

Necessary conditions for benchmarking next-generation AI in materials discovery

Daniel E Widdowson¹, Tatiana V Kurlina², Olga D Anosova¹,
Vitaliy A Kurlin^{1,3}

¹School of Computer Science and Informatics, University of Liverpool,
Ashton Street, Liverpool, L69 3DR, Merseyside, UK.

²Department of Mathematics, University College London, Gower Street,
London, WC1E 6BT, UK.

^{3*}Materials Innovation Factory, University of Liverpool, 51 Oxford
Street, Liverpool, L69 7ZD, Merseyside, UK.

Contributing authors: ; ; ; vitaliy.kurlin@liverpool.ac.uk;

Abstract

Current AI predicts materials by speeding up energy minimisation and training on large databases. However, scaling data generation prevents manual analysis by domain experts. If novelty is evaluated by chemical compositions or traditional representations in Crystallographic Information Files, or standard structure matchers based on manually chosen thresholds, one can easily generate millions of “new” compositions or files without discovering truly novel materials.

This paper uses a hierarchy of crystal invariants to continuously trace and quantify structural novelty based on the new mathematical definition of a crystal structure and asymptotic guarantees for invariant descriptors. The invariant-based analysis through cascade computations on a modest computer exposes the unexpected scale of duplication in the world’s largest materials databases, shown here in the high-profile case of GNoME datasets. To prevent current “embarrassment of over-prediction”, we propose necessary conditions for next-generation AI that should be automatically enforced before claiming any new materials.

Keywords: AI, materials discovery, periodic crystal, equivalence relation, rigid motion, isometry classification, complete invariant, continuous distance metric

*Freedom is the freedom to say that two plus two make four.
If that is granted, all else follows. (George Orwell, “1984”).*

1 Introduction to challenges of materials discovery

In 1988, John Maddox’s editorial in *Nature* [1] acknowledged that “one of the continuing scandals in the physical sciences is that it remains in general impossible to predict the structure of even the simplest crystalline solids from a knowledge of their chemical composition”. This paper will discuss why Crystal Structure Prediction remained “generally impossible” before the recent developments in Geometric Data Science [2].

The first reason is that chemical composition may not guarantee a unique structure. For example, diamond and graphite consist of pure carbon but have vastly different properties. The pharmaceutical industry is acutely aware of so-called “late-appearing” polymorphs whose crystal structure differs from an original prediction and might be useless for patients [3]. The broader problem is to predict, for a given composition, *all possible synthesisable materials* that are stable under application conditions.

In 2024, A. Cheetham and R. Seshadri [4] suggested that new compounds should “fulfill the trifecta of novelty, credibility, and utility”. L. Schoop and R. Palgrave [5] showed that materials discovery remains challenging because many synthesised disordered crystals are similar to known ordered structures in existing databases. Crystallographers debated different definitions of a crystal for dozens of years and concluded in 2021 that “a real-space definition mentioning periodicity is needed” [6].

The current definition [7] of crystals being the “same” states that “crystals are said to be isostructural if they have the same structure, but not necessarily the same cell dimensions nor the same chemical composition, and with a ‘comparable’ variability in the atomic coordinates to that of the cell dimensions and chemical composition”.

To understand what was missing in the past, recall that a *periodic crystal* is traditionally described in a Crystallographic Information File (CIF) by a linear basis of three vectors in $\mathbb{R}^3$, usually by their lengths and pairwise angles, and a *motif* of atoms with fractional coordinates in this basis. Then a periodic crystal is obtained from a motif by periodic translations along all linear combinations of three basis vectors.

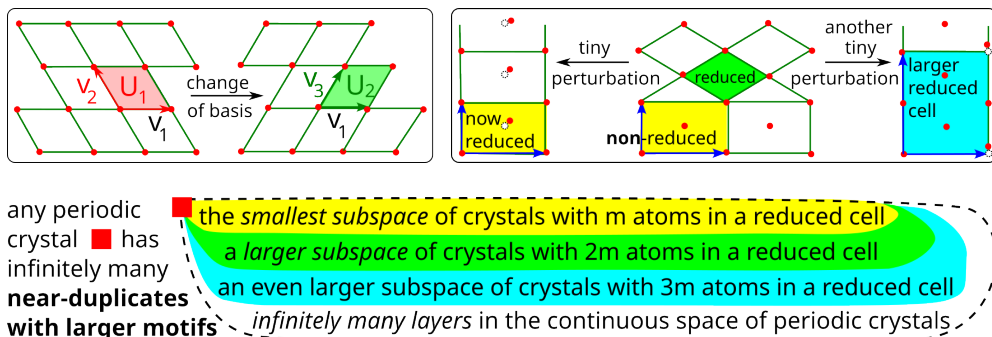


Fig. 1 **Top left:** a hexagonal lattice and any periodic crystal can be generated by many different cells. **Top right:** a reduced cell can discontinuously change under tiny trivial perturbations. **Bottom:** any periodic crystal belongs to infinitely many subspaces of crystals with motifs of multiple sizes.

Any basis generates the *unit cell*, but any periodic crystal can be represented by infinitely many pairs (unit cell, motif). Though a unit cell can be reduced to a canonical form, e.g. Niggli’s cell [8], it has been experimentally known since 1965 that Niggli’s cell discontinuously changes under noise [9, p. 80], see Fig. 1 (top right).

The “same structure” remained undefined in [7] until a *crystal structure* was introduced in [10, Definition 6] as a class of all crystals that can be exactly superimposed with all atoms by *rigid motion*, which is a composition of translations and rotations.

This definition does not allow any ‘comparable’ variability of crystal parameters. Indeed, if we allow all perturbations up to any small threshold $\varepsilon > 0$, sufficiently many trivial perturbations can make all crystals equivalent due to the transitivity axiom: if a crystal S is equivalent to a crystal Q , and Q is equivalent to a crystal T , then S is equivalent to T , briefly written as follows: if $S \sim Q \sim T$ then $S \sim T$.

The ambiguity under noise, illustrated in Fig. 1 (top right), has been known since ancient times as the *sorites paradox*: “When does a heap of sand stop being a heap while grains of sand are removed one by one?” [11]. Since atoms have real coordinates and are not restricted to a discrete grid, atomic positions are like “grains of sand” that can be slightly shifted again and again until a crystal becomes completely different. Using tolerances for a binary outcome (true/false) of matching only shifts the comparison from 0 to another threshold $\varepsilon > 0$ without resolving the sorites paradox.

For continuous classifications of real data, we can distinguish any tiny differences by a continuous distance metric d so that if periodic crystals S, Q are related by a small perturbation moving each atom up to a Euclidean distance ε , then their distance $d(S, Q)$ is small relative to ε . If there is a constant λ such that $d(S, Q) \leq \lambda\varepsilon$ for all perturbations up to a small $\varepsilon > 0$, this condition is called *Lipschitz continuity*.

The continuous nature of our world necessarily requires us to distinguish between all (infinitely many) rigid structures by continuous metrics. How is it really possible?

Since crystal structures are determined in a rigid form without a preferred coordinate system, there is no sense in distinguishing between their different representations $S \neq Q$ that are *rigidly equivalent*, i.e. S can be exactly matched with Q by rigid motion, denoted as $S \cong Q$. This rigid equivalence is an example of an *equivalence relation* (denoted as $\sim$) that should satisfy three basic axioms: (1) *reflexivity*: $S \sim S$, (2) *symmetry*: if $S \sim Q$ then $Q \sim S$, and (3) *transitivity*: if $S \sim Q \sim T$, then $S \sim T$. The last axiom is needed for any justified classification as a splitting into disjoint classes.

Any periodic crystal S generates its *crystal structure* in [10, Definition 6] as the equivalence class $[S] = \{\text{all crystals } Q \cong S\}$ consisting of all (infinitely many) crystals Q that are rigidly equivalent to a single representative S in a CIF. The resulting space of crystal structures is continuous because any perturbation of a crystal S almost always produces a slightly different structure $[Q] \neq [S]$, as in Fig. 1 (top right).

The rigid equivalence can be extended by adding compositions with mirror reflections (*isometry*) and uniform scaling (*dilation*), which makes the resulting spaces smaller by not distinguishing between mirror images, but keeps them continuous. If we view atoms as zero-sized points without chemical types, all simple cubic crystals of different properties, such as rock salt NaCl or KBr, become equivalent under dilation.

Hence, if we aim to distinguish between all functionally different (periodic) materials, we should study them in the largest continuous space, without splitting it into disjoint pieces by symmetries or chemical compositions. This huge space of all theoretical periodic structures opened the problem of finding a mathematical code or a *materials genome* [12] that should identify any material with all its properties [13].

2 Towards a materials genome of periodic crystals

A practical approach to resolve the *curse of dimensionality* in the continuous crystal space is to quickly narrow the search in a hierarchical way until a unique material is reliably identified. While many real objects, including molecules and crystals, are still compared by aligning their geometric representations in $\mathbb{R}^3$, Emmy Noether [14] initiated a more general and efficient invariant-based approach as early as 1918 [15].

For a fixed equivalence relation, such as rigid motion, an *invariant* I is a function that takes the same value on all equivalent objects. In other words, if crystals $S \cong Q$ are rigidly equivalent, a rigid invariant should satisfy $I(S) = I(Q)$. A simple invariant is the minimal volume $\text{vol}(S)$ of a (primitive) unit cell of a crystal S . Instead of computationally expensive alignment, we can conclude that crystals S, Q are not rigidly equivalent if any (single) invariant I takes different values on them: $I(S) \neq I(Q)$.

If an invariant I satisfies $I(S) = I(Q)$, we cannot conclude that $S \cong Q$ in general. If this inverse property always holds (if $I(S) = I(Q)$ then $S \cong Q$), the invariant I is called *complete*. The standard settings of a periodic crystal with a reduced cell form a complete invariant in theory [16]. In practice, any reduced cell discontinuously changes under almost any noise, as in Fig. 1 (top right). The practical requirement of continuity for periodic structures was introduced in 2020 [17] and led to a hierarchy of continuous invariants from computationally fast vectors [18] to complete ones [19].

For a periodic crystal $S \subset \mathbb{R}^3$ and any integer $k \geq 1$, our first invariant $\text{AMD}_k(S)$, called the *average minimum distance*, is the average distance to the k -th atomic neighbour in S , see Fig. 2. The sequence of $\text{AMD}_k(S)$ is computed up to a fixed number k of atomic neighbours within the full crystal S , without restricting to any unit cell.

More traditional atomic environments up to a fixed cut-off radius r can contain different numbers of atoms and change discontinuously when atoms are perturbed across a hard boundary. While the order of equidistant neighbours can change under noise, their ordered distances to a central atom are Lipschitz continuous in the sense that perturbing any points p, q up to ε changes their distance $|p - q|$ up to 2ε .

The last conclusion follows for any distance *metric* d that satisfies three axioms for all objects under a given equivalence: (1) $d(S, Q) = 0$ if and only if $S \sim Q$, (2) *symmetry*: $d(S, Q) = d(Q, S)$, and (3) *triangle inequality* $d(S, Q) + d(Q, T) \geq d(S, T)$. If the triangle inequality fails, outputs of the clustering algorithms k -means and DBSCAN may not be trustworthy [20]. The *only if* condition in the first axiom guarantees completeness because any non-equivalent crystals $S \not\sim Q$ have $d(S, Q) > 0$.

