Navigation maps of the material space for automated self-driving labs of the future

Daniel Widdowson[1,2] and Vitaliy Kurlin[1,2*]

[1]Computer Science Department, University of Liverpool, Ashton Street, Liverpool, L69 3BX, United Kingdom.
[2]Materials Innovation Factory, University of Liverpool, Oxford Street, Liverpool, L7 3NY, United Kingdom.

*Corresponding author(s). E-mail(s): vitaliy.kurlin@liverpool.ac.uk;

Abstract

With the advent of self-driving labs promising to synthesize large numbers of new materials, new automated tools are required for checking potential duplicates in existing structural databases before a material can be claimed as novel. To avoid duplication, we rigorously define the novelty metric of any periodic material as the smallest distance to its nearest neighbor among already known materials.

Using ultra-fast structural invariants, all such nearest neighbors can be found within seconds on a typical computer even if a given crystal is disguised by changing a unit cell, perturbing atoms, or replacing chemical elements. This real-time novelty check is demonstrated by finding near-duplicates of the 43 materials produced by Berkeley's A-lab in the world's largest collections of inorganic structures, the Inorganic Crystal Structure Database and Materials Project.

To help future self-driving labs successfully identify novel materials, we propose navigation maps of the materials space where any new structure can be quickly located by its invariant descriptors similar to a geographic location on Earth.

Keywords: materials space, crystal structure, isometry invariant, continuous metric

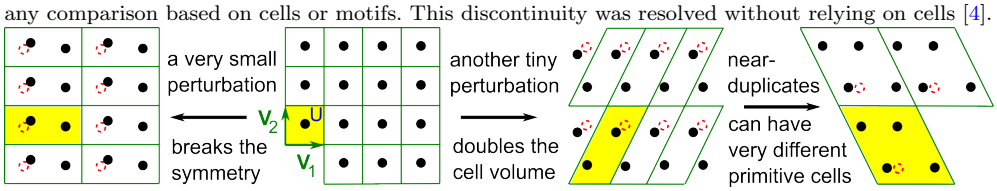
1 Introduction: how is the materials space defined?

The chemical space of all possible molecules is often estimated at the scale of 10^{60} [1]. Similar numbers are quoted for potential materials, though different polymorphs such as diamond and graphite have the same chemical composition and hence can only be distinguished by their geometry. When materials are claimed to be novel amongst

already known ones, we need to rigorously define what constitutes two materials being the "same or different" [2]. The definition of a *crystal structure* was finalised in the periodic case in [3], so we focus on ideal periodic crystals (briefly, *crystals*) as formalised below. When a material is disordered, we consider its closest periodic analogue.

A *crystal* is usually given by a basis of vectors $\boldsymbol{v_1}, \boldsymbol{v_2}, \boldsymbol{v_3}$ in Euclidean space $\mathbb{R}^3$ and a *motif* of atoms with chemical elements and fractional coordinates in this basis. If we forget about chemical elements, the atomic centres $p_1, \dots, p_m$ can be considered zero-sized points in the *primitive unit cell* $U = \{t_1\boldsymbol{v_1} + t_2\boldsymbol{v_2} + t_3\boldsymbol{v_3} \mid t_1, t_2, t_3 \in [0, 1)\}$ defined by the basis $\boldsymbol{v_1}, \boldsymbol{v_2}, \boldsymbol{v_3}$. In dimension 2, the second picture of Fig. 1 highlights the square cell U with the orthonormal basis $\boldsymbol{v_1}, \boldsymbol{v_2}$. Then the underlying *periodic point set* of any crystal consists of infinitely many points $p_i + c_1\boldsymbol{v_1} + c_2\boldsymbol{v_2} + c_3\boldsymbol{v_3}$ for $i = 1, \dots, m$ and integer coefficients $c_1, c_2, c_3 \in \mathbb{Z}$. Infinitely many different pairs of a basis (or a primitive cell) and a motif M generate pointwise identical crystals, see a detailed discussion of this ambiguity of the traditional definition in [3, section 2].

Fig. 1 Almost any tiny perturbation discontinuously scales up a primitive cell and makes unreliable any comparison based on cells or motifs. This discontinuity was resolved without relying on cells [4].



Because atoms vibrate [5, chapter 1], their fractional coordinates are always uncertain and will slightly deviate under repeated measurements even on the same instrument. Almost any displacement of one atom breaks the symmetry and can arbitrarily scale up a primitive unit cell as in Fig. 1. This discontinuity of a reduced cell [6] was experimentally reported in 1965 [7, p. 80] and remained unresolved until 2022 [4] when all periodic crystals in the Cambridge Structural Database (CSD) [8] were distinguished within two days (now within an hour) on a modest desktop computer. Several unexpected duplicates with identical geometries (almost to the last decimal place in all cell parameters and atomic coordinates) but with different chemistry are under investigation by five journals for data integrity [9, section 6].

Because crystal structures are determined in a rigid form, there is no sense in distinguishing crystal representations related by a *rigid motion* (a composition of translations and rotations in $\mathbb{R}^3$), which change a basis and atomic coordinates. On the other hand, there is no sense to fix any threshold $\varepsilon > 0$ that would allow us to call crystals the "same" if all their atomic centres (without chemical attributes) can be matched up to ε-perturbations. Indeed, any periodic point sets can be connected by sufficiently many ε-perturbations [9, Proposition 2.10], which makes the classification based on any threshold $\varepsilon > 0$ trivial due to the transitivity axiom saying that if S is equivalent to Q, and Q is equivalent to T, then S is equivalent to T [3, section 1].

Hence a rigorous way to classify crystals under rigid motion, is to define the *crystal structure* as a rigid class of periodic point sets, see [3, Definition 6]. Then any deviations of atomic positions are not ignored but continuously quantified by a distance metric between different rigid classes. This definition would remain impractical unless we can efficiently separate rigid classes by quickly computable *invariants* that are numerical properties preserved under rigid motion. The chemical composition written as percentages of chemical elements is such an invariant but is *incomplete* because many polymorphs have the same composition but can not be matched by rigid motion.

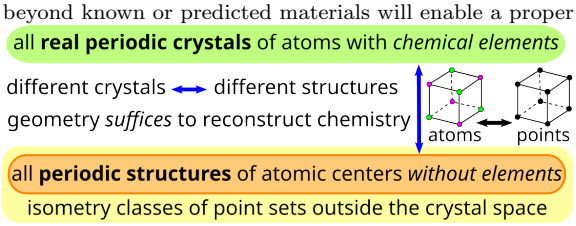
In the sequel, we will consider the sightly weaker equivalence of *isometry* (any distance-preserving transformation in $\mathbb{R}^3$), which is a composition of rigid motion and reflections. Because mirror images can be distinguished by a suitable sign of orientation, so the main difficulty is to classify periodic point sets under isometry.

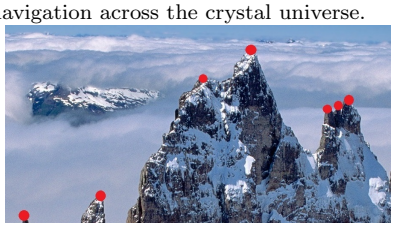
When comparing crystals as periodic sets of atomic centres without chemical attributes, it might seem that all chemistry is lost. However, the fact that all (more than 850 thousand) periodic crystals in the CSD (apart from the investigated duplicates) can be distinguished by isometry invariants in section 2 implies that no information is lost so that all chemistry under standard conditions such as temperature and pressure is in principle reconstructable from sufficiently precise atomic geometry.

This Crystal Isometry Principle (CRISP) first appeared in 2022 [9, section 7] and was inspired by Richard Feynman's Fig.1-7 in [5, chapter 1], which distinguished 7 cubic crystals by their cube size in the first lecture "Atoms and motion", see Fig. 2 (left). More importantly, when we consider atoms only as zero-sized points, we can study all periodic structures in a common continuous space below.

Definition 1 (space of periodic materials)**.** The *Crystal Isometry Space* CRIS($\mathbb{R}^3$) is the space of isometry classes of all periodic sets of points without atomic attributes.

Fig. 2 **Left**: the *Crystal Isometry Principle* says that all chemistry of any real periodic crystal under standard ambient conditions can be reconstructed from (the isometry class of) the periodic set of atomic centers given with precisely enough coordinates [4]. **Right**: most optimization methods output local optima without exploring the space around. De-fogging this *Crystal Isometry Space* CRIS($\mathbb{R}^3$) beyond known or predicted materials will enable a proper navigation across the crystal universe.





Because (the isometry classes of) any periodic set of points has a unique location in CRIS($\mathbb{R}^3$), all known materials can be considered 'visible stars' in this continuous universe, while any periodic crystal discovered in the future will appear at its own position like a 'new star', see Fig. 2 (right). Not every position in CRIS($\mathbb{R}^3$) is realizable by a material because inter-atomic distances cannot be arbitrary in the same way as

not every location on Earth is habitable. However, mapping the whole space CRIS($\mathbb{R}^3$) by invariant coordinates enables a proper exploration with a geographic-style map.

If we do not restrict the motif size, the space CRIS is infinite dimensional. However, if we consider all periodic sets with exactly m points in a motif, the resulting subspace CRIS($\mathbb{R}^3; m$) has dimension $3m+3$ due to m triples x, y, z of atomic coordinates and 6 parameters of a unit cell, of which 3 are neutralised by translations along basis vectors. Alternatively, we can define a unit cell by 3 basis vectors with 3 coordinates, of which 6 are neutralised by 3+3 parameters of translations and rotations in $\mathbb{R}^3$.

