Geographic-style maps with a local novelty distance help navigate in the material space

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

[1]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 the 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.

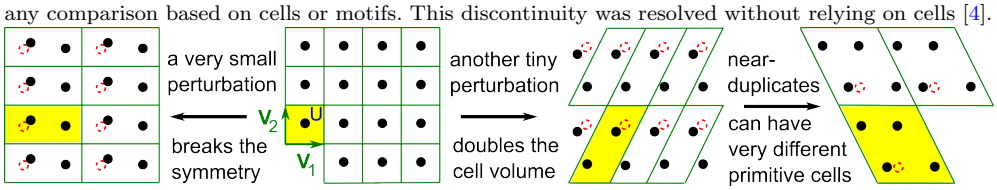
Keywords: materials space, crystal structure, isometry invariant, continuous metric

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 many 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 finalized in the periodic case in [3], so we focus on ideal periodic crystals (briefly, *crystals*) as formalized 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 centers $p_1, \dots, p_m$ 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].



Since atoms always 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].

Since 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 centers (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 mirror reflections. Since mirror images can be distinguished by a suitable sign of orientation, the main difficulty is to classify periodic point sets under isometry.

When comparing crystals as periodic sets of atomic centers 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.

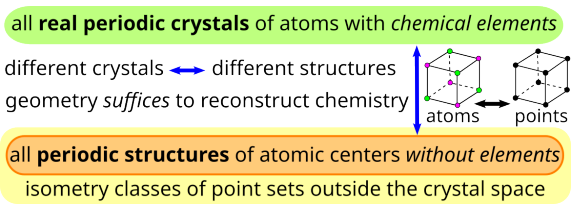
This Crystal Isometry Principle (CRISP) first appeared in 2022 [9, section 7] and was inspired by Richard Feynman's hint in Fig.1-7 [5, chapter 1], which distinguished 7 cubic crystals by their cube size in the first lecture "Atoms in motion", see Fig. 2 (left).

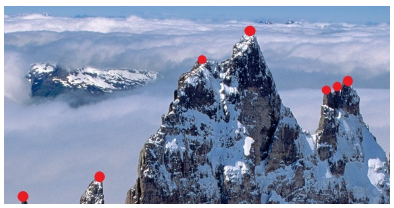
More importantly, when we consider atoms only as zero-sized points, we can study all periodic structures in a common space similar to the periodic table of all elements.

In the geographic analogy, the chemical composition can be compared to the altitude (the height above the sea level) of any location on Earth. If our geographic map is precise enough, we can determine the average temperature or any other property at every location. If we know the altitude (chemical composition) in addition to geographic coordinates (structural invariants), the property prediction will be easier.

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.





Since (the isometry class of) any periodic point set has a unique location in CRIS($\mathbb{R}^3$), all known materials can be considered 'visible stars' in this continuous universe. Any periodic crystal discovered in the future will appear at a new unique location like a 'new star', while all past crystals remain at the same locations. Until recently, optimizing complicated energy functions blindly climbed as a mountaineer to a high peak, illustrated in Fig. 2 (right), and stopping at many (approximations to) isolated peaks without even any reliable method to continuously measure distances between these peaks, while the remaining landscape was covered by clouds. Our vision is to map the full space CRIS($\mathbb{R}^3$) to enable a non-blind discovery of materials [10].

If we do not restrict the motif size, the space CRIS is infinitely dimensional. However, if we consider all periodic sets with 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 neutralized by translations along basis vectors. Alternatively, we can define a unit cell by 3 basis vectors with 3 coordinates, of which 6 are neutralized 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 in 14 disjoint subspaces of Bravais classes [11] but is now parametrized by complete invariants [12, 13]. Continuous maps of the simpler 3-dimensional space CRIS($\mathbb{R}^2; 1$) of 2D lattices recently appeared in [14], [15], [16].

The full space CRIS($\mathbb{R}^3$) $= \cup_{m=1}^{+\infty}$CRIS($\mathbb{R}^3; m$) is a union of infinitely many subspaces for $m = 1, 2, 3, \dots$ 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 of this work is the local novelty distance based on generically complete invariants, which identify closest neighbors of the 43 A-lab crystals in the Inorganic Crystal Structure Database (ICSD) [17] and Materials Project (MP) [18] within seconds on a desktop computer. The second contribution is the geographic-style maps showing how the ICSD and MP populate CRIS($\mathbb{R}^3$) in invariant coordinates.

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] and [19, Theorem 5.8]. Generic means any set apart from a singular subspace of measure 0, e.g. almost any noise makes every crystal generic.

The PDD is a matrix of inter-point distances and is stronger than the Pair Distribution Function (PDF) [20] in the sense that PDD can be simplified to PDF but distinguishes homometric structures [21] that have the same PDF [4, section 3].

Definition 2 (isometry invariant $\mathrm{PDD}(S;k)$)**.** 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 $d_1(p) \leq \dots \leq d_k(p)$ be the distances from p to its k nearest neighbors within the full infinite set S not restricted to any cell. The matrix $D(S;k)$ has m rows consisting of the distances $d_1(p_i), \dots, d_k(p_i)$ for $i = 1, \dots, m$. If any $l \geq 1$ 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 neighbors at the same distance but the k smallest distances (without any indices or types of neighbors) are always well-defined. The matrix $\mathrm{PDD}(S;k)$ has ordered columns (according to the index of neighbors) but unordered rows because points of a motif of S are unordered. The appendix computes a weighted PDD with atomic masses as extra weights of rows in $\mathrm{PDD}(S;k)$. Using only on atomic centers detects duplicates where chemical elements were artificially replaced without changing geometry, see [3, Table 1].

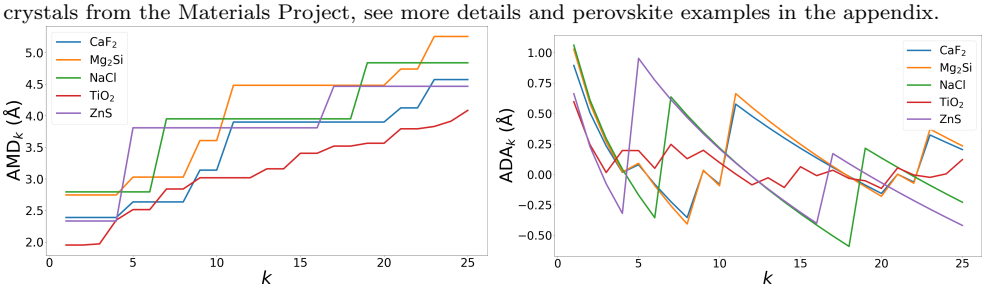
If the number k of neighbors increases to infinity, the asymptotic behavior of distances to neighbors is described in terms of the Point Packing Coefficient below.

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 behavior 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 $k \geq 1$. 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 $i, k \geq 1$.