In all our invariants, the number k of atomic neighbors is not treated as a parameter, but as a degree of resolution, similar to the number of decimal places on a calculator. Indeed, as k increases, the sequence of $\text{AMD}_k(S)$ expands by considering more distant atomic neighbors while all previous invariants remain unchanged.

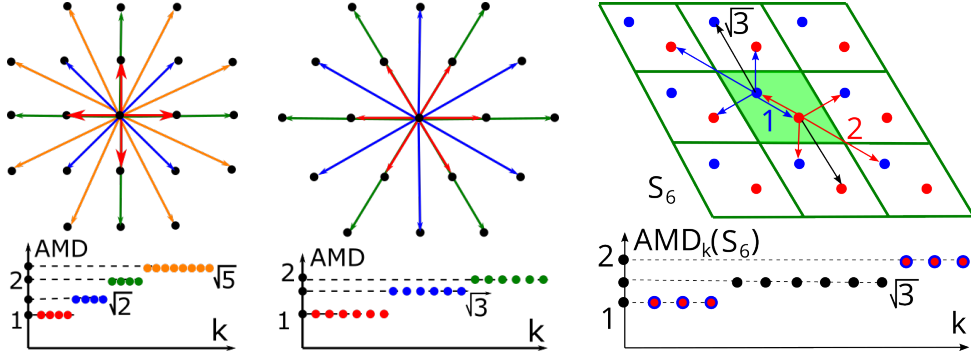


Fig. 2 **Left:** in the square lattice, the k -th neighbours of the origin and corresponding AMD_k are shown in the same colour. For example, the shortest axis-aligned distances $\text{AMD}_1 = \dots = \text{AMD}_4 = 1$ are in red, the longer diagonal distances $\text{AMD}_5 = \dots = \text{AMD}_8 = \sqrt{2}$ are in blue. **Middle:** in the hexagonal lattice, the shortest distances are in red: $\text{AMD}_1 = \dots = \text{AMD}_6 = 1$. **Right:** a honeycomb periodic point set has a 2-point motif and the first average distances $\text{AMD}_1 = \text{AMD}_2 = \text{AMD}_3 = 1$.

The key advantage of the invariant $\text{AMD}(S; k) = (\text{AMD}_1(S), \dots, \text{AMD}_k(S))$ and its stronger extensions below is Lipschitz continuity: if periodic crystals S, Q are related by shifting every atom up to ε , the Root Mean Square distance has the upper bound

$$\text{RMS}(\text{AMD}(S; k), \text{AMD}(Q; k)) = \sqrt{\frac{1}{k} \sum_{i=1}^k (\text{AMD}_i(S) - \text{AMD}_i(Q))^2} \leq 2\varepsilon \text{ for } k \geq 1.$$

The upper bound 2ε explains the physical meaning of RMS, which is measured in the same units as point coordinates: if $\text{RMS}(\text{AMD}(S; k), \text{AMD}(Q; k)) = d$, to exactly match S, Q by isometry, we should perturb the first k inter-atomic distances by d on average. Averaging over all the atoms in a motif produces an invariant that is independent of point ordering but theoretically incomplete. Instead of averaging, the Pointwise Distance Distribution (PDD), for each atom p in a motif, keeps the first k distances to its neighbours in a separate list [21], so the full invariant $\text{PDD}(S; k)$ is an unordered distribution of these equally weighted lists, see details in Definition A.10.

For a crystal S with a motif of m atoms, one can initially write $\text{PDD}(S; k)$ as a matrix with k ordered columns and m unordered rows. The honeycomb structure S_6 in Fig. 2 (right) has $m = 2$ points that can be matched (with the full S_6) by a global rotation. As a matrix, $\text{PDD}(S_6; k)$ for $k \geq 4$ has two identical rows starting with distances $1, 1, 1, \sqrt{3}, \dots$. It is convenient to collapse all identical (say, l) rows to a single one with the weight $\frac{l}{m}$, so $\text{PDD}(S_6; k)$ becomes a single vector of weight 1.

Interpreting $\text{PDD}(S; k)$ as a probability distribution of vectors in $\mathbb{R}^k$, we often apply the Earth Mover's Distance (EMD) from transportation theory [22], which optimises the cost of transforming distance lists in two PDDs by using RMS on $\mathbb{R}^k$.

Due to the Lipschitz continuity, if $\text{EMD}(\text{PDD}(S; k), \text{PDD}(Q; k)) = d$, to match the crystals S, Q by isometry, atoms of S should be mapped to atoms of Q so that the first k inter-atomic distances are perturbed by d overall [23, Theorem 4.2]. In other words, S and Q cannot be exactly matched if we shift every atom by less than $\frac{d}{2}$.

More significantly, $\text{PDD}(S; k)$ with a large enough k and a lattice of S suffices to reconstruct almost any periodic crystal $S \subset \mathbb{R}^3$, uniquely under isometry, see [23, Theorem 5.5]. Although this *generic completeness* does not guarantee completeness in theory, we have not found any non-isometric structures (denoted as $S \not\cong Q$) with $\text{PDD}(S; 100) = \text{PDD}(Q; 100)$ among all experimental crystals in modern databases.

In 1930, Linus Pauling [24] found non-isometric crystals that had identical diffraction patterns and were later called *homometric* [25]. Second-order PDD invariants distinguished the Pauling crystals and all other theoretically known periodic sets $S \not\cong Q$ with $\text{PDD}(S; k) = \text{PDD}(Q; k)$ for $k \geq 1$ in [26, Examples 3.4 and 4.2].

For a periodic crystal $S \subset \mathbb{R}^3$ with m atoms in a unit cell of volume $\text{vol}[U]$, the *Point Packing Coefficient* $\text{PPC}(S) = \sqrt[3]{\frac{\text{vol}[U]}{mV_3}}$ measures the (scaled) volume per atom, where $V = \frac{4}{3}\pi$ is the unit ball volume in $\mathbb{R}^3$. [18, Theorem 13] proved that the distance to the k -th atomic neighbour in S is asymptotically increasing as $\text{PPC}(S) \sqrt[3]{k}$ as $k \rightarrow +\infty$. Hence, if periodic crystals S, Q have $\text{PPC}(S) \neq \text{PPC}(Q)$, the invariants $\text{AMD}_k(S)$ and $\text{AMD}_k(Q)$ eventually diverge, so their RMS is dominated by contributions of the most distant neighbours. However, the most interesting differences appear for small k , such as the average bond length in molecular crystals for $k = 1$.

To minimise this effect, we adjusted $\text{AMD}_j(S)$ and all distances in the j -th column of $\text{PDD}(S; k)$ for $j \leq k$ by subtracting $\text{PPC}(S) \sqrt[3]{j}$. The resulting invariants $\text{ADA}(S; k)$ and $\text{PDA}(S; k)$ are called *average* and *pointwise deviations from asymptotic*, respectively. For molecular crystals, ADA_k behaves like waves fading to 0 in Fig. 3, which justifies why smaller values of k better separate crystals than larger values. For inorganic crystals, these waves have sharper “zigzags” in [27, Fig. 3].

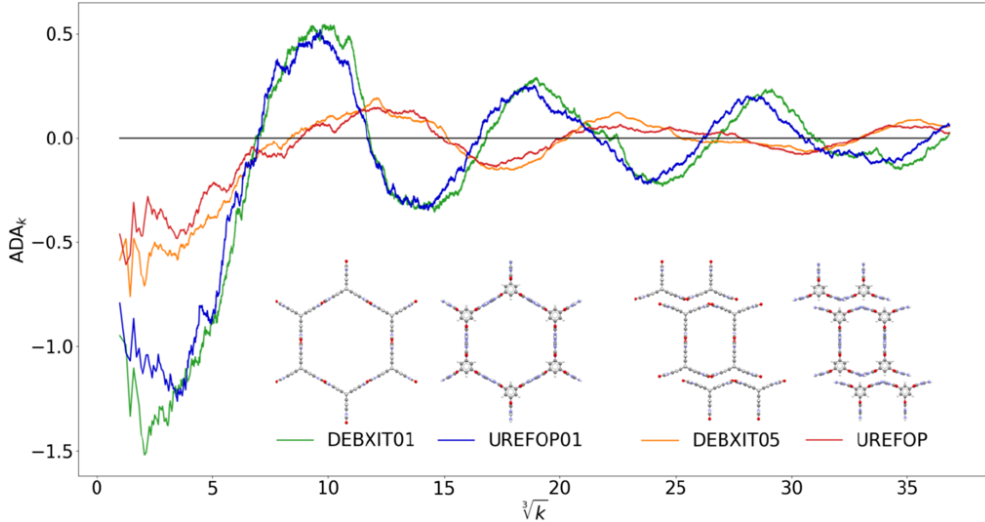


Fig. 3 The invariants ADA_k reveal the similarity between some crystals in the Cambridge Structural Database (CSD), though the molecules in crystals DEBXIT and UREFOP differ in size and chemistry.

All distance-based invariants do not distinguish mirror images, which motivated us to develop the stronger invariant (called *isoset*) whose completeness was proven under isometry [28] and then adapted to distinguish between mirror images under rigid motion [19]. All invariants and metrics above are computable in polynomial time in the number k of neighbours and the motif size m , for a fixed dimension n .

A limitation is an approximate algorithm for a metric d between complete isosets. This algorithm has an approximation factor of about 4 in $\mathbb{R}^3$ [19, Corollary 5.10] and guarantees that if the approximation of d is 0, then so is d , and hence the underlying crystals are rigidly equivalent. Similarly, if an approximation of d is small, then d can be at most 4 times larger, which suffices for detecting near-duplicates in practice.

Fig. 4 summarises progress from discrete to continuous classifications, including generically complete density functions [29], see historical details in [2, section 12.4].

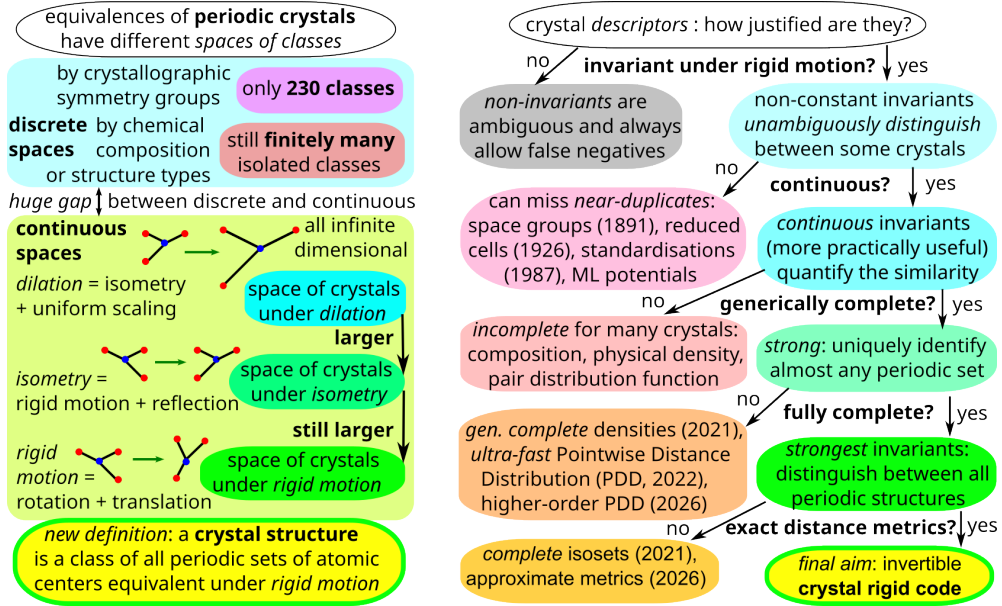


Fig. 4 **Left:** periodic crystals were previously studied by discrete equivalences but are experimentally determined as rigid structures that live in large *continuous spaces*. **Right:** the final aim of developing structural invariants is a *crystal rigid code* that is theoretically complete, continuous under tiny perturbations, and invertible back to a periodic crystal structure $S \subset \mathbb{R}^3$, uniquely under rigid motion.