In the partial case $m = 1$, CRIS($\mathbb{R}^3; 1$) is a continuous 6-dimensional space of 3D lattices, which was previously cut into 14 disjoint subspaces of Bravais classes [10] but is now parametrized by complete invariants [11]. Continuous maps of the simpler 3-dimensional space CRIS($\mathbb{R}^2; 1$) of 2D lattices recently appeared in [12], [13], [14].

The full space CRIS($\mathbb{R}^3$) $= \cup_{m=1}^{+\infty}$CRIS($\mathbb{R}^3; m$) is a union of infinitely many subspaces fr $m = 1, 2, 3, \dots$ such that any periodic set with m points in a cell is infinitesimally close to infinitely many subspaces of sets with $2m, 3m, \dots$ points in a primitive cell. Indeed, perturbations in Fig. 1 arbitrarily extend any given cell and make the extended cell primitive by a tiny displacement of any atom and all its translational copies. Crystals should be continuously compared only across multiple subspaces, not within one subspace CRIS($\mathbb{R}^3; m$) for a fixed number m of atoms.

Any database of periodic crystals is a finite sample from the continuous space CRIS($\mathbb{R}^3$). The first contribution is continuous maps of the world's five largest databases on CRIS($\mathbb{R}^3$) projected to various structural invariants. The second contribution is the local novelty distance based on generically complete invariants whose utility is demonstrated by identifying closest neighbors of the 43 A-lab crystals in the Inorganic Crystal Structure Database (ICSD) [15] and Materials Project (MP) [16].

2 Methods: invariant-based novelty distance metric

This section introduces a new metric LND (Local Novelty Distance) that satisfies all metric axioms and continuously quantifies in real time a deviation of any newly synthesized crystal from its nearest neighbor in an existing structural database.

2.1 Generically complete and continuous structural invariants

Definition 2 reminds us of the Pointwise Distance Distribution (PDD), which suffices together with a lattice to reconstruct any generic periodic point set $S \subset \mathbb{R}^3$ up to isometry by [4, Theorem 4.4]. Generic means any set apart from a singular subspace of measure 0, e.g. almost any tiny perturbation of atoms makes every crystal generic.

The PDD is a matrix of inter-point distances and is stronger than the Pair Distribution Function (PDF) [17] in the sense that PDD can be simplified to PDF but distinguishes homometric structures [18] that have the same PDF [4, section 3].

Definition 2 (isometry invariant PDD($S; k$))**.** Let $S \subset \mathbb{R}^n$ be a periodic point set with a motif $M = \{p_1, \dots, p_m\}$. Fix an integer $k \geq 1$. For every point $p_i \in M$, let

$d_1(p) \le \cdots \le d_k(p)$ be the distances from p to its k nearest neighbours within the full infinite set S not restricted to any cell. The matrix $D(S;k)$ has m rows consisting of the distances $d_1(p_i),\ldots,d_k(p_i)$ for $i=1,\ldots,m$. If any $l \ge 1$ rows are identical to each other, we collapse them into a single row and assign the weight l/m to this row. The resulting matrix of maximum m rows and $k+1$ columns including the extra (say, 0-th) column of weights is called the *Pointwise Distance Distribution* $\mathrm{PDD}(S;k)$.

In Definition 2, any point $p_i \in M$ can have several different neighbours at the same distance but the k smallest distances (without any indices or types of neighbours) are always well-defined. The matrix $\mathrm{PDD}(S;k)$ has ordered columns according to the index of neighbours but unordered rows because points of a motif of S are considered unordered. We can add chemical elements or atomic weights as extra attributes to the rows of $\mathrm{PDD}(S;k)$ but the pure geometric information will suffice in practice.

We can compare PDD matrices that have same number of columns and possibly different numbers of rows by interpreting $\mathrm{PDD}(S;k)$ as a distribution of unordered rows (or points in $\mathbb{R}^k$) with weights or probabilities. One metric on such weighted distributions is the Earth Mover's Distance (EMD), which was previously used in [19] for chemical compositions from the ICSD. If any point is perturbed up to ε in Euclidean distance, any inter-point distance changes up to 2ε.

This upper bound of 2ε formally follows from the triangle axiom of a distance metric, which is essential needed for reliable clustering. If the triangle axiom fails with any positive error, outputs of widely used clustering algorithms such as k-means and DBSCAN may not be trustworthy [20]. The EMD is a proper metric on weighted distributions (hence on PDD matrices) satisfying all metric axioms [21, Appendix].

If the number k of neighbours increases to infinity, the asymptotic behaviour of distances to neighbours is described in terms of the Point Packing Coefficient below.

Definition 3 (Point Packing Coefficient PPC)**.** Let $S \subset \mathbb{R}^3$ be a periodic point set with m atoms in a unit cell U. The *Point Packing Coefficient* is $\mathrm{PPC}(S) = \sqrt[3]{\dfrac{\mathrm{vol}(U)}{mV_3}}$, where $\mathrm{vol}(U)$ is the volume of U, $V_3 = \dfrac{4}{3}\pi$ is the volume of the unit ball in $\mathbb{R}^3$.

The distances in each row of $\mathrm{PDD}(S;k)$ asymptotically increase as $\mathrm{PPC}(S)\sqrt[3]{k}$ by [9, Theorem 13]. This asymptotic behaviour motivates the simplified invariants below.

Definition 4 (invariants AMD, ADA, PDA)**.** The *Average Minimum Distance* $\mathrm{AMD}_k(S)$ is the weighted average of the k-th column of $\mathrm{PDD}(S;k)$. The *Average Deviation from Asymptotic* is $\mathrm{ADA}_k(S) = \mathrm{AMD}_k(S) - \mathrm{PPC}(S)\sqrt[3]{k}$ for $k \ge 1$. The *Pointwise Deviation from Asymptotic* is the matrix $\mathrm{PDA}(S;k)$ obtained from $\mathrm{PDD}(S;k)$ by subtracting $\mathrm{PPC}(S)\sqrt[3]{k}$ from any distance in row i and column k for $i,k \ge 1$.

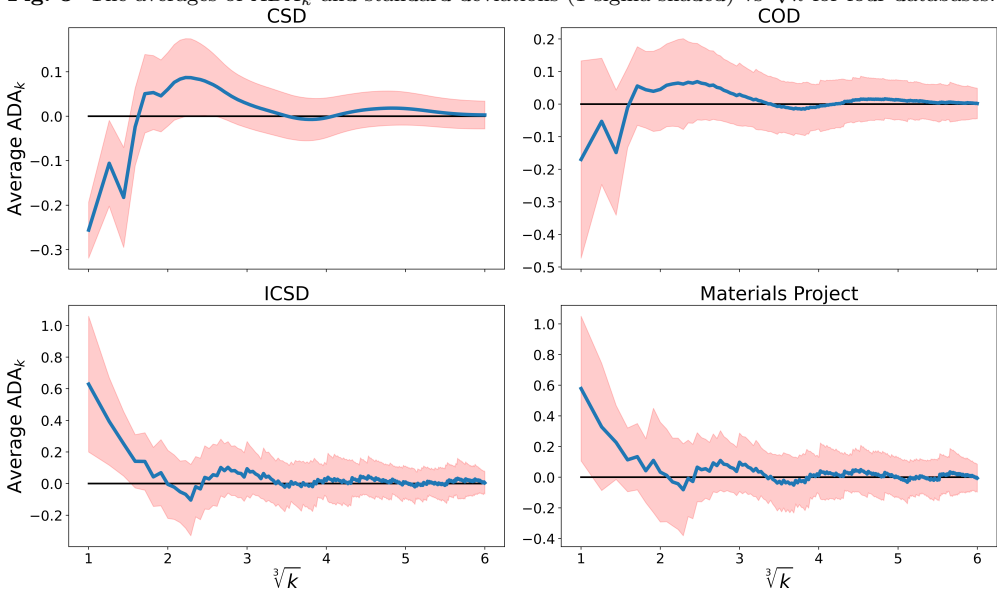
The invariants AMD_k and ADA_k form vectors of length k, e.g. set $\mathrm{AMD}(S;k) = (\mathrm{AMD}_1(S),\ldots,\mathrm{AMD}_k(S))$ and $\mathrm{ADA}(S;k) = (\mathrm{ADA}_1(S),\ldots,\mathrm{ADA}_k(S))$. These vectors can be compared by many metrics. The metric $L_\infty(u,v) = \max\limits_{i=1,\ldots,k} |\boldsymbol{u}_i - \boldsymbol{v}_i|$ for any vectors $\boldsymbol{u},\boldsymbol{v} \in \mathbb{R}^k$ preserves the intuition of atomic displacements in the following sense. If S is obtained from Q by perturbing every point up to a small ε,

then $L_\infty(\mathrm{AMD}(S;k),\mathrm{AMD}(Q;k)) \le 2\varepsilon$ by [9, Theorem 9]. Other distances such as Euclidean can be considered but will accumulate a larger deviation depending on k.

All invariants above and metrics on them are measured in the same units as original coordinates, i.e. in Angstroms for crystals given by Crystallographic Information Files (CIFs). The Point Packing Coefficient $\mathrm{PPC}(S)$ was defined as the cube root of the cell volume per atom (of the same radius 1Å) and can be interpreted as an average radius of balls 'packed' in a unit cell. So $\mathrm{PPC}(S)$ is roughly inverse proportional to the physical density but they are exactly related only when materials have the same average atomic mass (total mass of atoms in a unit cell divided by the cell volume).