Fig. 3 The average invariants AMD_k and ADA_k from Definition 4 for $k = 1, \dots, 25$ and five simple crystals from the Materials Project, see more details and perovskite examples in the appendix.

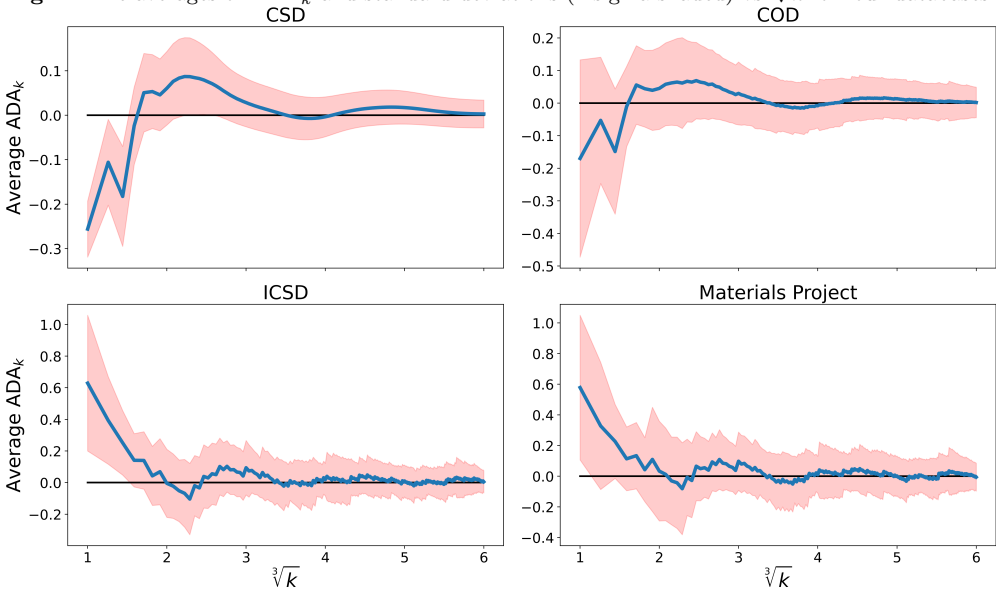


The invariants AMD_k and ADA_k form vectors of length k, e.g. set $\mathrm{AMD}(S;k) = (\mathrm{AMD}_1(S), \dots, \mathrm{AMD}_k(S))$ and $\mathrm{ADA}(S;k) = (\mathrm{ADA}_1(S), \dots, \mathrm{ADA}_k(S))$. These vectors can be compared by many metrics. The metric $L_\infty(u,v) = \max\limits_{i=1,\dots,k} |\boldsymbol{u}_i - \boldsymbol{v}_i|$ 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 $L_\infty(\mathrm{AMD}(S;k), \mathrm{AMD}(Q;k)) \leq 2\varepsilon$ 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. 4 reveals geometric differences between the mainly organic databases CSD and Crystallography Open Database (COD) [22] versus the more inorganic collections ICSD and MP.

Fig. 4 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. 4 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. 4.

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 experimental convergence $\mathrm{ADA}_k \to 0$ as $k \to +\infty$ in Fig. 4justifies computing the distance L_∞ between ADA vectors up to a reasonable k. We use $k = 100$ because all ADA_k for $k > 100$ are close to 0 (the range of 1 sigma between ± 0.2Å) in Fig. 4.

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 LND will be measured as a distance between $\mathrm{PDA}(S;k)$ and $\mathrm{PDA}(Q;k)$ for $Q \in D$.

We can compare PDD matrices that have the 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. Many similarities between discrete

distributions such as the Cramer - von Mises distance and the Kullback - Leibler (KL) divergence fail the axioms of a metric, which are prerequisites for convergence guarantees. 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 [24].

Though the KL divergence can be symmetrized to the Jensen-Shannon divergence whose square root becomes a metric [25], these divergences work best for distributions on the same finitely many discrete values, while PDD matrices of different crystals are unlikely to share any common rows (points in $\mathbb{R}^k$) consisting of k continuous distances.

We use the Earth Mover's Distance (EMD) on PDDs, see Definition 7 in the appendix because EMD is invariant under permutations of rows in PDDs and EMD satisfies all metric axioms [26, Appendix]. The EMD can be extended to more complicated Wasserstein metrics [27] but the simplest EMD behaves most nicely under bounded noise, motivated by atomic vibrations. If any point is perturbed up to ε, any inter-point distance (a value in PDD) can become smaller or larger but only up to 2ε, which will allow us to prove the same upper bound 2ε for EMD in Theorem 6.

Definition 5 (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* $\mathrm{LND}(S;D) = \min\limits_{Q\in D} \mathrm{EMD}(\mathrm{PDA}(S;k), \mathrm{PDA}(Q;k))$ is the shortest L_∞-based EMD distance from S to its nearest neighbor Q in the given crystal dataset D.

If S is already contained in the dataset D, then $\mathrm{LND}(S;D) = 0$, so S cannot be considered novel. Conversely, if $\mathrm{LND}(S;D) = 0$ then S highly likely belongs to S, because $\mathrm{PDD}(S;100)$ distinguished all non-duplicate periodic crystals in the CSD.

Also, 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]. $\mathrm{LND}(S;D)$ is based on PDAs instead of PDDs because distances to k-th neighbors in $\mathrm{PDD}(S;k)$ asymptotically increase as $\mathrm{PPC}(S)\sqrt[3]{k}$ by [9, Theorem 4.4]. If crystals S, Q have $\mathrm{PPC}(S) \neq \mathrm{PPC}(Q)$, the distance L_∞ between rows of PDDs equals the largest absolute difference of i-th distances, which likely happens for $i = k$. So subtracting $\mathrm{PPC}(S)\sqrt[3]{k}$ in Definition 4 makes any metric on PDAs more informative than on PDDs. If a newly synthesized periodic crystal S is a near-duplicate of some known $Q \in D$, 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 6. 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 6 is proved in Appendix A. The distance $\mathrm{LND}(S;D)$ is called *local* because Definition 5 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 [28], [29], and [30], [3, Tables 1-2 in section 6], respectively.

2.3 Insufficiency of past invariants and similarities of crystals

This subsection briefly reviews the past approaches to classify crystals. Some widely used similarities such as the Root Mean Square Deviation (RMSD) [31] deserve their own detailed discussions in another forthcoming work. Conventional settings were thoroughly developed to uniquely represent any periodic crystal in a reduced cell [32] and can be theoretically considered complete under rigid motion but discontinuously change under almost any perturbation of atoms in practice as shown in Fig. 1.

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 [33] inevitably uses thresholds and leads to a trivial classification due to the transitivity axiom, see [3, section 1].

The COMPACK algorithm [31] 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. Even for simple crystals based on a single molecule as is often the case in Crystal Structure Prediction [34], the choice of 14 (or any other number of) neighbors can be discontinuous when a central molecule has 14th and 15th neighbors 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 resolved in [35] but a choice of a single atomic environment in a periodic crystal remains discontinuous as shown in Fig. 1.