Algorithm 2.1 describes how a crystal structure S given by a CIF is uniquely identified by a hierarchy of invariants. This identification can involve finding the nearest neighbours of S in a given dataset D , such as the CSD [30], or several datasets. Stages 1-4 in Algorithm 2.1 are justified by [23, Theorem 4.4] and [19, Theorem 6.5]: a metric between weaker invariants is a lower bound for a metric between stronger ones.

The invariant $\text{AMD}(S; k)$ practically distinguishes all non-duplicate crystal structures in the CSD already for $k = 100$ [18]. In other words, if any periodic crystals

Algorithm 2.1 Nearest neighbours of a periodic crystal S in a known dataset D

Stage 1. $\text{ADA}(S; k)$ and $\text{PDA}(S; k)$ are computed in near-linear time of k, m [23, Theorem 3.10]. If the vectors $\text{ADA}(Q; k)$ are pre-computed for all periodic crystals $Q \in D$, they can be organised in a k-d tree [31], which speeds up the search for a fraction (say, 1%) of the nearest neighbours to $\text{ADA}(S; k)$ by the metric RMS.

Stage 2. For the nearest neighbours found by RMS on $\text{ADA}(S; k)$, we compute the EMD on $\text{PDA}(S; k)$. The EMD is exactly computed in time $O(m^3 \log m)$, where m is the maximum size PDA distributions [32], and is approximated in linear time [33].

Stage 3. For a fraction of the nearest neighbours found by EMD on $\text{PDA}(S; k)$, we compute the EMD adapted to higher-order $\text{PDA}^{\{h\}}(S; k)$, usually only for $h = 2$, to distinguish between potential homometric crystals that have equal PDAs.

Stage 4. For a fraction of the nearest neighbours found by EMD on $\text{PDA}^{\{h\}}(S; k)$, we compute the complete isoset, which is needed only in rare cases to distinguish between mirror images; there are four such pairs in the whole CSD, see [19, Fig. 7 (right)].

S, Q in the CSD have close $\text{AMD}(S; 100)$ and $\text{AMD}(Q; 100)$, for example, by RMS up to 10^{-10}Å , then they can be matched by isometry with a similar small perturbation. Here, one Angstrom ($1 \text{Å} = 10^{-10} \text{ meter}$) is roughly the smallest inter-atomic distance.

Stages 1-3 sufficed to find the nearest neighbours of all experimental crystals from [34, 35] in the Inorganic Crystal Structure Database (ICSD) and Materials Project (MP), including those that were missed in [5], see [27, Table 2] and [23, Tables 1-4].

After a news story [36] in December 2025 highlighted our research since 2021, the A-lab paper [34] was corrected in January 2026 by crossing out almost all the words “novelty” and “discovery”, and acknowledging that “the original claims of material novelty were subject to misinterpretation - their intention was to indicate that the materials were new to the prediction platform, not necessarily new to science” [37].

The next section shows that our invariants scale up to all-vs-all comparisons in the world’s largest databases, using Google’s GNoME as a high-profile example [38].

3 A mathematical “anatomy” of the GNoME data

This section applies the mathematical tools of section 2 to the GNoME data from [39]. We focus on the first version (called here GNoME1, dated 29 November 2023), which was published on the same day as the Nature paper [38], and the last version (called here GNoME2, dated 21 November 2024), see details of all versions in appendix B.

The main authors of [38], Amil Merchant and Ekin Dogus Cubuk, who also co-authored the A-lab paper [34], claimed “the discovery of 2.2 million new crystals – equivalent to nearly 800 years’ worth of knowledge”, which is quoted from [40].

While 2.2 million “new crystals” were not publicly shared, GNoME1 has 384,938 “most stable” crystals. GNoME2 has 554,054 entries, of which 301,413 are shared with GNoME1. Review [4, p. 3492] mentioned that up to 42 thousands entries in GNoME1 contain radioactive elements Ac, Np, Pa, Pm, Pu, and Tc, but missed Bismuth Bi (weakly radioactive), Thorium Th, and even Uranium U, which we count in Table 1.

number of entries with	1 radioactive element	2 elements	3 elements	total	rate
GNoME1	75,628	4,502	105	80,235	20.8%
GNoME2	102,217	5,664	139	108,020	19.5%

Table 1 Entries in GNoMEs 1&2 with radioactive elements. Full tables can be shared by request.

3.1 Comparing predictions in GNoMEs 1&2 as textual CIFs

For GNoME1, [10, Table 1] had one quadruple (with IDs 4135ff7bc7, 6370e8cf86, c6afea2d8e, e1ea534c2c), 43 triples, and 1,089 pairs of CIFs that are identical texts symbol-by-symbol, which gives $1 \times 4 + 3 \times 43 + 2 \times 1,089 = 2,311$ entries in total, though [38] claimed to “take the top 100 structures for any given composition and perform pairwise comparisons with pymatgen’s built-in structure matcher”.

When comparing only numerical data without chemical elements, the total number of duplicates rises to 3,248 entries, including a group of 9 CIFs in [10, Table 2], where some atoms were replaced without changing their positions. These replacements affect atomic interactions and should perturb the geometry at least slightly, so keeping all numbers in CIFs to six decimal places in Angstroms seems physically unrealistic.

group size: same CIF geometry	no rounding	rounding to 4 digits	rounding to 3 digits	rounding to 2 digits	rounding to 1 digit
10 or more	0	0	0	0	97
9	1	1	1	1	37
8	0	0	0	0	56
7	1	1	1	1	107
6	2	2	2	3	155
5	2	3	3	9	304
4	8	12	13	21	743
3	72	100	113	181	2,447
2	1,481	1,994	2,141	2,633	14,274
all entries	3,248	4,379	4,716	5,972	44,238
percentage	0.84%	1.14%	1.23%	1.55%	11.49%

Table 2 Groups of near-duplicate CIFs in **GNoME1** are obtained by rounding cell parameters (3 lengths and 3 side diagonals) and Euclidean coordinates of atoms in given CIFs, all in Angstroms.

This intuition was inspired by R. Feynman in [41, Fig.1-7] and stated as the *Crystal Isometry Principle* [18]: under fixed ambient conditions, any real periodic crystal is uniquely determined by a precise enough geometry of only atomic centres without chemical elements, now verified on the experimental databases [2, section 12.4].

Since atoms have fractional coordinates in CIFs, it is fairer to compare atomic positions after converting them into Euclidean coordinates, depending on a given cell basis. For GNoME1, Table 2 reports groups of CIFs with duplicated geometry after rounding Euclidean coordinates of atoms to 0.0001Å, 0.001Å, 0.01Å, and 0.1Å, where even the last tolerance is considered acceptable for noisy experimental data [42].

In columns 2-5 of Table 2, the largest group of 9 CIFs is the same as in [10, Table 2]. Table 4 shows two largest groups with identical geometry in CIFs of GNoME2.

group size: same CIF geometry	no rounding	rounding to 4 digits	rounding to 3 digits	rounding to 2 digits	rounding to 1 digit
8 or more	0	0	0	0	347
7	1	1	1	1	181
6	3	3	3	3	294
5	1	2	2	4	613
4	4	6	9	13	1,287
3	17	55	61	90	3,660
2	247	491	534	858	16,947
all entries	591	1,206	1,322	2,083	60,012
percentage	0.11%	0.22%	0.24%	0.38%	10.83%

Table 3 Groups of near-duplicate CIFs in **GNoME2** with rounded cell parameters (3 lengths and 3 side diagonals) and Euclidean coordinates of atoms in given CIFs, see groups of 7 and 6 in Table 4.

GNoME2 id	composition, group of 7	GNoME2 id	composition, group of 6
21a66e0029	Cd ₆ Os ₁ Ru ₁ Tb ₂ Y ₁₈	02c4cb55a6	Cd ₆ Dy ₁₅ Ru ₂ Tb ₅
28566d9d7c	Cd ₆ Dy ₁ Os ₁ Ru ₁ Y ₁₉	0affe9c149	Cd ₆ Dy ₁₈ Ru ₂ Tb ₂
511b93a2bd	Cd ₆ Os ₁ Ru ₁ Y ₂₀	100cfddef	Cd ₆ Dy ₁₇ Ru ₂ Tb ₃
7f9b693d17	Cd ₆ Os ₁ Ru ₁ Tb ₁ Y ₁₉	877c190805	Cd ₆ Dy ₁₆ Ru ₂ Tb ₄
885c00d261	Dy ₂ Cd ₆ Os ₁ Ru ₁ Y ₁₈	9ce48821cb	Cd ₆ Dy ₂₀ Ru ₂
8dfede557b	Cd ₆ Os ₁ Ru ₁ Tb ₁₅ Y ₅	b9e4b78276	Cd ₆ Dy ₁₉ Ru ₂ Tb ₁
d711c9e131	Cd ₆ Ho ₁ Os ₁ Ru ₁ Y ₁₉		

Table 4 Columns 1-2 and 3-4 show the largest groups of 7 and 6 CIFs, respectively, in **GNoME2** from Table 3 with all identical numbers (6+ decimal places) in their CIFs.

3.2 Comparing GNoMEs 1&2 by a hierarchy of invariants

The notations $\text{ADA}[k]$, $\text{PDA}[k]$ with a number k of atomic neighbours and $\text{isotet}[\alpha]$ with a radius α denote the relevant invariants as functions on all crystals. We will compare all crystals in GNoMEs 1&2 by Algorithm 2.1 using (1) RMS on $\text{ADA}[100]$, then (2) EMD on $\text{PDA}[100]$, then (3) EMD on $\text{PDA}^{\{2\}}[100]$, finally (4) EMD on $\text{isotet}[5]$, where the radius $\alpha = 5\text{\AA}$ suffices to include the nearest atomic neighbours.

For a threshold τ , a pair of periodic crystals is called a *geometric match* if all four metrics between their invariants are within τ . A geometric match is called *geochemical* if the optimal matching of atoms in the computation of EMD on isosets preserves all chemical elements. Matched entries form *groups* that are defined as connected components of a graph whose vertices are crystals, while edges join matched entries.