While $\mathrm{AMD}_k(S)$ monotonically increases in k, the invariants $\mathrm{ADA}_k(S)$ can be positive or negative as deviations around the asymptotic $\mathrm{PPC}(S)\sqrt[3]{k}$. Fig. 3 reveals geometric differences between the mainly organic databases CSD and Crystallography Open Database (COD) [22] versus the more inorganic collections ICSD and MP.

Fig. 3 The averages of ADA_k and standard deviations (1 sigma shaded) vs $\sqrt[3]{k}$ for four databases.



The first average of $\mathrm{ADA}_1 \in [-0.25, -0.17]$ in the top images of Fig. 3 can be explained by the presence of many hydrogen atoms, which have distances smaller than $\mathrm{PPC}(S)$ to their first neighbor in most organic materials. Indeed, hydrogens are usually bonded at distances less than 1.2Å, while $\mathrm{PPC}(S)$ is often larger than 1.2Å because most chemical elements have van der Waals radii above 1.2Å [23].

For inorganic materials, metal atoms or ions have relatively large distances to their first neighbors, so the average ADA_1 is in $[0.58, 0.62]$ in the bottom images of Fig. 3.

For all types of materials in Fig. 3, the value of ADA_k experimentally converges to 0 on average meaning that there is no need to substantially increase k because the important structural information emerges for smaller indices k of neighbors.

If we increase k, the matrix $\mathrm{PDD}(S;k)$ and hence the vector $\mathrm{ADA}(S;k)$ become longer by including distance data to further neighbors but all initial values remain the same. Hence we consider k not as a parameter that changes the output but as a degree of approximation similarly to the number of decimal places on a calculator.

The convergence $\mathrm{ADA}_k \to 0$ as $k \to +\infty$ justifies computing the distance L_∞ between ADA vectors up to a reasonable k. In practice, we use $k = 100$ because all ADA_k for $k \geq 100$ are close to 0 (the range of 1 sigma between ± 0.2Å) in Fig. 3.

2.2 Novelty distance based on practically complete invariants

This subsection introduces the Local Novelty Distance $\mathrm{LND}(S;D)$ of a periodic crystal S as a distance to the closest neighbor Q of S in a given dataset D.

The new distance LND is measured as a metric between the invariants $\mathrm{PDA}(S;k)$ and $\mathrm{PDA}(Q;k)$ for $k = 100$, motivated as follows. First, the invariants $\mathrm{PDD}(S;100)$ distinguished all non-duplicate periodic crystals in the CSD. Second, for a generic periodic set S (away from a measure 0 subspace), $\mathrm{PDD}(S;k)$ with a big enough k and a lattice of S suffices to reconstruct S uniquely under isometry in $\mathbb{R}^n$ by [4, Theorem 4.4]. Finally, distances to k-th neighbors in $\mathrm{PDD}(S;k)$ asymptotically increase as $\mathrm{PPC}(S)\sqrt[3]{k}$. If crystals S, Q have different $\mathrm{PPC}(S) \neq \mathrm{PPC}(Q)$, the distance L_∞ between corresponding rows of PDD matrices likely equals the expected largest difference in the final k-th colimn, which ignores all neighbors with smaller indices. Hence subtracting $\mathrm{PPC}(S)\sqrt[3]{k}$ in Definition 4 makes any metric on PDAs more informative than on PDDs. Definition 5 introduces a metric on PDA matrices.

Definition 5 (Earth Mover's Distance EMD [21])**.** Consider any matrix $\mathrm{PDA}(S;k)$ as a distribution of rows $R_i(S)$ with weights $w_i(S)$ for $i = 1, \dots, m(S)$ such that $\sum\limits_{i=1}^{m} w_i = 1$. The *Earth Mover's Distance* $\mathrm{EMD}(\mathrm{PDA}(S;k), \mathrm{PDA}(Q;k)) = \min\limits_{f_{ij}} \sum\limits_{i=1}^{m(S)} \sum\limits_{j=1}^{m(Q)} f_{ij} L_\infty(R_i(S), R_j(Q))$ is minimized for all real $f_{ij} \geq 0$ (called *flows*) subject to the conditions $\sum\limits_{i=1}^{m(S)} f_{ij} \leq w_j(Q)$, $\sum\limits_{j=1}^{m(S)} f_{ij} \leq w_i(S)$, $\sum\limits_{i=1}^{m(S)} \sum\limits_{j=1}^{m(Q)} f_{ij} = 1$.

The first condition $\sum\limits_{j=1}^{m(Q)} f_{ij} \leq w_i(S)$ means that not more than the weight $w_i(S)$ of the component $R_i(S)$ 'flows' into all components $R_j(Q)$ via 'flows' f_{ij} for $j = 1, \dots, m(Q)$. The second condition $\sum\limits_{i=1}^{m(S)} f_{ij} = w_j(Q)$ means that all 'flows' f_{ij} from $R_i(S)$ for $i = 1, \dots, m(S)$ 'flow' into $R_j(Q)$ up to the maximum weight $w_j(Q)$. The last condition $\sum\limits_{i=1}^{m(S)} \sum\limits_{j=1}^{m(Q)} f_{ij} = 1$ forces to 'flow' all rows $R_i(S)$ to all rows $R_j(Q)$.

Definition 6 (Local Novelty Distance $\mathrm{LND}(S;D)$)**.** Let D be a finite dataset of periodic point sets. Fix an integer $k \geq 1$. For any periodic point set S, the *Local Novelty Distance* $\mathrm{LND}(S;D) = \min\limits_{Q \in D} \mathrm{EMD}(\mathrm{PDA}(S;k), \mathrm{PDA}(Q;k))$vis the shortest L_∞ distance from S to its nearest neighbor Q in the given dataset D.

If S is already contained in the dataset D, then $\mathrm{LND}(S;D) = 0$, so S cannot be considered novel. More practically, a newly synthesized periodic crystal S can be a near-duplicate of some known Q. Then $\mathrm{LND}(S;D)$ is small as justified below.

The *packing radius* $r(Q)$ is the minimum half-distance between any points of Q.

Theorem 7. If S is obtained from a crystal Q in a dataset D by perturbing every point of Q up to $\varepsilon < r(Q)$, then $\mathrm{LND}(S;D) \leq 2\varepsilon$. To get S from a crystal $Q \in D$ with $\mathrm{LND}(S;D) < 2r(Q)$, some atom of Q should be perturbed by at least $0.5\mathrm{LND}(S;D)$.

Theorem 7 is proved in Appendix A. The distance $\mathrm{LND}(S;D)$ is called *local* because Definition 6 uses the first nearest neighbor of S in D. Another novelty of S can be characterized with respect to a global distribution of all crystals in D, which we will explore in a forthcoming work. The local novelty is more urgently needed to tackle the growing crisis of duplication in experimental and simulated databases, some of which were publicly rebutted in [24], [25], and [26], [3, Tables 1-2 in section 6], respectively.

2.3 Insufficiency of past invariants and similarities of crystals

This subsection only briefly reviews the past approaches to classify crystals and quantify their similarities. Some widely used similarities such as the Root Mean Square Deviation (RMSD) [27] deserve their own detailed discussions in another forthcoming work. The shape (isometry class) of a reduced cell with standard settings [28] were thoroughly developed to uniquely represent any periodic crystal. The resulting conventional representation can be theoretically considered a complete isometry invariant but discontinuously changes under almost any perturbation in practice.

Indeed, perturbations in Fig. 1 apply to any crystal and can arbitrarily extend a reduced cell to a larger cell whose size cannot be reduced. Searching for a small perturbation (pseudo-symmetry) to make a cell smaller [29] inevitably uses thresholds and leads to a trivial classification due to the transitivity axiom, see [3, section 1].

The COMPACK algorithm [27] outputs an RMSD quantity by comparing finite portions of only molecular crystals. Its implementation in Mercury also uses thresholds for acceptable deviations of atoms and angles. Even if these thresholds are ignored (made large), the algorithm chooses one molecule in a unit cell and 14 (by default) closest molecules around it. The resulting molecular group depends on a central molecule; for co-crystals containing geometrically different molecules or the same molecule in non-equivalent positions, matching these molecules by rigid motion does not match the full crystal. Even for simple crystals based on a single molecule as is often the case in Crystal Structure Prediction [30], the choice of 14 (or any other number of) neighbours can be discontinuous when a central molecule has 14th and 15th neighbours at the same distance. Selected clusters of molecules in two crystals require an optimal

alignment, which is a hard problem because atomic sets can contain numerous indistinguishable atoms, so the optimization must consider many potential permutations. This problem of exponentially many permutations was recently resolved in [31] but a choice of a single atomic environment in a crystal remains discontinuous.

Other similarities based on all atomic environments such as SOAP [32] and MACE [33] use a Gaussian deviation and a cut-off radius for interatomic interactions to convert a periodic set of discrete points to a complicated smooth function. This function decomposes into an infinite sum of spherical harmonics whose truncation up to a certain order becomes incomplete, which will be discussed in future work.

The PXRD similarity compares crystals through powder diffraction patterns that are identical for homometric structures [18], some of which were distinguished even by AMD_2 in [9, appendix A]. The PXRD as implemented in Mercury also fails the triangle inequality but runs faster than the RMSD and SOAP similarities.