Other similarities based on all atomic environments such as SOAP [36] and MACE [37] 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 finite order can become incomplete. Appendix A explains why the structure matcher from pymatgen [38] can classify many near-duplicates as "unique crystal structures".

The PXRD similarity compares crystals through powder diffraction patterns that are identical for all homometric structures [21], some of which were distinguished even by AMD_2 in [9, appendix A]. The PXRD as implemented in Mercury [39] 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 [40] for lattices and then extended in [41] 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 [42].

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 [43], 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 $\mathrm{PDA}(S;100)$. 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 nearest neighbors with small distances EMD, we selected the neighbor with the most similar composition as measured by element mover's distance [44] 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 EMD_{100}. All experiments were run on a desktop computer: AMD Ryzen 5 5600X (6-core), 32GB RAM, Python 3.9, see the Python code with instructions and examples in the supplementary materials. Table 1 shows running times, see a linear-time asymptotic of neighbor search in [45].

The GNoME (Graph Network Materials Exploration) [46] was trained on a snapshot of the Materials Project database (whose entries are partly sourced from the ICSD) from 2021 and made public 384,938 crystals. Berkeley's A-lab attempted to synthesize 58 of them and reported 43 [43], which were split into 36 "successes" and 7 "partial successes" (less than 50% of the weight of solute versus the weight of solution).

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

Table 1 Time (seconds) to complete each stage of the process of finding nearest neighbors in the ICSD and Materials Project for 43 A-lab crystals [43] on a modest desktop computer. The binary search used 6-cores for multiprocessing.

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	$(Ca_{0.43}La_{0.57})_2Ba \leftrightarrow Ca_{0.33}La_{0.67}(Ca_{0.5}Ba_{0.5})_2$
$CaCo(PO_3)_4$	300027	$Co_2P_4O_{12}$	0.172	Ca ↔ Co
$CaFe_2P_2O_9$	79735	$CaV_2P_2O_9$	0.073	Fe ↔ V
$CaGd_2Zr(GaO_3)_4$*	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	$SbHf \leftrightarrow Sn_{0.72}W_{0.28}$
$InSb_3(PO_4)_6$	72735	$Sb_2P_3O_{12}$	0.193	In ↔ Sb
$InSb_3Pb_4O_{13}$	49531	$Pb_2Ru_2O_{6.5}$	0.147	SbIn ↔ Ru
$K_2TiCr(PO_4)_3$	40307	$K_2P_3Ti_2O_{12}$	0.098	Cr ↔ Ti
$K_4MgFe_3(PO_4)_5$	263040	$Fe_4K_4P_5O_{20}$	0.139	Mg ↔ Fe
$K_4TiSn_3(PO_5)_4$	79650	$KPSnO_5$	0.094	Ti ↔ Sn
$KBaGdWO_6$	60499	$WCaBa_2O_6$	0.009	GdK ↔ CaBa
$KBaPrWO_6$	60499	$WCaBa_2O_6$	0.053	PrK ↔ 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 ↔ GaSiNa
$KNaP_6(PbO_3)_8$*	182501	$KNaP_6Pb_8O_{24}$	0.005	
$KNaTi_2(PO_5)_2$	68705	$KPTiO_5$	0.157	Na ↔ K
$KPr_9(Si_3O_{13})_2$*	153272	$KSi_6Pr_9O_{26}$	0.16	$(K_{0.1}Pr_{0.9})_2 \leftrightarrow PrK_{0.25}Pr_{0.75}$
$Mg_3MnNi_3O_8$	40584	$MnNi_6O_8$	0.020	Mg ↔ Ni
Mg_3NiO_4*	109086	$TaLi_3O_4$	0.000	$Mg_{0.75}Ni_{0.25} \leftrightarrow Li_{0.75}Ta_{0.25}$
$MgCuP_2O_7$*	69576	$Co_{0.92}Mg_{1.08}P_2O_7$	0.218	$Mg_{0.5}Cu_{0.5} \leftrightarrow Mg_{0.54}Co_{0.46}$
$MgNi(PO_3)_4$	37137	$NiZnP_4O_{12}$	0.132	Mg ↔ Zn
$MgTi_2NiO_6$	171584	$NiTiO_3$	0.047	Mg ↔ Ni
$MgTi_4(PO_4)_6$	419418	$MnTi_4P_6O_{24}$	0.133	Mg ↔ Mn
$MgV_4Cu_3O_{14}$	164189	$Cu_2V_2O_7$	0.146	Mg ↔ Cu
Mn_2VPO_7	20296	$Mn_2P_2O_7$	0.21	V ↔ P
$Mn_4Zn_3(NiO_6)_2$	625	$MgCu_2Mn_3O_8$	0.186	MnZnNi ↔ MgCuMn
$Mn_7(P_2O_7)_4$	67514	$Fe_7P_8O_{28}$	0.126	Mn ↔ 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	$POMn_2 \leftrightarrow Na_2Fe_2$
$Sn_2Sb_2Pb_4O_{13}$	49533	$PbNb_2Tl_{0.9}O_{6.45}$	0.209	SbSnPb ↔ NbTl
$Y_3In_2Ga_3O_{12}$	185862	$Y_3Ga_5O_{12}$	0.104	In ↔ Ga
$Zn_2Cr_3FeO_8$	196119	$ZnCr_2O_4$	0.022	Fe ↔ 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	$SbZr \leftrightarrow Ti_{0.5}Sb_{0.5}$

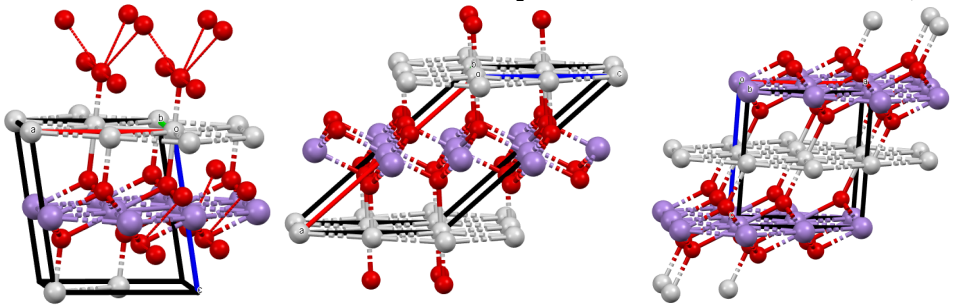
Table 2 Close neighbors of each A-lab crystal in the ICSD. The ICSD entry with the smallest element mover's distance [44] was selected from the list of 100 nearest neighbors by ADA_{100}. Disordered crystals are marked with an asterisk *.

Table I in [28] summarized four types of issues for the 36 "successes", where only 3 were marked as already reported structures. Table 2 lists geometric close matches that were automatically found in the ICSD for all 43 A-lab crystals within a few seconds.