Tables 5-6 count groups of geometric and geochemical matches in GNoMEs 1&2 for various thresholds τ . The statistics of duplication is summarised by the *excess* counting entries that would have to be deleted to leave one representative per group. The *redundancy* rate is the excess divided by the size of the dataset. Appendix B contains more details of the computed invariants, metrics, and their running times.

In comparisons, we used the entries in the ICSD and MP [43] only up to March 2021 because “all candidates for GNoME are derived from snapshots of databases made in March 2021, including the Materials Project” [38], see details in appendix B.2.

threshold τ , Å	number of groups of geometric matches	groups of size ≥ 2	entries in groups ≥ 2	excess entries	redundancy rate, %	largest group
10^{-10}	383,257	1,567	3,248	1,681	0.437	9
10^{-4}	382,446	2,324	4,816	2,492	0.647	9
10^{-3}	381,908	2,744	5,774	3,030	0.787	9
10^{-2}	337,060	19,610	67,488	47,878	12.438	352
geochemical matches						
10^{-10}	383,760	1,133	2,311	1,178	0.306	4
10^{-4}	383,468	1,417	2,887	1,470	0.382	4
10^{-3}	383,250	1,618	3,306	1,688	0.439	4
10^{-2}	380,488	4,172	8,622	4,450	1.156	12

Table 5 Near-duplicates, excess, and redundancy of the **GNoME1** dataset (384,938 entries).

threshold τ , Å	number of groups of geometric matches	groups of size ≥ 2	entries in groups ≥ 2	excess entries	redundancy rate, %	largest group
10^{-10}	553,736	273	591	318	0.057	7
10^{-4}	553,339	613	1,328	715	0.129	7
10^{-3}	552,905	958	2,107	1,149	0.207	7
10^{-2}	469,962	24,194	108,286	84,092	15.178	723

Table 6 Near-duplicates, excess, and redundancy of the **GNoME2** dataset (554,054 entries).

No pairs of entries of GNoME2 were a geochemical match for thresholds up to 0.01Å.

The top 4-row block of Table 7 compares GNoMEs 1&2 with the ICSD and MP using the EMD on invariants PDA[100], which distinguished all non-isometric crystals in the CSD and ICSD. The bottom 4-row block of Table 7 reports the Root Mean Square Deviation (rmsd_max) by StructureMatcher with default parameters [44]. This RMSD measures the deviation between all atoms in suitably extended and optimally aligned unit cells from two crystals, including non-rigid averaging of lattices in [45, step 10.b.i.ia]. The last column in Table 7 implies that over 60% of predictions in GNoMEs 1&2 are geometrically close to known crystals in the ICSD and MP up to 0.05Å.

Table 7 Statistics of distances from GNoMEs 1&2 to the ICSD and MP, see histograms in Fig. 5.

GNoME version	reference database	EMD on PDA[100] in Angstroms for quantiles					and under a threshold τ		
		5%	10%	25%	50%	75%	0.01Å	0.02Å	0.05Å
1	ICSD	0.0123	0.0166	0.0252	0.0376	0.0547	3.3%	15.2%	69.7%
2	ICSD	0.0124	0.0164	0.0241	0.0357	0.0516	3.0%	16.3%	73.1%
1	MP	0.0099	0.0139	0.0217	0.0334	0.0502	5.1%	21.3%	74.8%
2	MP	0.0096	0.0131	0.0204	0.0311	0.0466	5.5%	24.1%	78.4%
		RMSD by StructureMatcher in Å for quantiles					and under a threshold τ		
		5%	10%	25%	50%	75%	7.4%	22.6%	62.8%
1	ICSD	0.0069	0.0106	0.0188	0.0312	0.0484	7.4%	22.6%	62.8%
2	ICSD	0.0074	0.0108	0.0183	0.0301	0.0468	7.3%	24.2%	66.1%
1	MP	0.0057	0.0088	0.0158	0.0275	0.0446	10.8%	30.6%	71.1%
2	MP	0.0058	0.0086	0.0148	0.0258	0.0424	11.8%	34.1%	74.3%

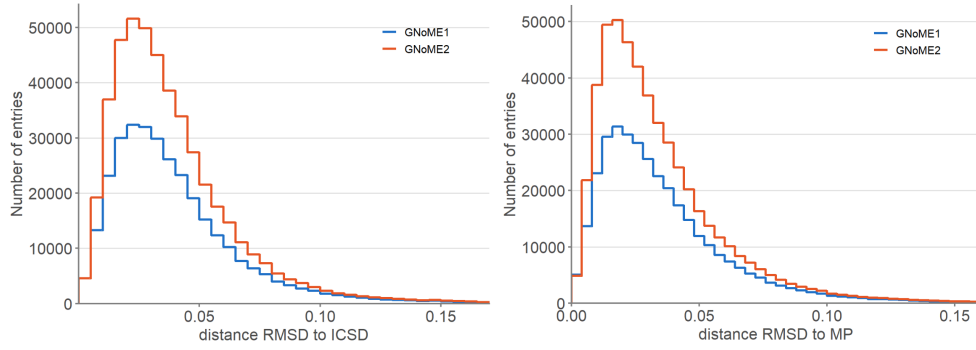


Fig. 5 The histograms of RMSD in Angstroms, obtained by StructureMatcher in pymatgen [44] between crystals in GNoMEs 1&2 and their neighbours in known databases. **Left:** ICSD. **Right:** MP.

3.3 Exposing the evolution of GNoME1 into GNoME2

Comparisons by crystal invariants reveal that GNoME1 largely evolved into GNoME2 by removing all groups of *geochemical matches* (with identical geometry and chemistry) by deleting full groups rather than by keeping one representative. In GNoME1 at the threshold $\tau = 10^{-10} \text{ \AA}$ (close to a floating-point error), of the 1,133 groups of geochemical matches, 1,010 groups (89%) disappeared and only 123 collapsed to a single entry in GNoME2. Loosening the threshold reverses the result: at $\tau = 10^{-2} \text{ \AA}$ only 1,817 of 4,172 groups (44%) were deleted, while 2,355 groups (56%) shrank.

GNoME2 introduced extra geometric duplication. Of 273 groups of geometric duplicates in GNoME2 at $\tau = 10^{-10} \text{ \AA}$, 179 groups were duplicated in GNoME1, but 62 grew out of GNoME1 singletons, and 32 consist entirely of newly added entries. At $10^{-2} \text{ \AA}$, the corresponding counts are 12,979, 5,047, and 6,168 of 24,194 groups.

$\tau, \text{ \AA}$	geometric groups	deleted	partly deduplicated	deduplicated	kept	grown
10^{-10}	1,567	1,064	324	12	160	7
10^{-5}	2,167	1,170	609	23	349	16
10^{-4}	2,324	1,239	689	23	357	16
10^{-3}	2,744	1,342	832	45	492	33
10^{-2}	19,610	2,693	3,938	845	7,494	4,640
geochemical groups						
10^{-10}	1,133	1,010	123	0	0	0
10^{-5}	1,274	1,083	191	0	0	0
10^{-4}	1,417	1,147	270	0	0	0
10^{-3}	1,618	1,234	384	0	0	0
10^{-2}	4,172	1,817	2,355	0	0	0

Table 8 What became of GNoME1 matched groups in GNoME2. Only groups of sizes $s_1 \geq 2$ in GNoME1 are listed. *Deleted*: every member dropped (new size $s_2 = 0$); *deduplicated*: exactly one member survives ($s_2 = 1$); *partially deduplicated*: $1 < s_2 < s_1$; *kept*: $s_2 = s_1 > 1$; *grown*: $s_2 > s_1$.

More than half of geometric redundancy in GNoME2 is inherited from GNoME1, i.e. 179 of 273 groups at $10^{-10} \text{ \AA}$ and 12,979 of 24,194 at $10^{-2} \text{ \AA}$. Only a minority

(32 groups at $10^{-10}\text{\AA}$ and 6,168 at $10^{-2}\text{\AA}$) consists entirely of newly added entries. Tables 8 and 12 summarise the fates of groups and matched entries, respectively, when GNoME1 evolved into GNoME2. Fig. 10 visualises the fates at several thresholds.

3.4 The most striking examples in the GNoMEs 1&2 datasets

In GNoME1, among 2,311 entries with repeated CIFs, the entries cdc06a1a2a and 0e2d8f26d6 stand out because each CIF has two pairs of repeated atoms Na at the same position in [19, appendix A, Fig. 15] and is not parsable by pymatgen [45].

In GNoME1, entry 83cc15ef46 (H_2O_{11}) consists of chaotically located molecules of water H_2O and oxygen O_2 in ratio 1 : 5, and is unlikely to be a new useful material. While the authors of [38] aimed to “introduce novel compositions in the dataset instead of similar structures that may be missed by structure matchers”, Fig. 6 shows that “new compositions” can be obtained from existing ones by just adding water (ice).

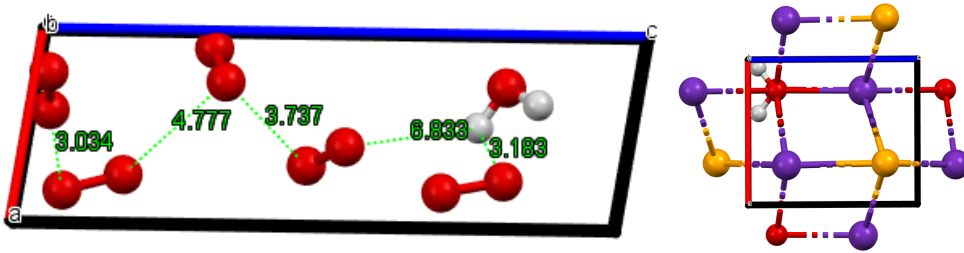


Fig. 6 Left: entry 83cc15ef46 (H_2O_{11}) in GNoME1. Right: entry 8175bee583 ($\text{H}_2\text{ORb}_2\text{Se}$) in both GNoMEs has a water molecule H_2O in the top left corner, all bonds are visualised by Mercury.

To further identify unusual entries, we used the smallest and largest distances to the first atomic neighbour, denoted by $d_{\min}$ and $d_{\max}$, respectively. The point density $\rho = \frac{m}{\text{vol}[U]}$ is the number m of motif points divided by the volume of the unit cell, given in atoms per nm^3 . The physical density is measured in g/cm^3 .

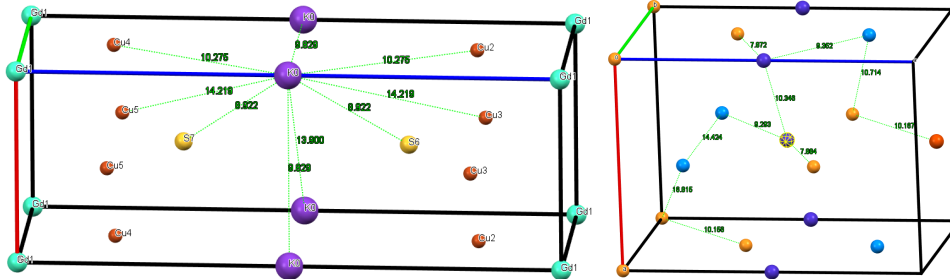


Fig. 7 Unrealistic predictions in GNoME1. Left: entry eed3027c1e has atom K with no neighbours up to $d_{\max} = 8.92\text{\AA}$. Right: entry 37f9f79871 has no pairs of atomic neighbours up to $d_{\min} = 7.87\text{\AA}$.