In summary, the past approaches through conventional representations and environment-based similarities separately focused on two important complementary properties: completeness and continuity. The problem of combining these two properties was first stated in [34] for lattices and then extended in [35] to a complete invariant isoset of any periodic point set and a continuous metric approximated with a small error factor by an algorithm whose time polynomially depends on the motif size [36].

3 Results: novelty of materials and navigation maps

This section describes how the 43 materials reported by A-lab can be automatically positioned relative to the ICSD and MP within the full materials space $\mathrm{CRIS}(\mathbb{R}^3)$.

Among the 43 materials whose CIFs are available in the supplementary materials in [37], only 32 are pure periodic without any disorder, 10 have *substitutional* disorder with one or more sites occupied by multiple atomic types, and one has *positional* disorder with an atom occupying any of 4 positions with occupancy 0.5.

Closest neighbors within the ICSD and Materials Project for each A-lab crystal were found as follows. Using binary search on $\mathrm{ADA}(S;100)$ vectors with the metric L_∞, we found the nearest 100 neighbors for each A-lab crystal within each database. These neighbors were then re-compared by Earth Mover's Distance on the stronger invariants $\mathrm{PDD}(S;100)$. This EMD metric also outputs which atomic types and/or occupancies were correctly matched and which were not. Since most A-lab crystals had several geometric nearest neighbors with small distances EMD, we selected the neighbor with the most similar composition as measured by element mover's distance [19], which are listed in Tables 2 and 4 below. The local novelty distance of each A-lab crystal is not more than the Earth Mover's Distance listed in the column $\mathrm{EMD},100$.

The time taken in each step of the process described above is given in Table 1 below. All experiments were performed on an average desktop computer: AMD Ryzen 5 5600X (6-core), 32GB RAM, Python 3.9.

Stage	ICSD, seconds	MP, seconds
Binary search on ADA(S; 100) in the full database	3.023	2.450
PDA(Q; 100) for 100 neighbors found by ADA	5.272	5.990
EMD on PDAs for 100 neighbors found by ADA	0.535	0.742
Elemental Mover's Distance (ElMD) for 100 neighbors	9.534	9.737

Table 1 Times (seconds) to complete each stage of the process of finding nearest neighbors in the ICSD and Materials Project for each A-lab crystal on a modest desktop computer.

3.1 Local novelty distances of the A-lab materials vs ICSD

Table 2 lists the nearest neighbors found in the ICSD for each A-lab crystal. We used a snapshot of the ICSD in 2019, while the GNoME AI project used a snapshot from 2021. Despite this, two A-lab crystals were found to already exist in the ICSD: $KNaP_6(PbO_3)_8$ has the closest neighbor ICSD 182501 reported in 2011 [38], and $MnAgO_2$ has the closest neighbor ICSD 670065 reported as a hypothetical structure in 2015 [39]. In particular, $MnAgO_2$ was one of three crystals that the later rebuttal said was synthesized successfully [24]. They state that the material was first reported in 2021 [40] (ICSD 139006), after the snapshot used to train the GNoME AI, and so was not included in the original training data and could be considered a success. Our findings show this crystal did in fact exist in the ICSD prior to the 2021 snapshot.

The pre-existing version of this crystal was not found by [24] using a unit cell search because the unit cell of ICSD 670065 significantly differs from that of the A-lab version or ICSD 139006, with the former listing its space group as A 2/m and the latter two having space group C 2/m. This missed near-duplicate further supports the robustness of a search based on continuous invariants independent of a unit cell, which can find near-duplicates despite disagreement on the correct space group.

Aside from the two structures above, all other A-lab crystals were found to have a geometric near-duplicate in the ICSD with a different composition. Many of these near-duplicates involve the substitution of only one atom, replacing a disordered site with a fully ordered one or adjusting the occupancy ratios of atoms at a site.

These structural analogues of A-lab's reported materials are not surprising as the GNoME AI [41] used atomic substitution on existing crystals to generate potential new ones without substantially changing the atomic geometry.

The fact that pre-existing structures in the ICSD were missed by the later rebuttal [24] suggests that a more robust method is needed for comparing structures in the aid of materials discovery.

3.2 Local novelty distances of the A-lab materials vs MP

The Materials Project contains a substantial number of theoretical structures, many of which are obtained by substituting atoms in existing structures with plausible alternatives, a strategy also employed by the GNoME AI which generated the crystals later synthesized by Berkeley's A-lab. Despite the substitution patterns used by GNoME being tuned to prioritize discovery and not repeat data, 42 of the 43 A-lab crystals were found to already exist in the Materials Project, all of which predate the March 2021 snapshot used to train the GNoME AI and hence were part of its training data.

A-lab name	ICSD ID	ICSD composition	EMD, 100	Site mismatches
Ba_2ZrSnO_6*	181433	$In_{0.5}Nb_{0.5}BaO_3$	0.003	$Zr_{0.5}Sn_{0.5} \leftrightarrow Nb_{0.5}In_{0.5}$
$Ba_6Na_2Ta_2V_2O_{17}$	97524	$Ba_6Na_2Ru_2V_2O_{17}$	0.092	Ta ↔ Ru
$Ba_6Na_2V_2Sb_2O_{17}$	97524	$Ba_6Na_2Ru_2V_2O_{17}$	0.081	Sb ↔ Ru
$Ba_9Ca_3La_4(Fe_4O_{15})_2$*	72336	$Ca_3La_4Fe_8Ba_9O_{30}$	0.192	$(Ca_{0.43}La_{0.57})_2Ba \leftrightarrow Ca_{0.33}La_{0.67}(Ca_{0.5}Ba_{0.5})_2$
$CaCo(PO_3)_4$	412558	MnP_2O_6	0.173	$CoCa \leftrightarrow Mn_2$
$CaFe_2P_2O_9$	407045	$CaVNiP_2O_9$	0.157	$Fe_2 \leftrightarrow VNi$
$CaGd_2Zr(GaO_3)_4$*	202850	$Ca_{0.95}Zr_{0.95}Gd_{2.05}Ga_{4.05}O_{12}$	0.123	$GaZrGdCa \leftrightarrow Ga_{0.52}Zr_{0.48}Ca_{0.32}Gd_{0.68}$
$CaMn(PO_3)_4$	412558	MnP_2O_6	0.132	Ca ↔ Mn
$CaNi(PO_3)_4$	37136	$NiCoP_4O_{12}$	0.204	Ca ↔ Co
$FeSb_3Pb_4O_{13}$*	65839	$CrSb_3Pb_{3.93}O_{13}$	0.086	$Fe_{0.25} \leftrightarrow Cr_{0.25}$
$Hf_2Sb_2Pb_4O_{13}$	84759	$W_{4.48}Sn_{11.5}Pb_{15.8}O_{51.9}$	0.086	$SbHf \leftrightarrow Sn_{0.72}W_{0.28}$
$InSb_3(PO_4)_6$	166834	$InSb_3P_6O_{24}$	0.21	$SbIn \leftrightarrow In_{0.5}Sb_{0.5}$
$InSb_3Pb_4O_{13}$	262198	$Bi_2Sn_2O_7$	0.439	$SbOIn_2Pb_2 \leftrightarrow Sn_2Bi_3O$
$K_2TiCr(PO_4)_3$	280999	$CrTiK_2P_3O_{12}$	0.044	$TiCr \leftrightarrow Ti_{0.61}Cr_{0.39}Ti_{0.39}Cr_{0.61}$
$K_4MgFe_3(PO_4)_5$	161484	$MgFe_3K_4P_5O_{20}$	0.075	$FeMg \leftrightarrow Mg_{0.25}Fe_{0.75}$
$K_4TiSn_3(PO_5)_4$	250088	$Ti_{0.253}Sn_{0.747}KPO_5$	0.086	$TiSn_2 \leftrightarrow Ti_{0.26}Sn_{0.74}$ $Ti_{0.24}Sn_{0.76}$
$KBaGdWO_6$	60499	$WCaBa_2O_6$	0.009	$GdK \leftrightarrow CaBa$
$KBaPrWO_6$	60499	$WCaBa_2O_6$	0.053	$PrK \leftrightarrow CaBa$
KMn_3O_6*	261406	$K_{0.463}MnO_2$	0.016	$K_{0.5} \leftrightarrow K_{0.695}$
$KNa_2Ga_3(SiO_4)_3$	411328	$SiNaGaO_4$	0.27	$SiGaK \leftrightarrow GaSiNa$
$KNaP_6(PbO_3)_8$*	182501	$KNaP_6Pb_8O_{24}$	0.005	
$KNaTi_2(PO_5)_2$	67539	$K_{0.05}Na_{0.95}TiPO_5$	0.2	$NaK \leftrightarrow Na_{0.95}K_{0.05}$
$KPr_9(Si_3O_{13})_2$*	153272	$KSi_6Pr_9O_{26}$	0.16	$(K_{0.1}Pr_{0.9})_2 \leftrightarrow PrK_{0.25}Pr_{0.75}$
$Mg_3MnNi_3O_8$	80306	$MnNi_3Mg_3O_8$	0.043	$MgNi \leftrightarrow Mg_{0.5}Ni_{0.5}$
Mg_3NiO_4*	60496	$Cu_{0.2}Mg_{0.8}O$	0.003	$Mg_{0.75}Ni_{0.25} \leftrightarrow Mg_{0.8}Cu_{0.2}$
$MgCuP_2O_7$*	69576	$Co_{0.92}Mg_{1.08}P_2O_7$	0.218	$Mg_{0.5}Cu_{0.5} \leftrightarrow Mg_{0.54}Co_{0.46}$
$MgNi(PO_3)_4$	37136	$NiCoP_4O_{12}$	0.141	$Mg \leftrightarrow Co$
$MgTi_2NiO_6$	171583	$NiMgTi_2O_6$	0.038	$MgNi \leftrightarrow Mg_{0.5}Ni_{0.5}$
$MgTi_4(PO_4)_6$	419418	$MnTi_4P_6O_{24}$	0.133	$Mg \leftrightarrow Mn$
$MgV_4Cu_3O_{14}$	69731	$MgCu_3V_4O_{14}$	0.11	$CuMg \leftrightarrow Mg_{0.25}Cu_{0.75}$
Mn_2VPO_7	20296	$Mn_2P_2O_7$	0.21	$V \leftrightarrow P$
$Mn_4Zn_3(NiO_6)_2$	625	$MgCu_2Mn_3O_8$	0.186	$MnZnNi \leftrightarrow MgCuMn$
$Mn_7(P_2O_7)_4$	67514	$Fe_7P_8O_{28}$	0.126	$Mn \leftrightarrow Fe$
$MnAgO_2$	670065	$MnAgO_2$	0.097	
$Na_3Ca_{18}Fe(PO_4)_{14}$	85103	$FeNa_3P_{14}Ca_{18}O_{56}$	0.153	$FeCa_2Na \leftrightarrow Ca_{0.5}Fe_{0.5}Na_{0.17}Ca_{0.83}$
$Na_7Mg_7Fe_5(PO_4)_{12}$	200238	$Na_2Fe_3P_3O_{12}$	0.229	$POMg_2 \leftrightarrow Na_3Fe$
$NaCaMgFe(SiO_3)_4$*	172120	$NaCaMgCrSi_4O_{12}$	0.075	$(MgFeNaCa)_{0.25} \leftrightarrow MgCrNaCa$
$NaMnFe(PO_4)_2$	200238	$Na_2Fe_3P_3O_{12}$	0.242	$POMn_2 \leftrightarrow Na_2Fe_2$
$Sn_2Sb_2Pb_4O_{13}$	262198	$Bi_2Sn_2O_7$	0.461	$SbOSnPb_2 \leftrightarrow SnBi_3O$
$Y_3In_2Ga_3O_{12}$	185862	$Y_3Ga_5O_{12}$	0.104	$In \leftrightarrow Ga$
$Zn_2Cr_3FeO_8$	196119	$ZnCr_2O_4$	0.022	$Fe \leftrightarrow Cr$
$Zn_3Ni_4(SbO_6)_2$*	180711	$Ti_{0.18}Zr_{0.33}ZnO_2$	0.162	$Ni_{0.66}Sb_{0.33} \leftrightarrow Ti_{0.17}Zn_{0.5}Zr_{0.33}$
$Zr_2Sb_2Pb_4O_{13}$	65054	$TiSbPb_{1.97}O_{6.5}$	0.12	$SbZr \leftrightarrow Ti_{0.5}Sb_{0.5}$