Two A-lab crystals were found to already exist in the ICSD with the same composition: $KNaP_6(PbO_3)_8$ matched ICSD 182501 reported in 2011 [47], and $MnAgO_2$

matched ICSD 670065 reported as a hypothetical structure in 2015 [48]. In particular, $MnAgO_2$ was one of three crystals that the later rebuttal said was synthesized successfully [28], and they go on to state that the material was first reported in 2021 [49] (ICSD 139006), after the snapshot used to train the GNoME, 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 [28] 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, see Fig. 5. Such cell-based search can always miss near-duplicates as in Fig. 1, while continuous invariants independent of a unit cell find near-duplicates despite disagreement on a space group, which breaks down under almost any noise.

Fig. 5 **Left**: $MnAgO_2$ synthesized by A-lab. **Middle**: ICSD entry 670065 with the same composition and EMD = 0.097Å found by structural invariants in Table 2, though its unit cell is very different from the cell of $MnAgO_2$. **Right**: another ICSD entry 139006 from 2021 matched by [28] and found by unit cell search, but is more distant from $MnAgO_2$ by EMD = 0.368Å on invariants PDA(S; 100).



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 [46] 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 [28] suggests that a more robust method is needed for comparing structures in the aid of materials discovery.

The Materials Project contains many theoretical structures, many of which are obtained by substituting atoms in experimental structures with plausible alternatives, a strategy also employed by the GNoME 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 and hence were part of its training data.

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. So all 43 A-lab crystals had already been hypothesized or synthesized prior to the beginning of the GNoME project, see Table 3.

The rebuttal paper [28] said that the crystal $Y_3In_2Ga_3O_{12}$ in Table 3 was one of the three new crystals to have been synthesized and provided the reference for this crystal to 2022 [56], again leading to the conclusion that the crystal was novel from the perspective of the GNoME AI trained on data from 2021. We found that the crystal $Y_3In_2Ga_3O_{12}$ 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 $Fe_{0.25}Sb_{0.75}$

A-lab name	Matching database entries	Source and date
$Ba_6Na_2Ta_2V_2O_{17}$	mp-1214664, Pauling file sd_1003187	[50], 2003
$Ba_6Na_2V_2Sb_2O_{17}$	mp-1214658, Pauling file sd_1003189	[50], 2003
$CaGd_2Zr(GaO_3)_4$	mp-686296, ICSD 202850	[51], 1988
$KNa_2Ga_3(SiO_4)_3$	mp-1211711, Pauling file sd_1707156	[52], 1982
$KNaP_6(PbO_3)_8$	ICSD 182501	[47], 2011
$KNaTi_2(PO_5)_2$	mp-1211611, Pauling file sd_1414297	[53], 1991
Mn_2VPO_7	mp-1210613, Pauling file sd_1322766	[54], 2000
$Y_3In_2Ga_3O_{12}$	mp-1207946, Pauling file sd_1704376	[55], 1964

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

A-lab name	MP ID	MP composition	EMD_{100}	Site mismatches
Ba_2ZrSnO_6*	1228067	Ba_2ZrSnO_6	0.025	$Zr_{0.5}Sn_{0.5} \leftrightarrow ZrSn$
$Ba_6Na_2Ta_2V_2O_{17}$	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	$Ca_{0.43}La_{0.57} \leftrightarrow Ca_3La_4$
$CaCo(PO_3)_4$	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	$Ga \leftrightarrow Zr$
$CaMn(PO_3)_4$	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	$Fe_{0.25}Sb_{0.75} \leftrightarrow FeSb_3$
$Hf_2Sb_2Pb_4O_{13}$	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	$Na_{0.25}K_{0.25}Pb_{0.5} \leftrightarrow NaKPb_2$
$KNaTi_2(PO_5)_2$	1211611	$KNaTi_2P_2O_{10}$	0.012	
$KPr_9(Si_3O_{13})_2$*	1223421	$KPr_9Si_6O_{26}$	0.009	$K_{0.1}Pr_{0.9} \leftrightarrow KPr_9$
$Mg_3MnNi_3O_8$	1222170	$Mg_3MnNi_3O_8$	0.029	
Mg_3NiO_4*	1099253	Mg_3NiO_4	0.002	$Mg_{0.75}Ni_{0.25} \leftrightarrow Mg_3Ni$
$MgCuP_2O_7$*	1041741	$MgCuP_2O_7$	0.093	$Mg_{0.5}Cu_{0.5} \leftrightarrow MgCu$
$MgNi(PO_3)_4$	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	$(MgFeNaCa)_{0.25} \leftrightarrow MgFeNaCa$
$NaMnFe(PO_4)_2$	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	$Ni_{0.67}Sb_{0.33} \leftrightarrow Ni_2Sb$
$Zr_2Sb_2Pb_4O_{13}$	1215826	$Zr_2Sb_2Pb_4O_{13}$	0.025	

Table 4 Close neighbors of each A-lab crystal in the Materials Project (MP). In each case, the MP entry with the smallest element mover's distance [44] was selected from the list of 100 nearest neighbors by ADA_{100}. Disordered crystals are marked with an asterisk *.

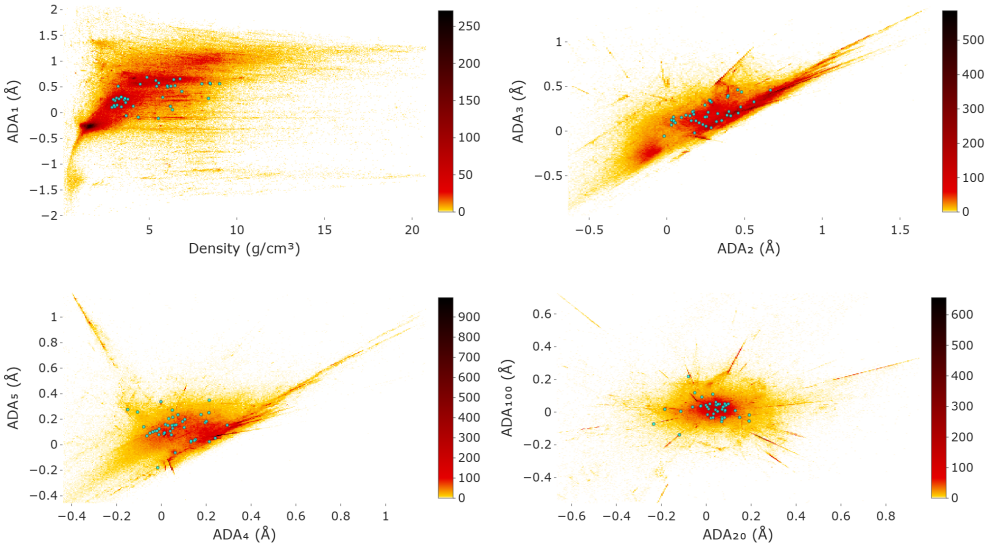
with multiplicity 4 replaced with $FeSb_3$. For completeness, this is noted in the site mismatches column of Table 4, listing all nearest neighbors in the Materials Project.