Significant deviations of these basic invariants from typical values in experimental datasets indicate that a crystal may not be physically realistic. In GNoME1, the sparsest entry 88e5e5c7a3 ($\text{S}_7\text{Gd}_2\text{Tb}_3$) has 1.29 atoms per nm^3 , forty times fewer than the median value in GNoME1, with a density of 0.18 g/cm^3 . Entry eed3027c1e ($\text{KGd}(\text{Cu}_2\text{S})_2$) has the largest $d_{\text{max}} = 8.92 \text{ \AA}$ with a K atom floating too far away from all others. Entry 37f9f79871 ($\text{P}_3\text{FeCo}_2\text{U}_2$) has 1.81 atoms/ nm^3 , density 0.28 g/cm^3 , and $d_{\text{min}} = 7.87 \text{ \AA}$, so atoms in this prediction have no close contact at all, see Fig. 7.

4 Scaling easy disguises of real crystal structures

Atomic vibrations and experimental noise inevitably produce slightly different cell parameters and atomic coordinates, whether that is due to different instruments, software, or conditions, such as temperature or pressure. A manual search for near-duplicates of 43 A-lab crystals in the ICSD by unit cell parameters missed several crystals that appeared in the ICSD earlier with different unit cells, see [27, Fig. 5].

The discontinuity of reduced cells under trivial perturbations, as in Fig. 1 (top right), implies that any comparison based on a unit cell or a single (even if large) finite subset can inevitably miss many near-duplicates, as we found with the GNoMEs.

More importantly, the discontinuity of cell-based representations allows anyone to disguise an existing experimental crystal by the following steps:

- (1) Shuffling (lines of) atoms and their indices in a CIF and shifting the origin.
- (2) Applying a linear transformation that changes and scales up a given unit cell.
- (3) Adding noise to the atom positions to make the new scaled-up cell primitive.

While the steps (1)-(2) keep the crystal structure rigidly equivalent to the original, a CIF substantially changes. Moreover, any raw data, such as diffraction patterns, can also be slightly perturbed. Step (3) discontinuously affects a primitive cell by size and hence makes all attempts at cell reduction obsolete by disguising a crystal beyond the capacity of cell-based tools, see details of StructureMatcher [45] in appendix B.2.

The packing similarity in Mercury compares a fixed number of molecules (15 by default) only in organic crystals. Adapting it to inorganic crystals would require exponentially many matchings of chemically identical atoms, as in the organic case, and still would not guarantee continuity because of a cut-off radius. For complete isosets with a cut-off, the discontinuity was resolved by a boundary-tolerant metric in [19, Definition 4.1] without additional parameters that are used in SOAP descriptors [46].

If novelty is evaluated only by chemical composition, one can replace chemical elements with similar ones. Hence, it is trivial to generate infinitely many different CIFs without essentially changing the underlying geometry of a crystal structure.

The second author implemented a white-box algorithm [47] that can generate 2.2 million new-looking, realistic crystals in about a day on a small laptop, using transparent school-level geometry instead of black-box AI, see details in appendix C.

The program can disguise any existing crystal so that its CIF looks completely different, and the final crystal is guaranteed to be physically realistic as a near-duplicate

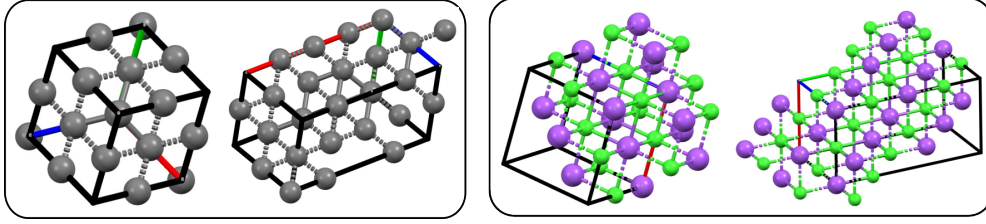


Fig. 8 An original crystal and its near-duplicate that has a different CIF with a larger primitive cell. **Left:** diamond, mp-394 in the Materials Project (MP). **Right:** NaCl, mp-22862 in the MP.

of the original, as in Fig. 8. The first presentation of this crystal disguise received a Science Slam award at the European Crystallographic Meeting in August 2024.

5 Necessary conditions for next-generation real AI

John Maddox [1] acknowledged the “general impossibility” of crystal structure prediction as a “scandal” in 1988, and Sally Price FRS [48] later rephrased it as the “embarrassment of over-prediction” because very few predictions are synthesised.

The far greater “scandal” was the absence of a rigorous definition of a crystal structure, which fuelled a “crisis of trust” in applied sciences, including paper mills [49] and large-scale retractions in crystallography [50], which continue to this day [51].

Indeed, simulations can produce numerous near-duplicates as slightly different approximations to the same local minimum of complicated energy [52]. The past approach was to report “meta-stable” predictions close to the boundary of the convex hull on points (composition, energy) within a small threshold like 1 meV/atom [39].

Any fixed threshold lets us generate infinitely many near-duplicates, all with different CIFs and similar geometries that should have energies close to the current boundary of the convex hull. For the “most stable” crystals exactly on the boundary, some fraction of their trivial perturbations have slightly different energies and likely appear on a “new” boundary in a high-dimensional space without predicting new materials. Past measures, such as SUN (stability, uniqueness, novelty), used cell-based comparisons but started to include the first continuous invariant $\text{AMD}[k]$ [53, 54].

After a (periodic) crystal structure was mathematically defined in June 2024 as a class of all crystal representations that are equivalent under rigid motion [10, Definition 6] and after the development of a hierarchy of continuous invariants [18, 19, 23, 26], the scale of artificial imitation in materials discovery has become much clearer.

To tackle the artificially created crisis in computational trial-and-error at scale, we propose the following necessary conditions for any next-generation explainable AI:

(1) **Input credibility** for periodic crystals should be guaranteed in terms of stable-under-noise invariants instead of discontinuous descriptors. Indeed, a cell-based CIF of an infinite periodic crystal is like a single snapshot of a complicated rigid structure, which can have infinitely many photographs taken from different points of view.

(2) **Chemical utility** should be evaluated by domain experts beyond comparing compositions as strings. Also, radioactive and toxic elements should be avoided.

(3) **Structural novelty** of any crystal S with respect to a dataset D of known crystals should be quantified by the continuous *Local Novelty Distance* $\text{LND}\langle d, I, D \rangle(S) = \min_{Q \in D} d(I(S), I(Q))$ for a metric d on a continuous invariant I of crystal structures.

(4) **The atomic geometry** should satisfy reasonable constraints by quantifying the average and maximum deviations from expected covalent and ionic bond lengths, depending on chemical elements, to avoid ‘new’ but physically unrealistic predictions.

The past work [27] introduced LND using EMD on the invariant PDA[100] to quantify the novelty of 43 predictions by the A-lab relative to the ICSD and MP. This LND has been adopted as the first stress test for AI-guided materials discovery [55].

Continuous structural novelty is needed to reliably filter out near-duplicates in any database. Without deduplicating an initial dataset, any splitting into 80:20 subsets for training and testing can keep near-duplicates in each part. Preventing this *data leakage* requires formalising the basic questions “Same or different? If different, by how much?” for all types of data under non-trivial equivalences in [2, Problem 1.4.5].

Conditions (1)-(4) are only minimal requirements for any justified AI and should be complemented in the future with more physics-based criteria and experimental evidence, regarding stability (synthesizability) and utility (functional properties).

In particular, there is no sense in maximising any version of LND without constraints on atomic geometry, see more unrealistic predictions in Fig. 15 and Table 21. After embracing Goodhart’s law “when a measure becomes a target, it ceases to be a good measure” [56], science can finally progress beyond optimising single-value functions, which are unlikely to characterise the synthesizability of real materials.

Concerning experimental validation of the GNoME, [57, p. 29638] said that “we first search for Cs_nABCl_6 structures that are computationally stable or close to stable in the combined zero-temperature convex hull of crystals from Materials Project, OQMD, and GNoME databases” and then reported successful in situ XRD for materials $\text{Cs}_2\text{LiCrCl}_6$, $\text{Cs}_2\text{LiFeCl}_6$, $\text{Cs}_2\text{LiIrCl}_6$, $\text{Cs}_2\text{LiRhCl}_6$, $\text{Cs}_2\text{LiRuCl}_6$, $\text{Cs}_2\text{NaDyCl}_6$ in [57, Table 1]. None of these compositions were found in GNoMEs 1&2. The only other compound, synthesised with impurities, was $\text{Cs}_2\text{LiVCl}_6$. For this composition, prediction 66ae3fec06 appeared in GNoME1 (not in GNoME2) and has the following nearest neighbours in the ICSD and MP with small EMDs on invariants PDA[100]:

ICSD 002064 (Cl_3CsCu), published in 1974 [58] and deposited in 1980, has two Cu atoms, which are replaced with Li and V to match 66ae3fec06 with $\text{EMD} \approx 0.064\text{\AA}$.

mp-1206187 ($\text{Cl}_6\text{LiRb}_2\text{Rh}$), published in 1979 [59], has atoms Rb and Rh, which are replaced with Cs and V, respectively, to match 66ae3fec06 with $\text{EMD} \approx 0.043\text{\AA}$.

The same paper [57, p. 29638] added that “we selected the compounds that did not have a matching ICSD entry (by composition) ... identifying compounds not in the ICSD is commonly considered a benchmark for experimental novelty”. To excel in this “novelty” benchmark, it suffices to just add water (ice), as illustrated in Fig. 6.

Another paper [60] experimentally verified a compound “MnFeCo₄Si₂ discovered by GNoME”. For this composition, the prediction d8afdc992d in both GNoMEs has a near-duplicate 675046 in the ICSD with EMD $\approx 0.022\text{\AA}$, deposited in 2018 [61].

Any finite dataset is a measure-zero subset of the space of all continuous data, including physical structures whose atoms have real coordinates. For more practical value, finitely many examples can help formulate conjectures that should be converted into justified guarantees, such as generic completeness and Lipschitz continuity.

For instance, 100 values of the invariant $\text{ADA}(S;100)$ sufficed to put all non-isometric crystals in major databases at their unique locations on a continuous map of the materials space. This Crystal Isometry Space, introduced in [27, Definition 1], extends Mendeleev’s table of chemical elements to unambiguously identify all periodic crystals, like houses on a geographic map through infinite-precision satellite cameras. After that, whether to assign different houses to one village (cluster) or another one using manually chosen thresholds is more an administrative than a scientific decision.

While both humans and AI can make mistakes, meaningful progress means learning from them. Just as Isaac Newton transformed the simple mistake of sitting under an apple tree into the law of gravity, we welcome collaborations that advance sustainable science towards rigorously justified solutions based on real Analytic Intelligence.