Table 2 Close neighbors of each A-lab crystal in the ICSD. The ICSD entry with the smallest element mover's distance [19] was selected from the list of 100 nearest neighbors by ADA_{100}. Disordered crystals are marked with an asterisk *.

As the Materials Project does not model disorder, no match was found for the positionally disordered KMn_3O_6. However, its nearest neighbor was found in the ICSD (with a change in occupancy), and so all 43 A-lab crystals had already been hypothesized or synthesized prior to the beginning of the GNoME project. Table 3 below lists the 8 crystals, which have already been synthesized.

A-lab name	Matching database entries	Source and date
$Ba_6Na_2Ta_2V_2O_{17}$	mp-1214664, Pauling file sd_1003187	[42], 2003
$Ba_6Na_2V_2Sb_2O_{17}$	mp-1214658, Pauling file sd_1003189	[42], 2003
$CaGd_2Zr(GaO_3)_4$	mp-686296, ICSD 202850	[43], 1988
$KNa_2Ga_3(SiO_4)_3$	mp-1211711, Pauling file sd_1707156	[44], 1982
$KNaP_6(PbO_3)_8$	ICSD 182501	[45], 2011
$KNaTi_2(PO_5)_2$	mp-1211611, Pauling file sd_1414297	[46], 1991
Mn_2VPO_7	mp-1210613, Pauling file sd_1322766	[47], 2000
$Y_3In_2Ga_3O_{12}$	mp-1207946, Pauling file sd_1704376	[48], 1964

Table 3 The 8 reportedly new crystals synthesized by A-lab found to already have been synthesized and uploaded to various databases.

$Y_3In_2Ga_3O_{12}$ in Table 3 was one of the three crystals agreed to have been synthesized by the later rebuttal paper [24], as discussed in Section 3.1. Their earliest reference for this crystal dates to 2022 [49], again leading to their conclusion that the crystal was novel from the perspective of the GNoME AI trained on data from 2021. We found that this crystal was reported in 1964 and uploaded to the Materials Project no later than 2018, and so would have been part of GNoME's training data.

The 10 substitutionally disordered A-lab crystals had matches in the Materials Project where disordered sites were replaced with multiple fully ordered sites of atoms in the same ratio; e.g. $FeSb_3Pb_4O_{13}$ matching mp-1224890 had a site $Fe_{0.25}Sb_{0.75}$ with multiplicity 4 replaced with $FeSb_3$. For completeness, this is noted in the site mismatches column of Table 4, listing all nearest neighbors in the Materials Project.

One pair of note is $CaGd_2Zr(GaO_3)_4$ & mp-686296, which have one atom swapped ($Ga \leftrightarrow Zr$). This Materials Project entry originates from ICSD 202850, listed in Table 2 as the closest neighbor in the ICSD. The ICSD entry has disorder on the sites where atoms were swapped, whereas the A-lab and Materials Project versions have no disorder. We conclude that this crystal is not new, as these atoms could have been swapped to match the A-lab crystal with a different ordering of the disordered ICSD entry.

The GNoME paper used the Pymatgen structure matcher [50] to filter out duplicate structures, whose first three steps are quoted below:

"1. Given two structures: s1 and s2

2. Optional: Reduce to primitive cells.

3. If the numbers of sites do not match, return False."

These steps are followed by several heuristic steps which involve finding deviations between atoms in the reduced unit cell. If step 2 above is optionally missed, step 3 can output False (no match) for identical crystals given with different non-primitive cells. If step 2 is enforced, step 3 will output False (no match) for any nearly identical

crystals, whose primitive cells can arbitrarily differ due to a tiny atomic displacement as in Fig. 1. For the above reasons, our findings show that this method of comparing structures was insufficient for comparing structures to filter out existing duplicates from the data, resulting in the AI silently reproducing data in the training set.