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.

Fig. 6 shows 2D projections (heat maps) of the ICSD and MP to pairs of analytically defined (data-independent) invariant coordinates, see continuous maps of the CSD and its subsets in [57]. The color of any pixel with coordinates (x, y) indicates the number of crystals whose continuous invariants coincide with (x, y) after discretization to pixels. To better visualize hot spots, we excluded some outliers, e.g. all crystals with densities higher than 21 g/cm^3. Subspaces of highly symmetric (cubic or primitive orthorhombic) crystals are visible as straight lines due to linear dependencies of inter-atomic distances in these subspaces. The projections in Fig. 6 can be considered universal maps of the continuous space of all crystal structures because any newly discovered crystal will appear at its unique location without affecting all known crystals.

Fig. 6 A-lab crystals are in cyan over the heat map of ICSD and MP in invariant coordinates.



4 Conclusions: fast navigation in the materials space

Definition 1 formalized the materials universe as the *Crystal Isometry Space* CRIS containing all known and not yet discovered crystals at unique locations determined by sufficiently precise geometry of atomic centers. Definition 5 introduced the Local Novelty Distance (LND) based on generically complete invariants of periodic point sets. The ultra fast LND quantifies the novelty of any synthesized material as a continuous distance to the nearest crystal structure (independent of chemical composition) from the world's largest databases within seconds on a modest desktop computer.

Tables 2 and 4 showed that structural near-duplicates of all A-lab crystals existed before the GNoME project and were seemingly part of its training data but were targeted for synthesis. As with car driving, navigation maps based on structural invariants in Fig. 6 are needed to guide synthesis without getting lost in the materials space.

The next step in exploring the materials space is to understand the structure-property relations by visualizing property values like mountainous landscapes in Fig. 2 (right). The invariant coordinates PDD (generically complete under isometry) helped predict properties of organic and inorganic materials [58–60] including synthesizability [61]. Similar maps of the protein universe used linear-time invariants [62], which detected thousands of unexpected duplicates in the Protein Data Bank [63].

Authors' contributions. VK developed Definitions 1 and 5 and wrote the paper. DW implemented the code and produced all tables. Both authors reviewed the manuscript.

Acknowledgements. This work was supported by the EPSRC 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). We thank Andy Cooper, Robert Palgrave and Leslie Schoop for fruitful discussions.

Data Availability. Data is provided within the supplementary information files.

References

[1] Orsi, M., Reymond, J.-L.: Navigating a 1e+60 chemical space of peptide/peptoid oligomers. Molecular Informatics **44**(1), 202400186 (2025)

[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 **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. Math. Comp. Chemistry **87**, 529–559 (2022)

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

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

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

[13] Bright, M.J., Cooper, A.I., Kurlin, V.A.: Welcome to a continuous world of 3-dimensional lattices. arxiv:2109.11538 (2021)

[14] Kurlin, V.: Mathematics of 2d lattices. Found. Comp. Mathematics **24**, 805–863 (2024)

[15] 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)

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

[17] 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)

[18] Jain, A., et al.: Commentary: The materials project: A materials genome approach to accelerating materials innovation. APL materials **1**(1) (2013)

[19] Widdowson, D., Kurlin, V.: Pointwise distance distributions for detecting near-duplicates in large materials databases. arxiv:2108.04798v3 (2021)

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

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

[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.: Van der Waals radii of elements. Inorganic mat. **37**, 871–885 (2001)

[24] 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 **19**, 7304–7319 (2024)

[25] Endres, D.M., Schindelin, J.E.: A new metric for probability distributions. IEEE Transactions on Information theory **49**(7), 1858–1860 (2003)

[26] 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)

[27] Givens, C.R., Shortt, R.M.: A class of wasserstein metrics for probability distributions. Michigan Mathematical Journal **31**(2), 231–240 (1984)

[28] 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)

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

[30] 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)

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

[32] 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)

[33] 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)

[34] Pulido, A., *et al.*: Functional materials discovery using energy–structure–function maps. Nature **543**(7647), 657–664 (2017)

[35] 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. In: Computer Vision and Pattern Recogn., pp. 1275–1284 (2023)

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

[37] 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)

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

[39] Macrae, C., *et al.*: Mercury 4.0: From visualization to analysis, design and prediction. Applied Crystallography **53**(1), 226–235 (2020)

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

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

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

[43] Szymanski, N., et al.: An autonomous laboratory for the accelerated synthesis of novel materials. Nature, 86–91 (2023)

[44] 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)

[45] Elkin, Y., Kurlin, V.: A new near-linear time algorithm for k-nearest neighbor search using a compressed cover tree. In: International Conference on Machine Learning (ICML), pp. 9267–9311 (2023)

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

[47] Azrour, M., *et al.*: Rietveld refinements and vibrational spectroscopic studies of $Na_{1-x}K_xPb_4(PO_4)_3$ lacunar apatites ($0 \leq x \leq 1$). Journal of Physics and Chemistry of Solids **72**(11), 1199–1205 (2011)

[48] Cerqueira, T., *et al.*: Identification of novel Cu, Ag, and Au ternary oxides from global structural prediction. Chemistry of Materials **27**(13), 4562–4573 (2015)

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

[50] Quarez, E., Abraham, F., Mentré, O.: Synthesis, crystal structure and characterization of new 12h hexagonal perovskite-related oxides $Ba_6M_2Na_2X_2O_{17}$ (m=ru, nb, ta, sb; x=v, cr, mn, p, as). J Solid State Chemistry **176**(1), 137–150 (2003)

[51] Julien, P., *et al.*: Structure cristalline du grenat $Gd_{3-x}Ca_xGa_{5-x}Zr_xO_{12}$. Comptes rendus de l'Académie des sciences. Série 2 **306**, 531–535 (1988)

[52] Selker, P., Klaska, R.: Struktur und hydrothermalsynthesen von beryllonittypen im system $(Na_K)OH{-}Al(OH)_3{-}Ga_2O_3{-}SiO_2{-}GeO_2$. Zeitschrift für Krist. **159**, 119–120 (1982)

[53] Crennell, S., Owen, J., Grey, C., Cheetham, A., Kaduk, J., Jarman, R.: 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)

[54] Yakubovich, O., Anan'eva, E., Dimitrova, O.: Crystal structure of the solid solution $Mn_2(P_xV_{1-x})(V_yP_{1-y})O_7$. Koord. Khimiya **26**(8), 586–591 (2000)

[55] 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)

[56] Li, C., Zhong, J.: Highly efficient broadband near-infrared luminescence with zero-thermal-quenching in garnet $Y_3In_2Ga_3O_{12}$: Cr3+ phosphors. Chemistry of Materials **34**(18), 8418–8426 (2022)

