by $(2,0),(0,1)$, the neighbor set $N(\Lambda) \subset \Lambda$ includes the 3rd neighbors $(0,\pm2)$ of the points $(0,\pm0.4) \in V(\Lambda;0)$.

Indeed, if Definition C.3 has a radius $r < 2$, then $\Lambda \cap \bar{B}(0;r) - \{0\} = \{(0,\pm1)\}$ but the y-axis does not generate $\mathbb{R}^2$. For $q = (0,0.4)$, considering all pairs (p_1,p_2) among the four possibilities $((0,\pm1),(\pm2,0))$, we find the basis distances $d_1(q) = 0.6 < d_2(q) = \sqrt{0.4^2+2^2}$ for the 2nd and 3rd lattice neighbors $p_1 = (0,1)$ and $p_2 = (\pm2,0)$ of q.

Lemma C.4 ($N(\Lambda)$ bounds)**.** *The neighbor set $N(\Lambda)$ of any lattice Λ is covered by $\bar{B}(0;2R(\Lambda))$, where the* packing radius *$R(\Lambda)$ is the minimum $R > 0$ such that $\cup_{p\in\Lambda}\bar{B}(p;R) = \mathbb{R}^n$.* ■

Proof of Lemma C.4. Any point p in the closure $\bar{V}(\Lambda;0)$ has $n+1$ lattice neighbors (within Λ) among the origin $0 \in \Lambda$ and at least $2(2^n-1)$ Voronoi neighbors of 0.

In $\mathbb{R}^n$, any vertex of the boundary of $\bar{V}(\Lambda;0)$ is equidistant to at least $n+1$ points of Λ (the origin 0 and its n Voronoi neighbors). The longest of these distances is the covering radius $R(\Lambda)$. The closed ball $\bar{B}(0;2R(\Lambda))$ covers all Voronoi neighbors of 0, hence all points of $N(\Lambda)$. □

The condition of a linear basis in Definition C.3 guarantees that $n+1$ linearly independent vectors $\vec{p}_1,\dots,\vec{p}_n$ uniquely identify a point q by their basis distances $d_1(q),\dots,d_n(q)$.

Definition C.5 (a distance-generic periodic point set)**.** *A periodic point set $S = \Lambda + M \subset \mathbb{R}^n$ with the origin $0 \in \Lambda \subset S$ is called* distance-generic *if the following conditions hold.*

(C.5a) $\vec{p},\vec{q}$ are not orthogonal for any points $p,q \in S \cap V(\Lambda;0)$.

(C.5b) For any vectors $\vec{u},\vec{v}$ between any two pairs of points in S, if $|\vec{u}| = l|\vec{v}| \leq 2R(\Lambda)$ for $l = 1,2$, then $\vec{u} = \pm l\vec{v}$ and $\vec{v} \in \Lambda$.

(C.5c) For any point $q \in V(\Lambda;0)$, let d_0 be the distance from q to its closest neighbor $p_0 = 0$ within Λ. Take any $p_1,\dots,p_n$ in the neighbor set $N(\Lambda)$ with distances $d_1 \leq \dots \leq d_n$ to q.

The $n+1$ spheres $C(p_i;d_i)$ with the centers p_i and radii d_i, $i = 0,\dots,n$, can meet at the single point $q \in V(\Lambda;0)$ only if $d_1 \leq \dots \leq d_n$ are the basis distances of q, hence $\vec{p}_1,\dots,\vec{p}_n$ form a linear basis of $\mathbb{R}^n$, only for at most two tuples $p_1,\dots,p_n \in N(\Lambda)$ symmetric in 0. ■

Condition (C.5b) means that all inter-point distances are distinct apart from necessary exceptions due to periodicity. Since any periodic set $S = \Lambda + M \subset \mathbb{R}^n$ is invariant under translations along vectors of its lattices Λ, condition (C.5b) for $|\vec{v}| \leq 2\mathrm{diam}[U]$ can be checked only for vectors from all points in the Voronoi domain $V(\Lambda;0)$ to all points in the extended domain $3V(\Lambda;0)$.

Condition (C.5b) allows us to recognize *lattice distances* from any point $p \in M$ to its lattice translates $\Lambda + p$ in the row of $\mathrm{PDD}(S;k)$ representing p. Indeed, only a lattice distance d appears in the row together with $2d$ (and possibly with higher multiples) by condition (C.5b). Any lattice distance d and its multiple are repeated twice in every row, because any lattice is centrally symmetric.