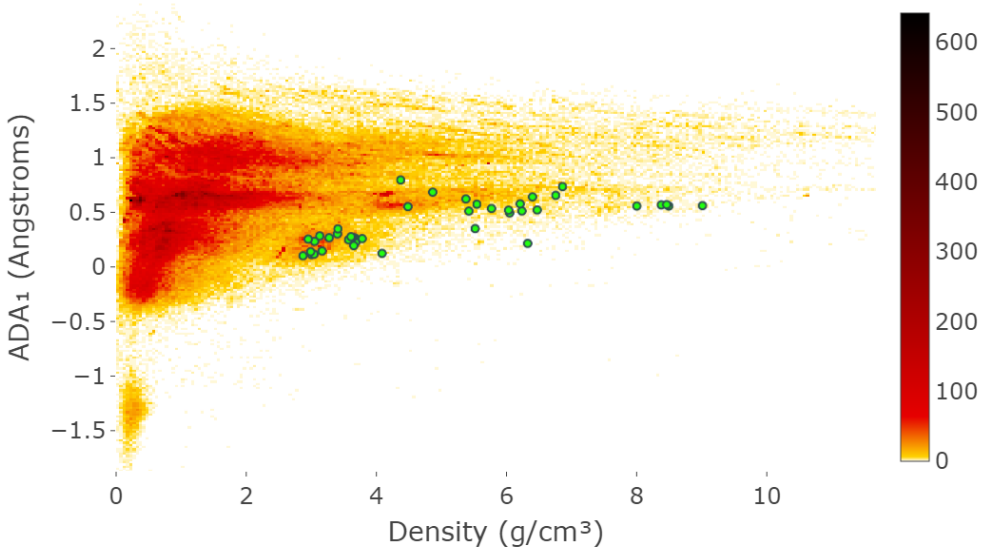
A-lab name	MP ID	MP composition	EMD_{100}	Site mismatches
Ba_2ZrSnO_6*	1228067	Ba_2ZrSnO_6	0.025	$Zr_{0.5}Sn_{0.5}$ ↔ ZrSn
$Ba_6Na_2Ta_2V_2O_{17}$	1214664	$Ba_6Na_2Ta_2V_2O_{17}$	0.029	
$Ba_6Na_2V_2Sb_2O_{17}$	1214658	$Ba_6Na_2V_2Sb_2O_{17}$	0.021	
$Ba_9Ca_3La_4(Fe_4O_{15})_2$*	1228537	$Ba_9Ca_3La_4Fe_8O_{30}$	0.136	$Ca_{0.43}La_{0.57}$ ↔ Ca_3La_4
$CaCo(PO_3)_4$	1045787	$CaCoP_4O_{12}$	0.090	
$CaFe_2P_2O_9$	1040941	$CaFe_2P_2O_9$	0.114	
$CaGd_2Zr(GaO_3)_4$*	686296	$CaGd_2ZrGa_4O_{12}$	0.069	Ga ↔ Zr
$CaMn(PO_3)_4$	1045779	$CaMnP_4O_{12}$	0.163	
$CaNi(PO_3)_4$	1045813	$CaNiP_4O_{12}$	0.151	
$FeSb_3Pb_4O_{13}$*	1224890	$FeSb_3Pb_4O_{13}$	0.027	$Fe_{0.25}Sb_{0.75}$ ↔ $FeSb_3$
$Hf_2Sb_2Pb_4O_{13}$	1224490	$Hf_2Sb_2Pb_4O_{13}$	0.012	
$InSb_3(PO_4)_6$	1224667	$InSb_3P_6O_{24}$	0.011	
$InSb_3Pb_4O_{13}$	1223746	$InSb_3Pb_4O_{13}$	0.029	
$K_2TiCr(PO_4)_3$	1224541	$K_2TiCrP_3O_{12}$	0.009	
$K_4MgFe_3(PO_4)_5$	532755	$K_4MgFe_3P_5O_{20}$	0.076	
$K_4TiSn_3(PO_5)_4$	1224290	$K_4TiSn_3P_4O_{20}$	0.014	
$KBaGdWO_6$	1523079	$KBaGdWO_6$	0.006	
$KBaPrWO_6$	1523149	$KBaPrWO_6$	0.012	
KMn_3O_6*	1223545	KMn_2O_4	0.439	Not a match
$KNa_2Ga_3(SiO_4)_3$	1211711	$KNa_2Ga_3Si_3O_{12}$	0.022	
$KNaP_6(PbO_3)_8$*	1223429	$KNaP_6Pb_8O_{24}$	0.174	$Na_{0.25}K_{0.25}Pb_{0.5}$ ↔ $NaKPb_2$
$KNaTi_2(PO_5)_2$	1211611	$KNaTi_2P_2O_{10}$	0.012	
$KPr_9(Si_3O_{13})_2$*	1223421	$KPr_9Si_6O_{26}$	0.009	$K_{0.1}Pr_{0.9}$ ↔ KPr_9
$Mg_3MnNi_3O_8$	1222170	$Mg_3MnNi_3O_8$	0.029	
Mg_3NiO_4*	1099253	Mg_3NiO_4	0.002	$Mg_{0.75}Ni_{0.25}$ ↔ Mg_3Ni
$MgCuP_2O_7$*	1041741	$MgCuP_2O_7$	0.093	$Mg_{0.5}Cu_{0.5}$ ↔ MgCu
$MgNi(PO_3)_4$	1045786	$MgNiP_4O_{12}$	0.018	
$MgTi_2NiO_6$	1221952	$MgTi_2NiO_6$	0.009	
$MgTi_4(PO_4)_6$	1222070	$MgTi_4P_6O_{24}$	0.075	
$MgV_4Cu_3O_{14}$	1222158	$MgV_4Cu_3O_{14}$	0.060	
Mn_2VPO_7	1210613	Mn_2VPO_7	0.125	
$Mn_4Zn_3(NiO_6)_2$	1222033	$Mn_4Zn_3Ni_2O_{12}$	0.054	
$Mn_7(P_2O_7)_4$	778008	$Mn_7P_8O_{28}$	0.123	
$MnAgO_2$	996995	$MnAgO_2$	0.098	
$Na_3Ca_{18}Fe(PO_4)_{14}$	725491	$Na_3Ca_{18}FeP_{14}O_{56}$	0.031	
$Na_7Mg_7Fe_5(PO_4)_{12}$	1173791	$Na_7Mg_7Fe_5P_{12}O_{48}$	0.028	
$NaCaMgFe(SiO_3)_4$*	1221075	$NaCaMgFeSi_4O_{12}$	0.026	$(MgFeNaCa)_{0.25}$ ↔ MgFeNaCa
$NaMnFe(PO_4)_2$	1173592	$NaMnFeP_2O_8$	0.032	
$Sn_2Sb_2Pb_4O_{13}$	1219056	$Sn_2Sb_2Pb_4O_{13}$	0.025	
$Y_3In_2Ga_3O_{12}$	1207946	$Y_3In_2Ga_3O_{12}$	0.008	
$Zn_2Cr_3FeO_8$	1215741	$Zn_2Cr_3FeO_8$	0.014	
$Zn_3Ni_4(SbO_6)_2$*	1216023	$Zn_3Ni_4Sb_2O_{12}$	0.092	$Ni_{0.67}Sb_{0.33}$ ↔ Ni_2Sb
$Zr_2Sb_2Pb_4O_{13}$	1215826	$Zr_2Sb_2Pb_4O_{13}$	0.025	

Table 4 Close neighbors of each A-lab crystal in the Materials Project. The Materials Project entry with the smallest element mover's distance [19] was selected from the list of 100 nearest neighbors by ADA_{100}. Disordered crystals are marked with an asterisk *.

3.3 A-lab crystals on continuous heatmaps of the ICSD and MP

All figures in this section show scatter plots of the 42 CIFs from the A-lab over heatmaps of the ICSD and MP. The only excluded CIF is the positionally disordered crystal KMn_3O_6. On every map, the colour of any pixel with coordinates (x, y) indicates the number of crystals whose continuous invariants coincide with (x, y) after discretisation. To better visualise the hot spots of the maps, we excluded some outliers in the ICSD and MP, e.g. all crystals with densities higher than 10 g/cm^3 in Fig. 4.

Fig. 4 The scatter plot of A-lab crystals over the ICSD and MP in the coordinates (density, ADA_1).



4 Conclusion: where to go in the materials space?

This paper introduced the materials space in Definition 1 as the *Crystal Isometry Space* containing all known and not yet discovered crystals at unique locations determined by sufficiently precise geometry of atomic centers without atomic types.

Definition 6 introduced the Local Novelty Distance (LND) based on generically complete invariants of periodic point sets. This LND shows how far away any periodic crystal is from its nearest neighbor in a given dataset. The ultra fast speed of LND allows us to find nearest neighbors from the world's largest databases within seconds on a modest desktop computer. Future work will develop another distance characterizing a global novelty or similarity of a crystal relative to a dataset.

Our finding that 42 of the 43 A-lab crystals existed prior to the GNoME project and were seemingly part of its training data but were later targeted for synthesis by the

Fig. 5 The scatter plot of A-lab crystals over the ICSD and MP in the coordinates (PPC, ADA_1).

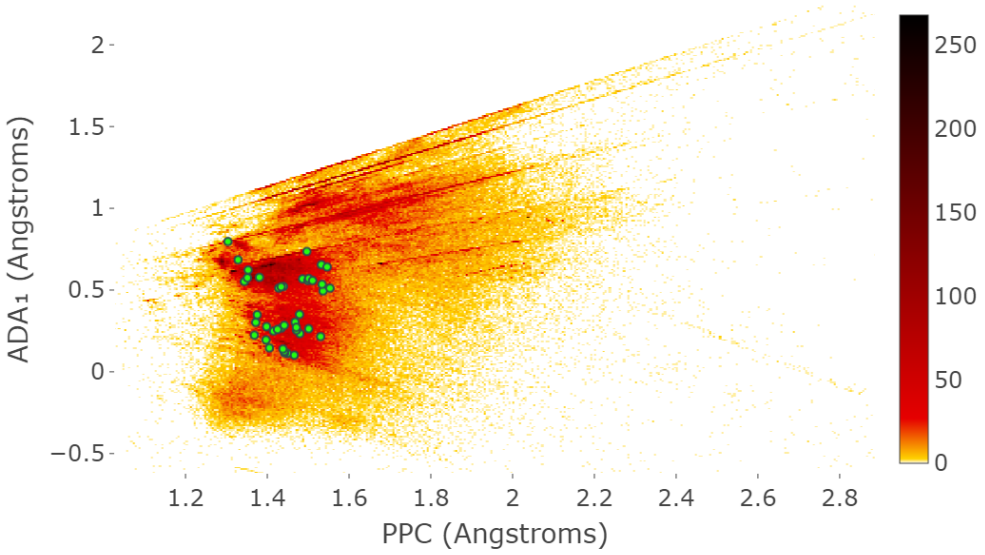


Fig. 6 The scatter plot of A-lab crystals over the ICSD and MP in the coordinates (ADA_2, ADA_3).