[57] Widdowson, D., Kurlin, V.: Continuous invariant-based maps of the cambridge structural database. Crystal Growth and Design **24**, 5627–5636 (2024)

[58] Ropers, J., Mosca, M., Anosova, O., Kurlin, V., Cooper, A.: Fast predictions of lattice energies by continuous isometry invariants of crystal structures. In: Data Analytics and Management in Data Intensive Domains, pp. 178–192 (2022)

[59] Balasingham, J., Zamaraev, V., Kurlin, V.: Material property prediction using graphs based on generically complete isometry invariants. Integrating Materials and Manufacturing Innovation **13**, 555–568 (2024)

[60] Balasingham, J., Zamaraev, V., Kurlin, V.: Accelerating material property prediction using generically complete invariants. Scientific Rep. **14**, 10132 (2024)

[61] Schwalbe-Koda, D., Widdowson, D., Pham, T.A., Kurlin, V.: Inorganic synthesis-structure maps in zeolites with machine learning and crystallographic distances. Digital Discovery **2**, 1911–1924 (2023)

[62] Anosova, O., Gorelov, A., Jeffcott, W., Jiang, Z., Kurlin, V.: A complete and bi-continuous invariant of protein backbones under rigid motion. MATCH Comm. Math. Comp. Chemistry **94**, 97–134 (2025)

[63] Wlodawer, A., Dauter, Z., Rubach, P., Minor, W., Jaskolski, M., Jeffcott, W., Jiang, Z., Anosova, O., Kurlin, V.: Duplicate entries in the protein data bank: how to detect and handle them. Acta Crystallographica Section D **81** (2025)

[64] Widdowson, D.: Crystal Invariants in Python. https://github.com/dwiddo/AMD

[65] 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 Extra examples and navigation maps

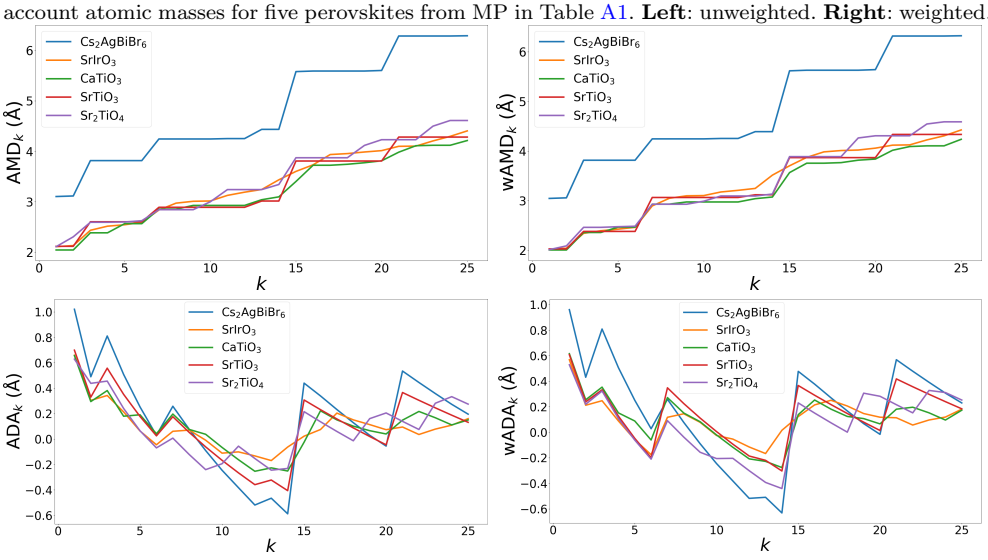
This appendix includes extra examples of invariant computations for 5 perovskites in addition to 5 simple crystals in Fig. 3, see corresponding entries from the MP in Table A1, instructions for running the Python code for all invariants, and high-resolution navigation maps. The zip folder with supplementary information includes the Python code and tables with PDD and PDA matrices for the $5 + 5$ example crystals.

Fig. A1 and Table A2 compare unweighted (based on atomic centers) and weighted versions (including atomic masses) of invariants. The atomic masses generally increase EMD distances but the geometry of atomic centers already captures chemistry.

simple crystals	rock salt	rutile	zincblende	fluorite	antifluorite
composition	NaCl	TiO_2	ZnS	CaF_2	Mg_2Si
MP entry id	22862	2657	10695	2741	1367
perovskites	cubic	hexagonal	tetragonal	double	Ruddlesden-Popper
composition	$SrTiO_3$	$SrIrO_3$	$CaTiO_3$	$Cs_2AgBiBr_6$	Sr_2TiO_4
MP entry id	5229	17097	3442	1078250	5532

Table A1 Names, compositions and IDs of 5+5 crystals whose invariants are in Fig. 3 and A1.

Fig. A1 The invariants AMD_k and ADA_k from Definition 4 and their weighted versions taking into account atomic masses for five perovskites from MP in Table A1. **Left**: unweighted. **Right**: weighted.



The instructions below will reproduce the distances between the structures in Fig. 5, whose CIFs are publicly available at these URLs:

- icsd_670065.cif: https://www.ccdc.cam.ac.uk/structures/Search?Ccdcid=670065&DatabaseToSearch=Published
- icsd_139006.cif: https://www.ccdc.cam.ac.uk/structures/Search?Ccdcid=139006&DatabaseToSearch=Published
- MnAgO2.cif: Under `Supplementary-Data\Structure_Files` in the supplementary data at https://www.nature.com/articles/s41586-023-06734-w#Sec14

The supplementary materials also include the script `compare_MnAgO2.py` for the distances. With Python 3.9+, install average-minimum-distance (1.5.3) [64] with

```
pip install average-minimum-distance==1.5.3
```