All conditions of Definition C.5 can be written as algebraic equations via coordinates of motif points and basis vectors of a unit cell. Almost all $n+1$ spheres in $\mathbb{R}^n$ have no common points, so condition (C.5c) forbids very singular situations, which can be practically checked since the neighbor set $N(\Lambda)$ is finite for any lattice Λ containing 0. Hence any periodic point set can be made distance-generic by almost any perturbation of points and lattice basis.

Proof of Theorem 4.4. Assuming that $\mathrm{PDD}(S;k)$ is realizable by a periodic point set $S=\Lambda+M$, we will reconstruct all motif points $p\in V(\Lambda;0)$, uniquely up to the central symmetry of $\mathbb{R}^n$ with respect to 0. The given number m of points in a unit cell U of S is a common multiple of all denominators in rational weights of the rows in the given matrix $\mathrm{PDD}(S;k)$. Enlarge $\mathrm{PDD}(S;k)$ replacing every row of a weight w by mw identical rows of weight $\frac{1}{m}$.

One can assume that the origin $0\in\Lambda$ belongs to the motif M of S and is represented by the first row of $\mathrm{PDD}(S;k)$. If $\mathrm{PDD}(S;k)$ has $m\geq 2$ rows, we will reconstruct all other $m-1$ points of S within the open Voronoi domain $V(\Lambda;0)$. No points of S can be on the boundary of $V(\Lambda;0)$ due to condition (C.5b) on distinct distances.

Remove from each row of $\mathrm{PDD}(S;k)$ all *lattice distances* between any points of Λ. Then every remaining distance is between only points $p,q\in S$ such that $p-q\notin\Lambda$. Any point $q\in S\cap V(\Lambda;0)-\{0\}$ has its first lattice neighbour 0 at the distance $d_0=|q|$ and a lexicographically smallest list of basis distances $d_1(q)<\dots<d_n(q)$ from q to its further n lattice neighbors $p_1,\dots,p_n\in N(\Lambda)\subset\Lambda-0$ such that the vectors $\vec{p}_1,\dots,\vec{p}_n$ form a basis of $\mathbb{R}^n$. All basis distances are distinct due to (C.5b). By Lemma C.4 they appear once in both rows of the points $0,q\in S$ in $\mathrm{PDD}(S;k)$ after $d_0=|q|$.

Though the basis distances of q may not be the n smallest values appearing after $d_0=|q|$ in the first and second rows of 0 and q, we will try all n-distance subsequences $d'_1<\dots<d'_n$ shared by both rows. Similarly, we cannot be sure that $n+1$ closest neighbors of q in Λ form an affine basis of $\mathbb{R}^n$. Hence we try all n-tuples of points $p_1,\dots,p_n\in N(\Lambda;0)$ whose vectors form a linear basis of $\mathbb{R}^n$.

For all finitely many choices above, we check if the $n+1$ spheres $S(p_i;d'_i)$, which are 1D circles (for $n=2$) or 2D spheres (for $n=3$), meet at a single point in $V(\Lambda;0)$, which is the reconstructed q.

Condition (C.5c) guarantees that these $n+1$ spheres can intersect at a single point in the open Voronoi domain $V(\Lambda;0)$ only if the three conditions of Definition C.5 hold. Firstly, the vectors $\vec{p}_1,\dots,\vec{p}_n$ should form a linear basis of $\mathbb{R}^n$. Secondly, if some distances $d_1<\dots<d_n$ are the basis distances from q to $p_1,\dots,p_n$ only if this list is the lexicographically smallest over all tuples $\{p_1,\dots,p_n\}\subset N(\Lambda)$ that form a linear basis. Thirdly, the single-point intersection happens only for two subsets $\{p_1,\dots,p_n\}\subset N(\Lambda)$ related by the central symmetry with respect to 0. This symmetry is an isometry preserving the lattice Λ and the distances $d_0<d_1<\dots<d_n$. Making a choice by this inevitable ambiguity, we uniquely identify a point $q\in S\cap V(\Lambda;0)-\{0\}$ relative to the fixed Λ.

If the matrix $\mathrm{PDD}(S;k)$ has $m\geq 3$ rows, any further point $p\in(S-\{0,q\})\cap V(\Lambda;0)$ will be uniquely determined as follows. Similarly to the point q above, we determine a position of p using its basis distances $d_0(p)<d_1(p)<\dots<d_n(p)$ to points $0=p_0,p_1,\dots,p_n\in N(\Lambda)$. At the end of reconstruction, we have a final choice between $\pm p$ symmetric with respect to the origin 0.