We thank Richard Catlow FRS, Andy Cooper FRS, Sally Price FRS, Simon Billinge, Ram Seshadri, Robert Palgrave, Leslie Schoop, and Mike Glazer for fruitful discussions. This work was funded by the Royal Society Short Industry Fellowship (SIF/R2/252048) at the Cambridge Crystallographic Data Center, which re-implemented the PDD invariants for inclusion in the release of Mercury in 2027.

References

- [1] Maddox, J. Crystals from first principles. *Nature* **335**, 201 (1988).
- [2] Anosova, O. & Kurlin, V. Geometric Data Science (290+ pages, the latest version at <https://kurlin.org/geometric-data-science-book.pdf>). *arxiv:2512.05040* (2025).
- [3] Morissette, S. L., Soukasene, S., Levinson, D., Cima, M. J. & Almarsson, Ö. Elucidation of crystal form diversity of the hiv protease inhibitor ritonavir by high-throughput crystallization. *Proceedings of the National Academy of Sciences* **100**, 2180–2184 (2003).
- [4] Cheetham, A. K. & Seshadri, R. Artificial intelligence driving materials discovery? Perspective on the article: Scaling deep learning for materials discovery. *Chemistry of Materials* **36**, 3490–3495 (2024).
- [5] Leeman, J. *et al.* Challenges in high-throughput inorganic materials prediction and autonomous synthesis. *PRX Energy* **3**, 011002 (2024).
- [6] Brock, C. Change to the definition of “crystal” in the IUCr Online Dictionary of Crystallography.

www.iucr.org/news/newsletter/etc/articles?issue=151351&result_138339_result_page=17 (2021).

- [7] Isostructural crystals in the International Union of Crystallography dictionary. <https://dictionary.iucr.org/Isostructural.crystals>.
- [8] Niggli, P. *Krystallographische und strukturtheoretische Grundbegriffe* Vol. 1 (Akademische verlagsgesellschaft mbh, 1928).
- [9] Lawton, S. & Jacobson, R. The reduced cell and its crystallographic applications. Tech. Rep., Ames Lab, Iowa State University (1965).
- [10] Anosova, O., Kurlin, V. & Senechal, M. The importance of definitions in crystallography. *International Union of Crystallography Journal* **11**, 453–463 (2024).
- [11] Oms, S. & Zardini, E. *The sorites paradox* (Cambridge University Press, 2019).
- [12] de Pablo et L. New frontiers for the materials genome initiative. *npj Computational Materials* **5**, 1–23 (2019).
- [13] Nosengo, N. The material code. *Nature* **533**, 22–25 (2016).
- [14] Rowe, D. E. *Emmy Noether - Mathematician Extraordinaire* (Springer, 2021).
- [15] Noether, E. Invariant variation problem. *Königlich Gesellschaft der Wissenschaften Göttingen Nachrichten Mathematik-physik Klasse* **2**, 235–267 (1918).
- [16] Parthé, E. et al. *TYPIX standardized data and crystal chemical characterization of inorganic structure types* (Springer Science & Business Media, 2013).
- [17] Mosca, M. & Kurlin, V. Voronoi-based similarity distances between arbitrary crystal lattices. *Crystal Research and Technology* **55**, 1900197 (2020).
- [18] Widdowson, D., Mosca, M. M., Pulido, A., Cooper, A. I. & Kurlin, V. Average minimum distances of periodic point sets - foundational invariants for mapping all periodic crystals. *MATCH Commun. Math. Comput. Chem.* **87**, 529–559 (2022).
- [19] Anosova, O., Widdowson, D. & Kurlin, V. Recognition of near-duplicate periodic patterns by continuous metrics with approximation guarantees. *Pattern Recognition* **171**, 112108 (2026).
- [20] Rass, S. et al. Metricizing the euclidean space towards desired distance relations in point clouds. *IEEE Trans. Information Forensics and Security* **19**, 7304–7319 (2024).
- [21] Widdowson, D. & Kurlin, V. Resolving the data ambiguity for periodic crystals. *Advances in Neural Information Processing Systems* **35**, 24625–24638 (2022).

- [22] Kantorovich, L. Mathematical methods of organizing and planning production. *Management science* **6**, 366–422 (1960).
- [23] Widdowson, D. & Kurlin, V. Pointwise distance distributions for detecting near-duplicates in large materials databases. *SIAM Journal on Applied Mathematics* **86**, 898–918 (2026).
- [24] Pauling, L. & Shappell, M. D. The crystal structure of bixbyite and the c-modification of the sesquioxides. *Zeitschrift für Kristallographie-Crystalline Materials* **75**, 128–142 (1930).
- [25] Patterson, A. Homometric structures. *Nature* **143**, 939–940 (1939).
- [26] Widdowson, D. & Kurlin, V. Higher-order, generically complete, continuous, and polynomial-time invariants of crystalline materials. *Computational and Applied Mathematics (arXiv:2509.15088)* (2027).
- [27] Widdowson, D. & Kurlin, V. Geographic-style maps with a local novelty distance help navigate in the materials space. *Scientific Reports* **15**, 27588 (2025).
- [28] Anosova, O. & Kurlin, V. An isometry classification of periodic point sets. *Proceedings of Discrete Geometry and Mathematical Morphology* 229–241 (2021).
- [29] Edelsbrunner, H., Heiss, T., Kurlin, V., Smith, P. & Wintraecken, M. The density fingerprint of a periodic point set. *Proceedings of SoCG* **189**, 32:1–32:16 (2021).
- [30] Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The Cambridge Structural Database. *Structural Science* **72**, 171–179 (2016).
- [31] Gieseke, F., Heinermann, J., Oancea, C. & Igel, C. Buffer kd trees: processing massive nearest neighbor queries on GPUs. *Intern. Conf. Machine Learning* 172–180 (2014).
- [32] Orlin, J. B. A faster strongly polynomial minimum cost flow algorithm. *Operations research* **41**, 338–350 (1993).
- [33] Shirdhonkar, S. & Jacobs, D. Approximate earth mover’s distance in linear time. *Conference on Computer Vision and Pattern Recognition* 1–8 (2008).
- [34] Szymanski, N. J. *et al.* An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature* 86–91 (2023).
- [35] Zeni, C. *et al.* A generative model for inorganic materials design. *Nature* **639**, 624–632 (2025).
- [36] Chawla, D. S. Duplicate structures haunt crystallography databases. *Chemical and Engineering News* (2025). URL <https://cen.acs.org/research-integrity/Duplicate-structures-haunt-crystallography-databases/103/web/2025/12>.

- [37] Szymanski *et al.* Author correction: An autonomous laboratory for the accelerated synthesis of inorganic materials. all corrections are highlighted at https://static-content.springer.com/esm/art%3a10.1038%2f41586-025-09992-y/mediaobjects/41586_2025_9992_moesm1_esm.pdf. <https://www.nature.com/articles/s41586-025-09992-y> (2026).
- [38] Merchant, A. *et al.* Scaling deep learning for materials discovery. *Nature* 80–85 (2023).
- [39] Google. Materials discovery: GNoME datasets (accessed 18 september 2026). https://github.com/google-deepmind/materials_discovery/blob/main/DATASET.md (2023).
- [40] Merchant, A. & Cubuk, E. D. Millions of new materials discovered with deep learning (accessed 18 september 2026). <https://deepmind.google/discover/blog/millions-of-new-materials-discovered-with-deep-learning/> (2023).
- [41] Feynman, R. *The Feynman lectures on physics: the new millennium edition* Vol. 1 (Basic Books, 2011).
- [42] Allen, F. H., Cole, J. & Howard, J. A systematic study of coordinate precision in x-ray structure analyses. ii. predictive estimates of esd’s for the general-atom case. *Foundations of Crystallography* **51**, 112–121 (1995).
- [43] Materials project v2021.03.22 (2021). URL <https://docs.materialsproject.org/changes/database-versions>.
- [44] Ong, S. P. *et al.* Python materials genomics (pymatgen): A robust, open-source Python library for materials analysis. *Computational Materials Science* **68**, 314–319 (2013).
- [45] Pymatgen StructureMatcher. https://pymatgen.org/pymatgen.core.html#pymatgen.core.structure_matcher.StructureMatcher.
- [46] Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Physical Review B—Condensed Matter and Materials Physics* **87**, 184115 (2013).
- [47] Kurlina, T. Disguise a crystal in python (2026). URL <https://github.com/tatiana1666/Disguise-A-Crystal>.
- [48] Price, S. L. Is zeroth order crystal structure prediction (CSP_0) coming to maturity? What should we aim for in an ideal crystal structure prediction code? *Faraday Discussions* **211**, 9–30 (2018).
- [49] Bimler, D. Better living through coordination chemistry: A descriptive study of a prolific papermill that combines crystallography and medicine. <https://www.researchsquare.com/article/rs-1537438/v1> (2022).

- [50] Harrison, W., Simpson, J. & Weil, M. Editorial. *Acta Cryst E* **66**, e1–e2 (2010).
- [51] Kurlin, V. Duplicate dilemma: resolved in 2022 after being open at least since 1965. <https://www.iucr.org/news/newsletter/volume-34/number-2/duplicate-dilemma-resolved-in-2022-after-being-open-at-least-since-1964> (original version at <https://kurlin.org/news/duplicate-dilemma-resolved.pdf>) (2026).
- [52] Mardirossian, N. & Head-Gordon, M. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Molecular physics* **115**, 2315–2372 (2017).
- [53] Negishi, M., Park, H., Mastej, K. O. & Walsh, A. Continuous sun (stable, unique, and novel) metric for generative modeling of inorganic crystals. *Machine Learning: Science and Technology* **7**, 035064 (2026).
- [54] Babu, A., Gouvêa, R. A. & Rignanese, G.-M. Toward automated discovery with generative models multimodal learning and closed loop workflows in inverse materials design. *Cell Reports Physical Science* **7**, 103561 (2026).
- [55] Zhang, X. *et al.* New crystal structures hide in plain sight: A stress test for ai-guided materials discovery. *arXiv:2603.05613* (2026).
- [56] Fire, M. & Guestrin, C. Over-optimization of academic publishing metrics: observing goodhart’s law in action. *GigaScience* **8**, giz053 (2019).
- [57] Miura, A. *et al.* Efficient exploratory synthesis of quaternary cesium chlorides guided by in silico predictions. *Journal of the American Chemical Society* **146**, 29637 (2024).
- [58] Kroese, C., Maaskant, W. & Verschoor, G. The high-temperature structure of CsCuCl₃. *Structural Science* **30**, 1053–1056 (1974).
- [59] Meyer, G. & Dietzel, E. Zur struktursystematik der elpasolith-familie]: Neue chloride A1+2B1+M3+Cl6 (A1+= Rb, Tl, In, K; B1+= Li, Ag, Na). *Revue de Chimie Minerale* **16**, 189–202 (1979).
- [60] Naganuma, S. & Kitagawa, J. Experimental investigation of magnetic properties of MnFeCo₄Si₂ discovered by gnome. *Journal of Magnetism and Magnetic Materials* 174056 (2026).
- [61] Abderrahim, B. *et al.* Half-metallic magnetism of quaternary heusler compounds co₂fe_xmn_{1-x} si ($x = 0, 0.5$, and 1.0): First-principles calculations. *Journal of Superconductivity and Novel Magnetism* **29**, 277–283 (2016).
- [62] Sacchi, P. *et al.* Same or different – that is the question: identification of crystal forms from crystal structure data. *Cryst Eng Comm* **22**, 7170–7185 (2020).