Run the script with

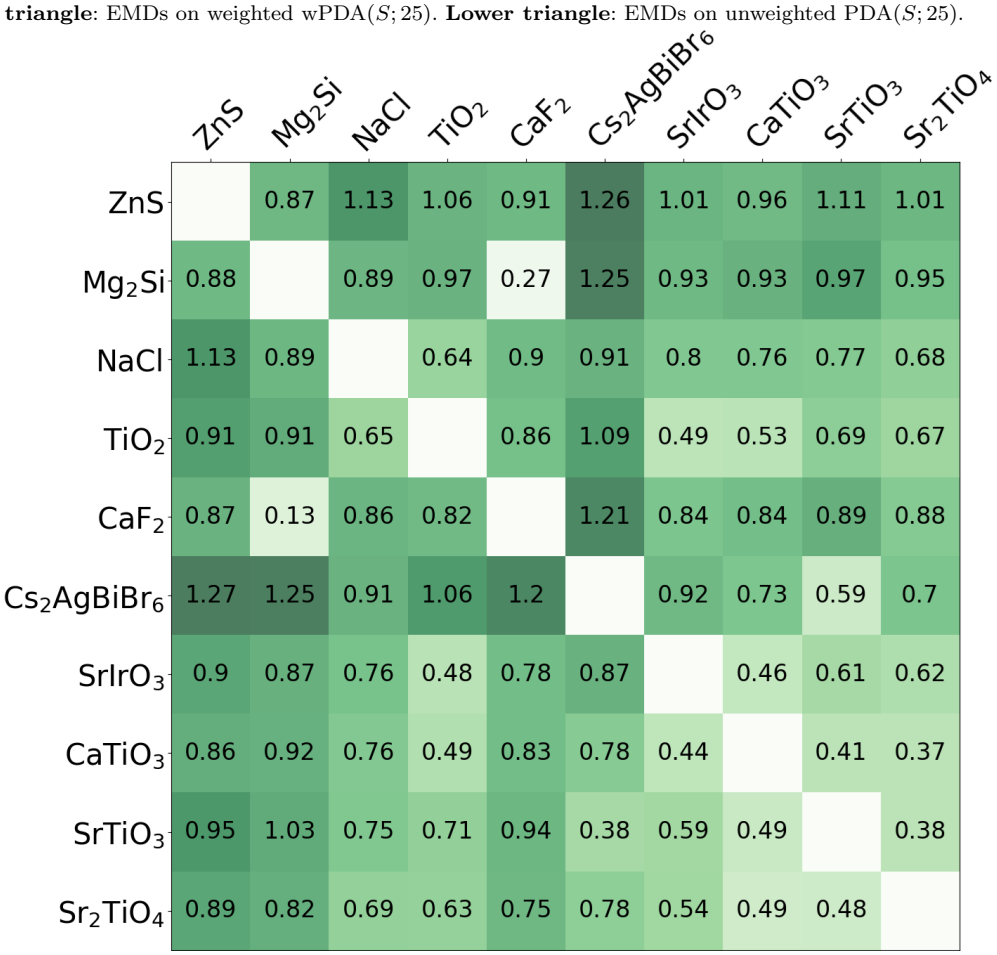
```
python compare_MnAgO2.py
```

The expected output is:

```
>> EMD(PDA(MnAgO2, 100), PDA(icsd_670065, 100)) = 0.0975A
>> EMD(PDA(MnAgO2, 100), PDA(icsd_139006, 100)) = 0.3675A
```

The GNoME paper used the Pymatgen structure matcher [38], which cannot filter out near-duplicate structures according to the quoted steps 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 to filter out existing duplicates from materials databases, resulting in the AI silently reproducing near-duplicates from the training set.

Fig. A2 Distances in Angstroms between 5 simple crystals and 5 perovskites from Table A1. **Upper triangle**: EMDs on weighted wPDA(S; 25). **Lower triangle**: EMDs on unweighted PDA(S; 25).



Definition 7 can use arbitrary weights of points so that the total weight within a unit cell is 1 after normalization. When all equal rows of PDD(S; k) are collapsed to a single row, the discontinuity in Fig. 1 is properly resolved. The unweighted version of EMD on PDAs below uses zero-sized points at atomic centers with equal weights, which can be multiplied (before normalization) by atomic masses or ionic radii.

Definition 7 (Earth Mover's Distance EMD [26])**.** Consider any matrix PDA(S; k) as a distribution of rows $R_i(S)$ with weights $w_i(S)$ for $i = 1, \dots, m(S)$ such that $\sum_{i=1}^{m} w_i = 1$. The *Earth Mover's Distance* $\text{EMD}(\text{PDA}(S;k), \text{PDA}(Q;k)) = \min_{f_{ij}} \sum_{i=1}^{m(S)} \sum_{j=1}^{m(Q)} f_{ij} L_\infty(R_i(S), R_j(Q))$ is minimized for all real $f_{ij} \geq 0$ (called *flows*) subject to the conditions $\sum_{i=1}^{m(S)} f_{ij} \leq w_j(Q)$, $\sum_{j=1}^{m(S)} f_{ij} \leq w_i(S)$, $\sum_{i=1}^{m(S)} \sum_{j=1}^{m(Q)} f_{ij} = 1$.

The first condition $\sum_{j=1}^{m(Q)} f_{ij} \leq w_i(S)$ 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 $\sum_{i=1}^{m(S)} f_{ij} = w_j(Q)$ 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 $\sum_{i=1}^{m(S)} \sum_{j=1}^{m(Q)} f_{ij} = 1$ forces to 'flow' all rows $R_i(S)$ to all rows $R_j(Q)$.

Proof of Theorem 6. 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 [65, 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 5.

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. □

Fig. A3 A-lab crystals in cyan over the ICSD and MP heatmap in the coordinates (density, ADA_1).

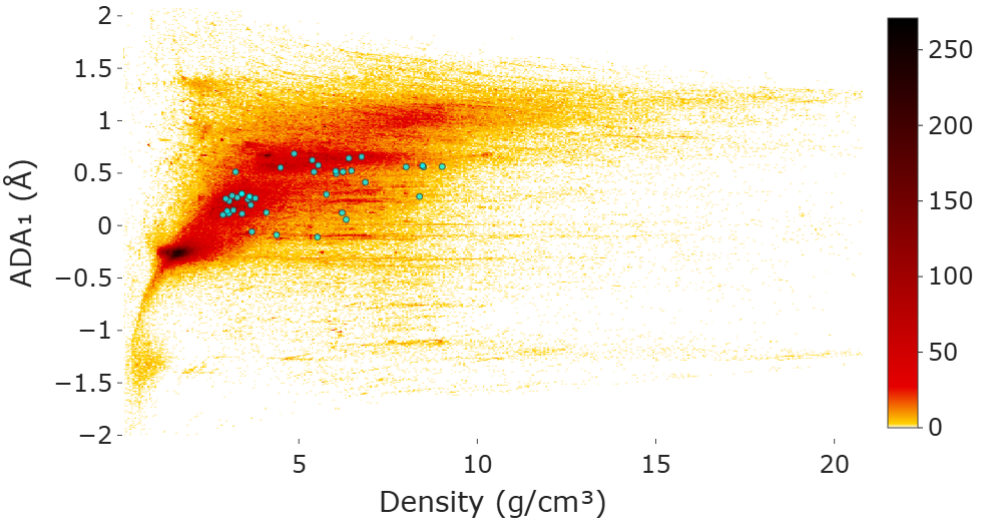


Fig. A4 A-lab crystals in cyan over the ICSD and MP heatmap in the coordinates (PPC, ADA_1).