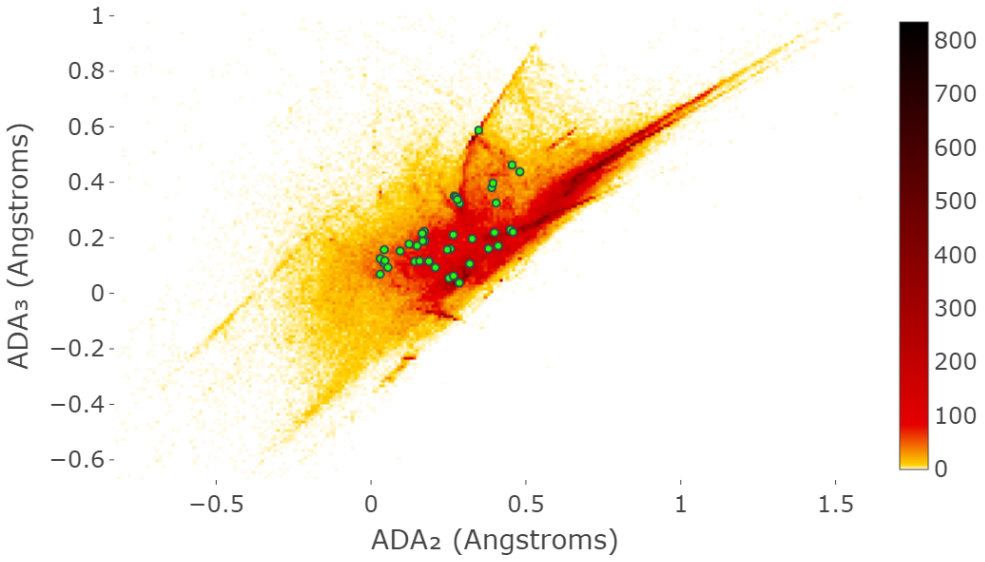


Fig. 7 The scatter plot of A-lab crystals over the ICSD and MP in the coordinates (ADA_5, ADA_4).

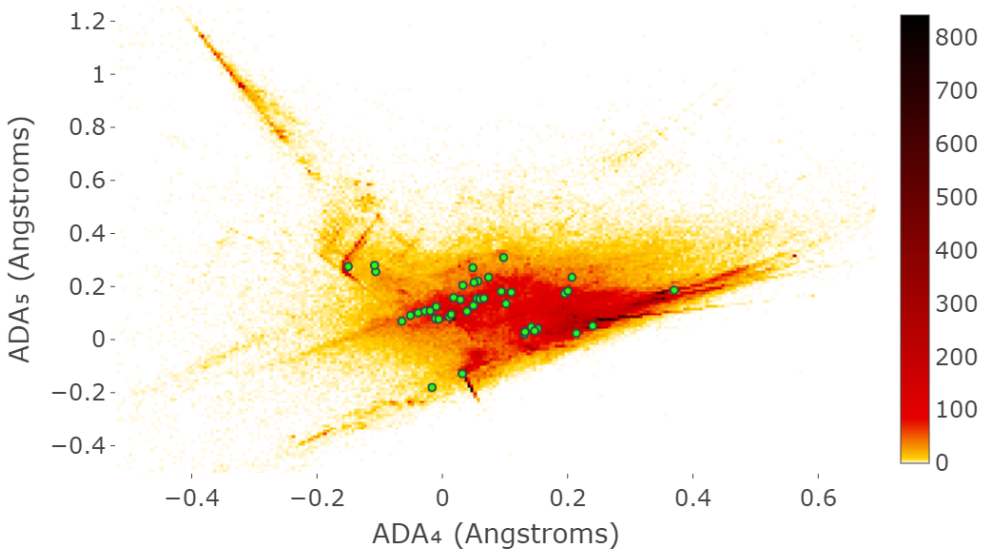


Fig. 8 The plot of A-lab crystals over the ICSD and MP in the coordinates (ADA_1, ADA_{20}).

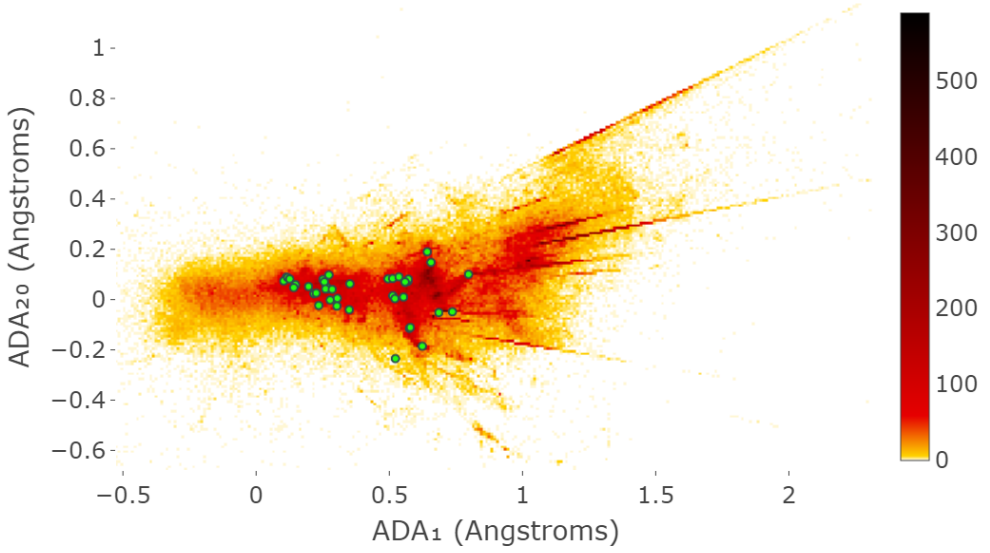
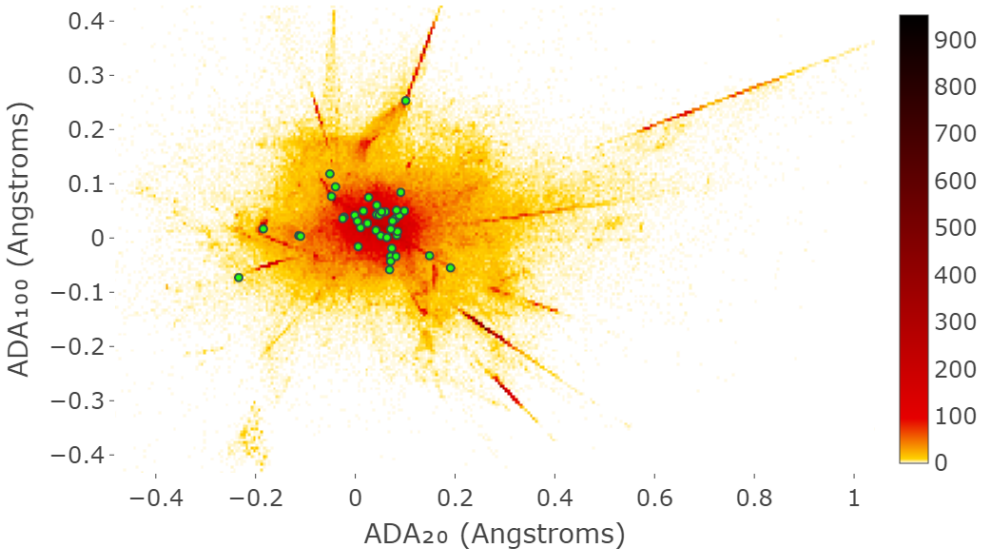


Fig. 9 The plot of A-lab crystals over the ICSD and MP in the coordinates (ADA_{20}, ADA_{100}).



A-lab shows that both the AI's pipeline and the later selection process for materials to target for synthesis would have benefited from the introduction of isometry invariants to find nearest neighbors. As theoretical structures generated by GNoME were re-introduced into its training set, these duplicates can pollute the training data and introduce bias, but could be filtered out by continuous invariants. It is also crucial that the selection process for targets to synthesize by an automated laboratory avoids pre-existing crystals to justify the discovery of truly novel materials. The next step in exploring the materials space CRIS($\mathbb{R}^3$) is to understand the structure-property relations by visualize property values like mountainous landscapes in Fig. 2 (right).

This work was supported by the EPSRC New Horizons grant "Inverse design of periodic crystals" (EP/X018474/1) and the Royal Society APEX fellowship "New geometric methods for mapping the space of periodic crystals" (APX/R1/231152) of the second author. We thank Andy Cooper FRS (the director of Materials Innovation Factory, Liverpool, UK), Robert Palgrave and Leslie Schoop for fruitful discussions of A-lab crystals and any reviewers for their valuable time and helpful suggestions.

References

[1] Orsi, M., Reymond, J.-L.: Navigating a 1e+60 chemical space (2024) https://doi.org/10.26434/chemrxiv-2024-bqd8c

[2] Sacchi, P., Lusi, M., Cruz-Cabeza, A.J., Nauha, E., Bernstein, J.: Same or different – that is the question: identification of crystal forms from crystal structure

data. CrystEngComm **22**(43), 7170–7185 (2020)

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

[4] Widdowson, D., Kurlin, V.: Resolving the data ambiguity for periodic crystals. Advances in Neural Information Processing Systems (NeurIPS) **35**, 24625–24638 (2022)

[5] Feynman, R.: The Feynman Lectures on Physics vol. 1, (1971)

[6] Niggli, P.: Krystallographische und Strukturtheoretische Grundbegriffe vol. 1. Akademische verlagsgesellschaft mbh, ??? (1928)

[7] Lawton, S., Jacobson, R.: The reduced cell and its crystallographic applications. Technical report, Ames Lab., Iowa State Univ. of Science and Tech., US (1965)

[8] Ward, S.C., Sadiq, G.: Introduction to the cambridge structural database – a wealth of knowledge gained from a million structures. CrystEngComm **22**(43), 7143–7144 (2020)

[9] Widdowson, D., Mosca, M.M., Pulido, A., Cooper, A.I., Kurlin, V.: Average minimum distances of periodic point sets - foundational invariants for mapping all periodic crystals. MATCH Comm. in Math. and in Computer Chemistry **87**, 529–559 (2022)