- [63] Coles, S. J., Threlfall, T. L. & Tizzard, G. J. The same but different: isostructural polymorphs and the case of 3-chloromandelic acid. *Crystal Growth & Design* **14**, 1623–1628 (2014).
- [64] 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).
- [65] Billinge, S. & Jensen, K. *Atomic pair distribution function analysis: a primer* Vol. 22 (Oxford University Press, 2023).
- [66] Kurlin, V. Complete and continuous invariants of 1-periodic sequences in polynomial time. *SIAM J Mathematics of Data Science* **7**, 1643–1663 (2025).
- [67] Materials project v2026.04.13 (2026). URL <https://docs.materialsproject.org/changes/database-versions>.
- [68] Kurlin, V. Complete invariants of atomic clouds under rigid motion with lip-schitz continuous metrics in a polynomial time. *MATCH Communications in Mathematical and in Computer Chemistry* **97**, 379–428 (2027).
- [69] Hall, S. R., Allen, F. H. & Brown, I. D. The crystallographic information file (cif): a new standard archive file for crystallography. *Acta Crystallographica Section A* **47**, 655–685 (1991).
- [70] Fishman, G. *Monte Carlo: concepts, algorithms, and applications* (Springer Science & Business Media, 2013).

A Details of mathematics for periodic crystals

A.1 Mathematical definitions of crystallographic concepts

We start from a classical representation as in a Crystallographic Information File (CIF) and remind of the new concept of a crystal structure [10] in Definition A.4.

Definition A.1 (lattice, unit cell, periodic crystal). **(a)** Vectors $\vec{v}_1, \dots, \vec{v}_n \in \mathbb{R}^n$ form a basis if any vector in $\mathbb{R}^n$ can be written as $\vec{v} = \sum_{i=1}^n t_i \vec{v}_i$ for unique $t_1, \dots, t_n \in \mathbb{R}$.

(b) Any basis $\vec{v}_1, \dots, \vec{v}_n \in \mathbb{R}^n$ defines the lattice $\Lambda = \{ \sum_{i=1}^n c_i \vec{v}_i \mid c_1, \dots, c_n \in \mathbb{Z} \}$ and the unit cell $U = \{ \sum_{i=1}^n x_i \vec{v}_i \mid 0 \leq x_i < 1, i = 1, \dots, n \} \subset \mathbb{R}^n$.

(c) For any finite motif of points $M \subset U$, the sum $S = M + \Lambda = \{ \vec{p} + \vec{v} \mid p \in M, v \in \Lambda \}$ is called a periodic point set. In dimension $n = 3$, if every point of S has an atomic type, which is a chemical element, possibly with a charge and or a mass number (for isotopes of the same element), then S can be called a periodic crystal.

A periodic crystal $S \subset \mathbb{R}^3$ can generate many periodic point sets. We can take points at all atomic centres of S , or points only at centres of specific atoms, or centres of mass of molecules, or any characteristic points that form a periodic set.

After data objects such as periodic crystals are introduced, the most important question of data science is *Same or different?* [62, 63], which appeared even in titles of crystallographic papers. An objective answer to this question requires us to define what crystals can be called *the same* by choosing a suitable equivalence relation.

Definition A.2 (equivalence relation, equivalence class). *(a) For any given objects, an equivalence is a binary relation (denoted by $S \sim Q$) satisfying three axioms:*

- (1) reflexivity: *any object $S \sim S$;*
- (2) symmetry: *if $S \sim Q$ then $Q \sim S$;*
- (3) transitivity: *if $S \sim Q$ and $Q \sim T$ then $S \sim T$.*

(b) The equivalence class $[S]$ of an object S consists of all objects Q equivalent to S .

(c) One equivalence $S \sim_1 Q$ is called weaker or coarser than another $S \sim_2 Q$ (or $S \sim_2 Q$ is stronger or finer than $S \sim_1 Q$) if $S \sim_2 Q$ always implies that $S \sim_1 Q$.

Any equivalence relation guarantees a well-defined classification by splitting all objects into disjoint equivalence classes. Indeed, if two classes $[S]$ and $[T]$ share a common object Q , then $[S] = [T]$ by the transitivity axiom. If an equivalence is not defined *a priori* or axioms fail, all subsequent claims are hard to objectively justify.

For periodic crystals, one simple equivalence is by chemical composition: $S \sim Q$ if S, Q have the same ratios of chemical elements. Under this equivalence, all carbon allotropes such as diamond and graphite with vastly different properties fall into the same equivalence class. To distinguish all polymorphs, we need a stronger (finer) equivalence that splits all periodic crystals into a wider variety of smaller classes.

On the other hand, a traditional CIF specifying a unit cell and a motif of atoms defines a too strict and impractical equivalence because one can shift all atoms of the motif by the same short vector even within a fixed unit cell and get a very different CIF of the same structure. Hence a CIF is similar to a single snapshot (a 3D photo) but the same object can have infinitely many photos in different coordinate systems.

Crystallography tried to overcome this ambiguity by reducing a cell to a canonical form, e.g. Niggli cell [8], and then by putting atoms at the most symmetric (Wyckoff) positions inside the fixed cell. However, all reduced cells discontinuously change under almost any noise, which was reported experimentally in 1965 [9, p. 80] and theoretically proved in 2022 [18, Theorem 15] even for lattices with empty motifs and any potential reduced cell. Though this discontinuity was practically resolved in 2022 [21, Theorems 4.3-4.4], many materials databases still accumulate near-duplicates.

A practical mid-point compromise between coarse equivalences (by chemical composition or another property) and too strict coincidences (by cell-based CIFs or other coordinate-dependent representations) is a *rigid motion* (transformation) justified by the fact that structures of solid crystalline materials are determined in a rigid form.

Example A.3 (rigid motion and isometry). A rigid motion is a composition of translations and rotations in $\mathbb{R}^n$. An isometry is a distance-preserving map $\mathbb{R}^n \rightarrow \mathbb{R}^n$. A dilation is a composition of isometry and uniform scaling in $\mathbb{R}^n$.

In $\mathbb{R}^n$, any isometry is either a rigid motion or a composition of a mirror reflection with a suitable rigid motion and is also called a *congruence* in Euclidean geometry, though an *isometry* makes sense in more general metric spaces with non-Euclidean distances. We use the term a *rigid motion* rather than a rigid transformation because any such map $f_1 : \mathbb{R}^n \rightarrow \mathbb{R}^n$ can be included into a continuous family (motion) of transformations $f_t : \mathbb{R}^n \rightarrow \mathbb{R}^n$, $t \in [0, 1]$, connecting a given f_1 with the identity map $f_0 = \text{id}$ that fixes every point of $\mathbb{R}^n$. Moreover, any isometry can also be called a rigid transformation in the sense that all inter-point distances are preserved. The longer alternative name for a rigid motion is an *orientation-preserving* isometry.

Under rigid motion, moduli spaces of crystals become naturally continuous because atomic coordinates have real continuous values, not restricted to a discrete grid.

Definition A.4 (periodic structures and crystal structures). In any dimension $n \geq 1$, a (rigid) periodic structure is a class of periodic sets $S \subset \mathbb{R}^n$ of unlabelled points under rigid motion in $\mathbb{R}^n$. In dimension $n = 3$, if every point of a periodic set $S \subset \mathbb{R}^3$ has an atomic type, the class of S under rigid motion is called a crystal structure.

Under almost any noise or atomic vibration [41], the class of a periodic crystal under rigid motion becomes slightly different. Similarly, almost any movement of a physical object on the surface of Earth changes its geographic coordinates (latitude and longitude). Crystal structures also live in continuous spaces defined below.

Definition A.5 (crystal spaces). The moduli space under a given equivalence is the set of all equivalence classes. For periodic sets of all atomic centers in a crystal, the moduli space under rigid motion is called the Crystal Rigid Space (CRIS), the moduli space under isometry is called the Crystal Isometry Space (CIMS), the moduli space under dilation is called the Crystal Dilation Space (CRDS).

Fig. 4 (left) visualises spaces of crystals starting from finite collections of space-group types (230 in dimension 3) and chemical compositions, which form a much larger but still a finite and discrete set because ratios of (integer numbers of) chemical elements cannot change continuously while atomic coordinates certainly can. The space of rigid classes (CRIS) is the largest in practice because there is no sense in excluding rotations or translations from an equivalence of crystals. The weaker equivalence of isometries, which includes mirror reflections, defines a smaller (roughly speaking, twice smaller) space of isometry classes (CIMS) by putting any mirror images into the same class under isometry. The even weaker equivalence of dilation puts all uniformly scaled periodic sets, e.g. all cubic lattices, into the same class in the even smaller space CRDS. Any non-uniform scaling produces different classes under dilation.

In Definition A.5 all points (zero-sized atomic centres) are considered unlabelled (indistinguishable). If we add atomic types to points, we lose the ability to compare crystals with different chemical compositions. Though it might seem that all chemistry is lost without atomic types, Richard Feynman’s table in Figure 1-7 of his first lecture [41] was the Eureka moment for the last author in May 2021 to realise that precise

enough geometry of unlabelled atomic centres should suffice to uniquely determine all chemical elements under fixed ambient conditions, such as temperature and pressure.

Our physical intuition says that we cannot change atomic types at fixed geometric positions because interatomic interactions will shift atomic centres at least slightly (with two or three decimal places in Angstroms), which should be detectable by modern instruments. To rigorously justify that geometry can determine chemistry, we need reliable tools to practically distinguish all periodic point sets, not only known crystals.

The moduli spaces in Definition A.5 are not only continuous but are also infinite-dimensional, which makes distinguishing (separating) all periodic point sets a notoriously hard problem. If we fix (an upper bound of) a number m of atoms in motifs of crystals, the resulting crystals spaces become finite-dimensional.

In $\mathbb{R}^3$, the spaces $\text{CRIS}(\mathbb{R}^3; m)$ and $\text{CIMS}(\mathbb{R}^3; m)$ have the dimension $3m$ due to 3 coordinates for each of m atoms and 6 parameters of a unit cell, which are neutralised by 6 parameters of rigid motion and isometry. In practice, noise and atomic vibrations do not care about a fixed number m of atoms and can easily scale up a primitive cell by any integer factor. Hence, every periodic crystal with m atoms in a unit cell is infinitesimally close to all (infinitely many) subspaces of crystals whose number of atoms in a unit cell is any integer multiple of m , see Fig. 1 (bottom).

A.2 Mathematical definitions of invariants and metrics

The full problem statement needs a few concepts that are important in all sciences, e.g. Definition A.6 helps rigorously answer the question *Same or different?* For simplicity, we fix the main equivalence as rigid motion in $\mathbb{R}^n$, but most concepts and the subsequent geo-mapping problem make sense for other equivalences and objects.