Since the second point q is already fixed, the third point p is also restricted by the distance $|p-q|$ appearing once only in the second and third rows of $\mathrm{PDD}(S;k)$. The distance $|p-q|$ doesn't help to resolve the ambiguity between $\pm p$ only if q belongs to the bisector of points equidistant to $\pm p$. In this case, $p,0,q$ form a right-angle triangle, which is forbidden by condition (C.5a). Hence p and any further point of $S\cap V(\Lambda;0)$ is uniquely determined by q and Λ. □

Proof of Theorem 5.1. It follows from [25, Theorem 14], which finds k nearest neighbors of m points in (a motif of) S before averaging these k-th distances for $\mathrm{AMD}(S;k)$. Instead of averaging, to get $\mathrm{PDD}(S;k)$, it remains only need to lexicographically sort m lists of ordered distances in time $O(km\log m)$. Indeed, a comparison of two ordered lists of the length k takes $O(k)$ time. □

References

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

[2] O Anosova and V Kurlin. Introduction to periodic geometry and topology. *arXiv:2103.02749*, 2021.

[3] O Anosova and V Kurlin. An isometry classification of periodic point sets. In *Proceedings of Discrete Geometry and Mathematical Morphology*, 2021.

[4] Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B*, 87(18):184115, 2013.

[5] Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of chemical physics*, 134(7):074106, 2011.

[6] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *Transactions PAMI*, 24(4):509–522, 2002.

[7] David Bimler. Better living through coordination chemistry: A descriptive study of a prolific papermill that combines crystallography and medicine. 2022.

[8] M. Bouniaev and N. Dolbilin. Regular and multi-regular t-bonded systems. *J. Information Processing*, 25:735–740, 2017.

[9] M. Boutin and G. Kemper. On reconstructing n-point configurations from the distribution of distances or areas. *Adv. Appl. Math.*, 32(4):709–735, 2004.

[10] Peter Brass and Christian Knauer. Testing the congruence of d-dimensional point sets. In *Proceedings of SoCG*, pages 310–314, 2000.

[11] Matthew Bright, Andrew I Cooper, and Vitaliy Kurlin. Geographic-style maps for 2-dimensional lattices. *arxiv:2109.10885*, 2021.

[12] Mathew J. Bryant, Simon N. Black, Helen Blade, Robert Docherty, Andrew G.P. Maloney, and Stefan C. Taylor. The csd drug subset: The changing chemistry and crystallography of small molecule pharmaceuticals. *Journal of Pharmaceutical Sciences*, 108(5):1655–1662, 2019.

[13] J. Chisholm and S. Motherwell. Compack: a program for identifying crystal structure similarity using distances. *J. Applied Crystal.*, 38:228–231, 2005.

[14] Scott Cohen and Leonidas Guibas. The earth mover's distance: Lower bounds and invariance under translation. Technical report, Stanford University, 1997.

[15] J Conway and N Sloane. Low-dimensional lattices. vi. voronoi reduction of three-dimensional lattices. *Proceedings Royal Society A*, 436(1896):55–68, 1992.

[16] William IF David, Kenneth Shankland, Ch Baerlocher, LB McCusker, et al. *Structure determination from powder diffraction data*, volume 13. 2002.

[17] de Pablo et L. New frontiers for the materials genome initiative. *npj Computational Materials*, 5(1):1–23, 2019.

[18] N Dolbilin and M Bouniaev. Regular t-bonded systems in R^3. *European Journal of Combinatorics*, 80:89–101, 2019.

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

[20] N Dym and S Kovalsky. Linearly converging quasi branch and bound algorithms for global rigid registration. In *Proceedings ICCV*, pages 1628–1636, 2019.

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

[22] Herbert Edelsbrunner and Raimund Seidel. Voronoi diagrams and arrangements. *Discrete & Computational Geometry*, 1(1):25–44, 1986.

[23] A Pulido et al. Functional materials discovery using energy–structure maps. *Nature*, 543:657–664, 2017.

[24] C Groom et al. The cambridge structural database. *Acta Cryst B*, 72(2):171–179, 2016.

[25] D Widdowson et al. Average minimum distances of periodic point sets. *MATCH Communications in Math. Comp. Chemistry*, 87:529–559, 2022.

[26] H Maron et al. Point registration via efficient convex relaxation. *ACM Trans. on Graphics*, 35(4):1–12, 2016.

[27] H Pottmann et al. Integral invariants for robust geometry processing. *Comp. Aided Geom. Design*, 26(1):37–60, 2009.