[10] Bravais, A.: Memoir on the systems formed by points regularly distributed on a plane or in space. J. École Polytech. **19**, 1–128 (1850)

[11] Kurlin, V.: A complete isometry classification of 3D lattices. arxiv:2201.10543 (2022)

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

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

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

[15] Zagorac, D., Müller, H., Ruehl, S., Zagorac, J., Rehme, S.: Recent developments in the inorganic crystal structure database: theoretical crystal structure data and related features. Journal of applied crystallography **52**(5), 918–925 (2019)

[16] Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., et al.: Commentary: The materials project: A materials genome approach to accelerating materials innovation. APL materials **1**(1) (2013)

[17] Terban, M.W., Billinge, S.J.: Structural analysis of molecular materials using the pair distribution function. Chemical Reviews **122**, 1208–1272 (2022)

[18] Patterson, A.: Homometric structures. Nature **143**, 939–940 (1939)

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

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

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

[22] Gražulis, S., Daškevič, A., Merkys, A., Chateigner, D., Lutterotti, L., Quiros, M., Serebryanaya, N.R., Moeck, P., Downs, R.T., Le Bail, A.: Crystallography open database (cod): an open-access collection of crystal structures and platform for world-wide collaboration. Nucleic acids research **40**(D1), 420–427 (2012)

[23] Batsanov, S.S.: Van der waals radii of elements. Inorganic materials **37**(9), 871–885 (2001)

[24] Leeman, J., Liu, Y., Stiles, J., Lee, S.B., Bhatt, P., Schoop, L.M., Palgrave, R.G.: Challenges in high-throughput inorganic materials prediction and autonomous synthesis. PRX Energy **3**(1), 011002 (2024)

[25] Chawla, D.S.: Crystallography databases hunt for fraudulent structures. ACS Central Science **9**, 1853–1855 (2024) https://doi.org/10.1021/acscentsci.3c01209

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

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

[28] Parthé, E., Gelato, L., Chabot, B., Penzo, M., Cenzual, K., Gladyshevskii, R.: TYPIX Standardized Data and Crystal Chemical Characterization of Inorganic Structure Types. Springer, ??? (2013)

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

[30] Pulido, A., Chen, L., Kaczorowski, T., Holden, D., Little, M.A., Chong, S.Y., Slater, B.J., McMahon, D.P., Bonillo, B., Stackhouse, C.J., *et al.*: Functional

materials discovery using energy–structure–function maps. Nature **543**(7647), 657–664 (2017)

[31] Widdowson, D., Kurlin, V.: Recognizing rigid patterns of unlabeled point clouds by complete and continuous isometry invariants with no false negatives and no false positives. Proceedings of Computer Vision and Pattern Recognition, 1275–1284 (2023)

[32] Bartók, A.P., Kondor, R., Csányi, G.: On representing chemical environments. Physical Review B **87**(18), 184115 (2013)

[33] Kovács, D.P., Batatia, I., Arany, E.S., Csányi, G.: Evaluation of the mace force field architecture: From medicinal chemistry to materials science. The Journal of Chemical Physics **159**(4) (2023)

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

[35] Anosova, O., Kurlin, V.: An isometry classification of periodic point sets. In: LNCS (Proceedings of DGMM), vol. 12708, pp. 229–241 (2021)

[36] Anosova, O., Kurlin, V.: Recognition of near-duplicate periodic patterns by continuous metrics with approximation guarantees. arxiv:2205.15298 (2022)

[37] Szymanski, N.J., Rendy, B., Fei, Y., Kumar, R.E., He, T., Milsted, D., McDermott, M.J., Gallant, M., Cubuk, E.D., Merchant, A., et al.: An autonomous laboratory for the accelerated synthesis of novel materials. Nature, 86–91 (2023)

[38] Azrour, M., Azdouz, M., Manoun, B., Essehli, R., Benmokhtar, S., Bih, L., El Ammari, L., Ezzahi, A., Ider, A., Hou, A.A.: Rietveld refinements and vibrational spectroscopic studies of na1-xkxpb4(po4)3 lacunar apatites ($0 \leq x \leq 1$). Journal of Physics and Chemistry of Solids **72**(11), 1199–1205 (2011) https://doi.org/10.1016/j.jpcs.2011.06.013

[39] Cerqueira, T.F.T., Lin, S., Amsler, M., Goedecker, S., Botti, S., Marques, M.A.L.: Identification of novel cu, ag, and au ternary oxides from global structural prediction. Chemistry of Materials **27**(13), 4562–4573 (2015) https://doi.org/10.1021/acs.chemmater.5b00716

[40] Griesemer, S.D., Ward, L., Wolverton, C.: High-throughput crystal structure solution using prototypes. Physical Review Materials **5**(10), 105003 (2021)

[41] Merchant, A., Batzner, S., Schoenholz, S.S., Aykol, M., Cheon, G., Cubuk, E.D.: Scaling deep learning for materials discovery. Nature, 80–85 (2023)

[42] Quarez, E., Abraham, F., Mentré, O.: Synthesis, crystal structure and characterization of new 12h hexagonal perovskite-related oxides ba6m2na2x2o17 (m= ru,

nb, ta, sb; x= v, cr, mn, p, as). Journal of Solid State Chemistry **176**(1), 137–150 (2003)

[43] JULIEN POUZOL, M., JAULMES, S., LE HUI, Z.: Structure cristalline du grenat gd3-xcaxga5-xzrxo12. Comptes rendus de l'Académie des sciences. Série 2, Mécanique, Physique, Chimie, Sciences de l'univers, Sciences de la Terre **306**(8), 531–535 (1988)

[44] Selker, P., Klaska, R.: Struktur und hydrothermalsynthesen von beryllonittypen im system (na, k) oh-al (oh) 3-ga2o3-sio2-geo2. Zeitschrift für Kristallographie **159**(1-4), 119–120 (1982)

[45] Azrour, M., Azdouz, M., Manoun, B., Essehli, R., Benmokhtar, S., Bih, L., El Ammari, L., Ezzahi, A., Ider, A., Hou, A.A.: Rietveld refinements and vibrational spectroscopic studies of na1- xkxpb4 (po4) 3 lacunar apatites (0≤ x≤ 1). Journal of Physics and Chemistry of Solids **72**(11), 1199–1205 (2011)

[46] Crennell, S.J., Owen, J.J., Grey, C.P., Cheetham, A.K., Kaduk, J.A., Jarman, R.H.: Isomorphous substitution in non-linear optical ktiopo 4. powder diffraction and magic angle spinning nuclear magnetic resonance study of (k 1/2 na 1/2) tiopo 4 and (rb 1/2 na 1/2) tiopo 4. Journal of Materials Chemistry **1**(1), 113–119 (1991)

[47] Yakubovich, O., Anan'eva, E., Dimitrova, O.: Crystal structure of the solid solution mn 2 (p x v 1-x)(v y p 1-y) o 7. Koordinatsionnaya Khimiya **26**(8), 586–591 (2000)

[48] Schmitz-Dumont, O., Moulin, N.: Farbe und konstitution bei anorganischen feststoffen. vi. über die lichtabsorption des dreiwertigen chroms in indiumhaltigen wirtsgittern mit granatstruktur. Zeitschrift für anorganische und allgemeine Chemie **330**(5-6), 259–266 (1964)

[49] Li, C., Zhong, J.: Highly efficient broadband near-infrared luminescence with zero-thermal-quenching in garnet y3in2ga3o12: Cr3+ phosphors. Chemistry of Materials **34**(18), 8418–8426 (2022)

[50] Pymatgen structure matcher. https://pymatgen.org/pymatgen.analysis.html#module-pymatgen.analysis.structure_matcher

[51] Edelsbrunner, H., Heiss, T., Kurlin, V., Smith, P., Wintraecken, M.: The density fingerprint of a periodic point set. In: Proceedings of SoCG, pp. 32–13216 (2021)

Appendix A Proof of invariant distance properties

Proof of Theorem 7. Let S be obtained from a periodic point set $Q \subset \mathbb{R}^n$ by perturbing every point of Q up to Euclidean distance ε, which is smaller than a minimum half-distance between any points of Q. Then S, Q have a common lattice by [51,

Lemma 4.1] and hence the same number m of points in a common unit cell, and equal Point Packing Coefficients $\mathrm{PPC}(S) = \mathrm{PPC}(Q)$ from Definition 3.

Since Definition 4 uses the L_∞ metric on rows of PDAs, the Earth Mover's Distance is unaffected by subtracting the same term $\mathrm{PPC}\sqrt[3]{k}$, so $\mathrm{EMD}(\mathrm{PDD}(S;k), \mathrm{PDD}(Q;k)) = \mathrm{EMD}(\mathrm{PDA}(S;k), \mathrm{PDA}(Q;k))$. Then [9, Theorem 4.3] implies that $\mathrm{EMD}(\mathrm{PDA}(S;k), \mathrm{PDA}(Q;k)| \leq 2\varepsilon$. The minimum for all sets Q in a finite dataset D can not be larger, so $\mathrm{LND}(S;D) \leq 2\varepsilon$ by Definition 6.

Conversely, assume that S is obtained from $Q \in D$ by perturbing every atom of Q up to Euclidean distance $\varepsilon < 0.5\mathrm{LND}(S;D) < r(Q)$. The previously proved inequality implies that $\mathrm{LND}(S;D) \leq 2\varepsilon < \mathrm{LND}(S;D)$, which is a contradiction. □