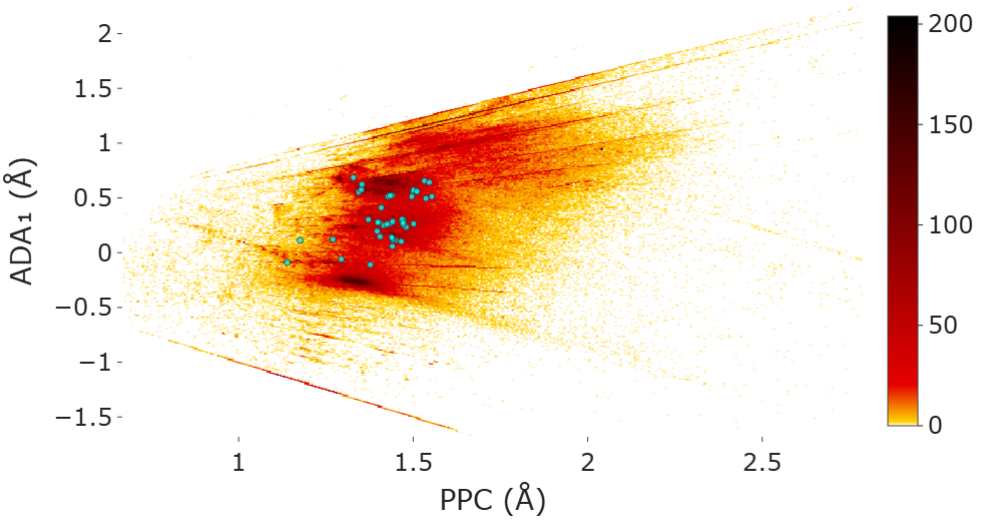


Fig. A5 A-lab crystals in cyan over the ICSD and MP heatmap in the coordinates (ADA_2, ADA_3).

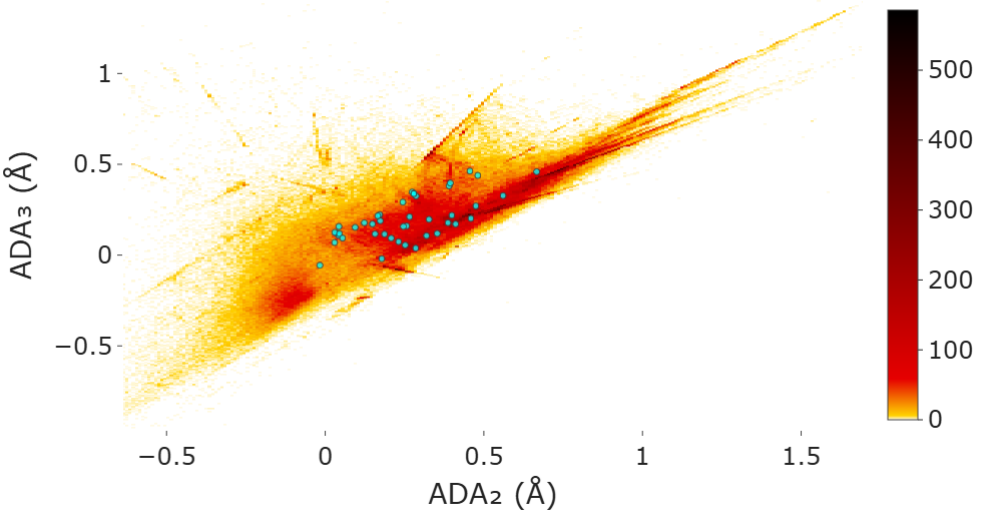


Fig. A6 A-lab crystals in cyan over the ICSD and MP heatmap in the coordinates (ADA_4, ADA_5).

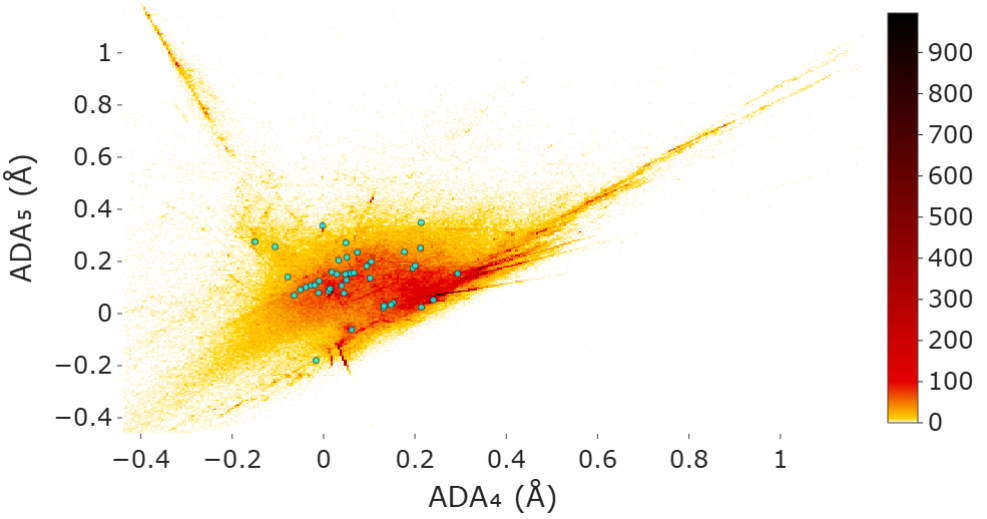


Fig. A7 A-lab crystals in cyan over the ICSD and MP heatmap in the coordinates (ADA_1, ADA_{20}).

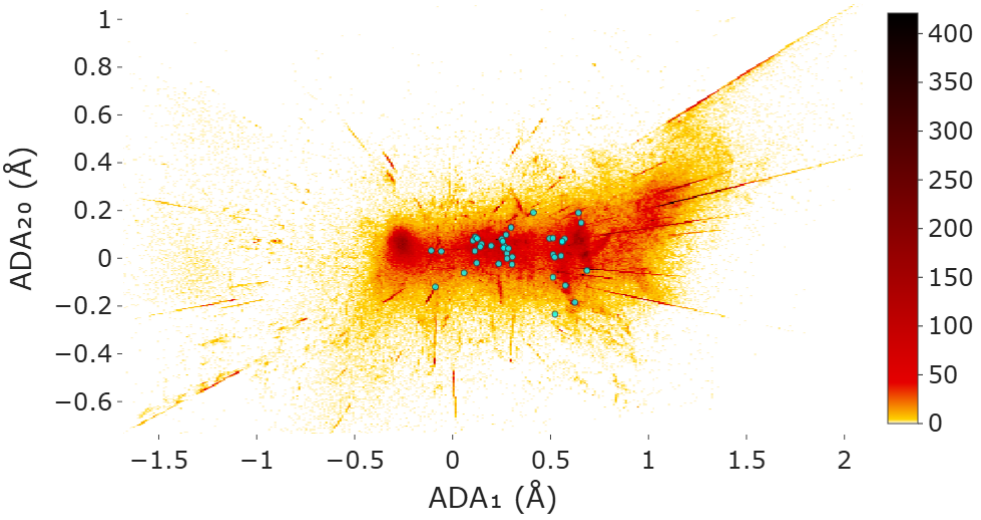


Fig. A8 A-lab crystals over the ICSD and MP heatmap in the coordinates (ADA_{20}, ADA_{100}).