[28] Hans-Georg Carstens et al. Geometrical bijections in discrete lattices. *Combinatorics, Probability and Computing*, 8:109–129, 1999.

[29] J Yang et al. Go-icp: A globally optimal solution to 3d icp point-set registration. *Transactions PAMI*, 38:2241–2254, 2015.

[30] R Sato et al. Fast and robust comparison of probability measures in heterogeneous spaces. *arXiv:2002.01615*, 2020.

[31] S Manay et al. Integral invariants for shape matching. *Transactions PAMI*, 28(10):1602–1618, 2006.

[32] S Pozdnyakov et al. Incompleteness of atomic structure representations. *Phys. Rev. Lett.*, 125:166001, 2020.

[33] Y Aflalo et al. On convex relaxation of graph isomorphism. *Proceedings of the National Academy of Sciences*, 112(10):2942–2947, 2015.

[34] Giuseppe Fadda and Giovanni Zanzotto. On the arithmetic classification of crystal structures. *Acta Cryst. A: Foundations*, 57(5):492–506, 2001.

[35] Cosmin Grigorescu and Nicolai Petkov. Distance sets for shape filters and shape recognition. *IEEE transactions on image processing*, 12(10):1274–1286, 2003.

[36] Grünbaum and Moore. The use of higher-order invariants in the determination of generalized Patterson cyclotomic sets. *Acta Cryst A*, 51:310–323, 1995.

[37] T Hahn, U Shmueli, and J Arthur. *Internat. tables for crystallography*, volume 1. 1983.

[38] Maria Angeles Junquera Pulido, Linjiang Chen, Tomasz Kaczorowski, Daniel Holden, Marc A Little, Samantha Y Chong, Benjamin J Slater, David Mcmahon, Baltasar Bonillo, Chloe J Stackhouse, et al. Additional computational data (related to" functional materials discovery using energy–structure–function maps" manuscript). 2017.

[39] Heuna Kim and Günter Rote. Congruence testing of point sets in 4 dimensions. *arXiv:1603.07269*, 2016.

[40] V Kurlin. Mathematics of 2-dimensional lattices, 2022.

[41] Vitaliy Kurlin. Density functions of periodic sequences. *arxiv:2205.02226*, 2022.

[42] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426*, 2018.

[43] Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of Comp. Mathematics*, 11(4):417–487, 2011.

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

[45] Robert Osada, Thomas Funkhouser, Bernard Chazelle, and David Dobkin. Shape distributions. *ACM Transactions on Graphics (TOG)*, 21(4):807–832, 2002.

[46] Behnam Parsaeifard and Stefan Goedecker. 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.

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

[48] AL Patterson. Homometric structures. *Nature*, 143:939–940, 1939.

[49] Sergey N Pozdnyakov and Michele Ceriotti. Incompleteness of graph convolutional neural networks for points clouds in three dimensions. *arXiv:2201.07136*, 2022.

[50] Daniel Probst and Jean-Louis Reymond. Visualization of very large high-dimensional data sets as minimum spanning trees. *Cheminformatics*, 12:1–13, 2020.

[51] Joseph Rosenblatt and Paul D Seymour. The structure of homometric sets. *SIAM Journal on Algebraic Discrete Methods*, 3(3):343–350, 1982.

[52] 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.

[53] Mauro R Ruggeri and Dietmar Saupe. Isometry-invariant matching of point set surfaces. In *3DOR*, pages 17–24. Citeseer, 2008.

[54] Pietro Sacchi, Matteo Lusi, Aurora J Cruz-Cabeza, Elisa Nauha, and Joel Bernstein. Same or different–that is the question: identification of crystal forms from crystal structure data. *CrystEngComm*, 22(43):7170–7185, 2020.

[55] N Shervashidze. Weisfeiler-lehman graph kernels. *J. Machine Learning Research*, 12(9), 2011.

[56] S Shirdhonkar and D Jacobs. Approximate earth mover's distance in linear time. In *Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[57] ALJ Spek. Single-crystal structure validation with the program platon. *Journal of applied crystallography*, 36(1):7–13, 2003.

[58] Maxwell W Terban and Simon JL Billinge. Structural analysis of molecular materials using the pair distribution function. *Chemical Reviews*, 122:1208–1272, 2022.

[59] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[60] Peter Zwart, Ralf Grosse-Kunstleve, Andrey Lebedev, Garib Murshudov, and Paul Adams. Surprises and pitfalls arising from (pseudo) symmetry. *Acta Cryst. D*, 64:99–107, 2008.