Definition A.6 (invariants and equivariants). *(a) An invariant I is a function that takes as input any periodic point set $S \subset \mathbb{R}^n$ and outputs a value $I(S)$ that remains the same under any rigid motion f in $\mathbb{R}^n$, i.e. $I(f(S)) = I(S)$. Alternatively, if any periodic point sets S, Q can be exactly matched by rigid motion (denoted as $S \cong Q$), then $I(S) = I(Q)$. In other words, I is a descriptor with no false negatives that are pairs of different representations of rigidly equivalent sets: $S \cong Q$ with $I(S) \neq I(Q)$.*

(b) An equivariant E is a function that takes as input any periodic point set $S \subset \mathbb{R}^n$ and outputs a value $E(S)$ that controllably changes under any rigid motion f in $\mathbb{R}^n$, i.e. $E(f(S)) = T_f(E(S))$, where a transformation T_f depends on f but not on S .

If S is a periodic crystal, Definition A.6 considers only rigid motions that also respect atomic types of S . Our physical intuition says that this restriction is not practically needed, because replacing atomic types should change relative geometric positions of atomic centers at least slightly. A rigorous justification of this conclusion requires complete invariants, which will be introduced in Definition A.7 below.

The center of mass of all atoms in a unit cell U (as a vector expressed in a basis of U) changes under any translation and hence is not invariant. This center is also not equivariant in all dimensions $n \geq 1$. The periodic sequence $S = \{0, 1\} + 2\mathbb{Z}$ has the centre $c = 0.5(0 + 1) = 0.5$ in the unit cell $[0, 2)$. The shift by $\varepsilon > 0$ to the right gives another representation $\{\varepsilon, 1 + \varepsilon\} + 2\mathbb{Z}$ of S (under translations) whose motif has the

centre $c(\varepsilon) = 0.5(\varepsilon + 1 + \varepsilon) = 0.5 + \varepsilon$ obtained by shifting $c = 0.5$ by ε to the right. The shift by $\varepsilon > 0$ to the left gives one more representation $\{2 - \varepsilon, 1 - \varepsilon\} + 2\mathbb{Z}$, where $-\varepsilon$ was replaced with $2 - \varepsilon$ within the unit cell $[0, 2)$. The new motif has the centre $c(-\varepsilon) = 0.5(2 - \varepsilon + 1 - \varepsilon) = 1.5 - \varepsilon$, which is not related to $c = 0.5$ by any ε -shift.

The basis vectors of a unit cell are equivariant but are not invariant and hence cannot distinguish any crystal structures under rigid motion. In general, all linear combinations of equivariant vectors are also equivariant, so equivariants form a much wider collection than invariants, which explains that equivariant neural networks can be easily trained to fit any finite data. Invariants are much rarer in comparison with equivariants and hence do not allow easy overfitting.

For example, the physical density of a periodic crystal S is a scalar quantity that can be computed from any CIF of S . Density satisfies the invariance, so a given crystal structure has one and only one density. Hence, if crystals have different densities, their underlying structures are also different (cannot be exactly matched by rigid motion). The important consequence of Definition A.6(a) is that any non-constant invariant distinguishes some (possibly, not all) crystals under rigid motion.

Definition A.7 (complete vs generically complete invariants). *(a) An invariant I is called complete if any periodic point sets $S, Q \subset \mathbb{R}^n$ with equal invariant values $I(S) = I(Q)$ can be exactly matched by rigid motion in $\mathbb{R}^n$.*

(b) An invariant I is called generically complete if the completeness property $I(S) = I(Q) \Rightarrow S \cong Q$ holds for almost all periodic point sets $S, Q \subset \mathbb{R}^n$ except sets that occupy a measure-zero subspace in the full space $\text{CRIS}(\mathbb{R}^n)$ of rigid classes.

The completeness property means that our desired structure representation has a one-to-one mapping. Given the value $I(S)$ of a complete invariant I computed from a structure S , there is only one structure that corresponds to this $I(S)$. Completeness is often a difficult property to satisfy. We note in passing that, whilst it is a desirable property, non-complete continuous invariants, such as density, atomic pair distribution function (PDF), and even X-ray powder diffraction patterns (XRPD), are useful invariants! Nonetheless, with that said, complete invariants are desirable to develop.

Definition A.8 (metrics vs pseudometrics). *(a) A distance metric d on values of an invariant I is a function satisfying the following metric axioms:*

- (1) coincidence: $d(a, b) = 0$ if and only if $a = b$;
- (2) symmetry: $d(a, b) = d(b, a)$; and
- (3) triangle inequality: $d(a, b) + d(b, c) \geq d(a, c)$ for any values a, b, c .

(b) If condition (1) is replaced with the weaker one including one-directional implication (1') $d(a, a) = 0$, then d is called a pseudo-metric.

Definition A.9 (ε -perturbation and Lipschitz continuity). *(a) For any $\varepsilon > 0$, periodic point sets $S, Q \subset \mathbb{R}^n$ are called ε -perturbations of each other if there is a bijection $g : S \rightarrow Q$ such that $|g(p) - p| \leq \varepsilon$ for any point $p \in S$.*

(b) An invariant I is called Lipschitz continuous in a metric d if there is a constant λ such that, for any small enough $\varepsilon > 0$, if any periodic point sets $S, Q \subset \mathbb{R}^n$ are ε -perturbations of each other, then $d(I(S), I(Q)) \leq \lambda\varepsilon$.

A.3 A hierarchy of invariants: from the fastest to complete ones

We give simplified definitions of invariants and refer to more general versions in past papers. Definition A.10 introduces the main invariant PDD (Pointwise Distance Distribution) and its average vector AMD of Average Minimum Distances.

Definition A.10 (AMD and PDD invariants). *Let $S \subset \mathbb{R}^n$ be any periodic set with a motif M of m points. Fix an integer $k \geq 1$ and a point $p \in M$. Let $d_1(p) \leq \dots \leq d_k(p)$ be the ordered list of Euclidean distances from p to its k nearest neighbours within the whole set S . The unordered collection of these m lists with equal weights $\frac{1}{m}$ for all points $p \in M$ is the Pointwise Distance Distribution $\text{PDD}(S; k)$. The vector $\text{AMD}(S; k)$ consists of the k averages $\text{AMD}_j(S) = \frac{1}{m} \sum_{p \in M} d_j(p)$ for $j = 1, \dots, k$.*

The PDD and AMD invariants can be defined for any discrete set $S = M$ of points in a metric space [23, Definition 3.1] and are invariant under isometry in $\mathbb{R}^n$ [23, Theorem 3.3]. If any number (say, l) of distance lists in $\text{PDD}(S; k)$ coincide, then can be collapsed into a single list with the weight $\frac{l}{m}$. Since all lists have the same length k , it is convenient to write the Pointwise Distance Distribution $\text{PDD}(S; k)$ as a matrix with (up to m) unordered rows and k ordered columns, plus an extra (0-th) column of weights. The columns of $\text{PDD}(S; k)$ are ordered by distance values, but rows are unordered because points of S are unlabelled. We also interpret $\text{PDD}(S; k)$ as a distribution of weighted points $(d_1(p), \dots, d_k(p))$ in $\mathbb{R}^k$, one per point $p \in M$.

One of the simple distance metrics on such distributions is Earth Mover’s Distance (EMD), which was previously applied to chemical compositions [64]. Briefly, EMD optimally splits and matches rows of two PDDs even if these distributions have different sizes and the same total (normalised) weight 1. For distributions of a maximum size m , the EMD can be computed exactly [32] in time $O(m^3 \log m)$ and approximately with a constant factor in time $O(m)$ [33, section 3]. If any points $p, q \in S$ are perturbed within their ε -neighbourhoods in $\mathbb{R}^n$, then the Euclidean distance between them changes up to 2ε due to the triangle inequality in Definition A.8(a). Extending this basic fact, [21, Theorem 4.3] proved that the PDD is continuous in the EMD with Lipschitz constant 2 for all periodic point sets and small enough perturbation $\varepsilon > 0$.

$\text{PDD}(S; k)$ may look similar to the Pair Distribution Function (PDF), which is a distribution of all inter-atomic distances [65]. This infinite discrete distribution is often normalised by frequency and further smoothed by Gaussian blurring to avoid discontinuities under scaling up and perturbing a unit cell in Fig. 1 (top right). Even if PDF is split by atomic types, the resulting invariant is weaker than the PDD, which is split by all (isometrically non-equivalent) points, hence depending only on geometry rather than on chemistry, see [21, Example 3.3] and the subsequent discussion of PDF.

The key advantage of PDD is its generic completeness [21, Theorem 4.4] in the sense of Definition A.7(b). Detailed proofs of all results about the PDD are in [23], which also extended this invariant to more general l -periodic sets $S \subset \mathbb{R}^n$ with periodicity in l directions within $\mathbb{R}^n$ for any $1 \leq n$. For $l = 1$, complete and continuous invariants of 1-periodic high-dimensional sequences in $\mathbb{R}^n$ were designed in [66].

Definition A.11 extends the PDD to a hierarchy of higher-order invariants $\text{PDD}^{\{h\}}$ that capture geometric information about relative positions of $h + 1 \geq 3$ atoms [26].

Definition A.11 (higher-order $\text{PDD}^{\{h\}}(S; k)$). *Let $S \subset \mathbb{R}^n$ be any periodic set with a motif M of m points. Fix a point $p \in M$ and integers $h, k \geq 1$. Consider any h distinct points $p_1, \dots, p_h \in S \setminus \{p\}$ and the h -order average $\frac{2}{h(h+1)} \sum_{0 \leq i < j \leq h} |p_i - p_j|$ of pairwise distances between the points $p = p_0, p_1, \dots, p_h$. Let $a(p; h, 1) \leq \dots \leq a(p; h, k)$ be the list of k smallest averages for the fixed point p and variable points $p_1, \dots, p_h \in S \setminus \{p\}$. The unordered collection of these lists with equal weights $\frac{1}{m}$ for all points $p \in M$ is called the h -order Pointwise Distance Distribution $\text{PDD}^{\{h\}}(S; k)$.*

The invariant $\text{PDD}^{\{2\}}(S; k)$ distinguished all (infinitely many) known counterexamples to the completeness of the PDD through manual computations in [26, Example 3.4], which demonstrates its simplicity in addition to strength. All invariants $\text{PDD}^{\{h\}}(S; k)$ are continuous with Lipschitz constant 2 [26, Theorem 4.3] and computable in polynomial time, for the fixed h, n [26, Theorem 5.10].

Definition A.12 (α -clusters and isoset $I(S; \alpha)$). *For a point p in a periodic point set $S \subset \mathbb{R}^n$ and any radius $\alpha \geq 0$, the α -cluster $C(S, p; \alpha)$ is the set of all vectors $\vec{q} - \vec{p}$ such that $q \in S$ and $|\vec{q} - \vec{p}| \leq \alpha$. The isoset $I(S; \alpha)$ is the unordered collection of all isometry classes $[C(S, p; \alpha)]$ of α -clusters with equal weights for all points p in a motif M of S . If we replace isometry with rigid motion, we get the oriented isoset $I^o(S; \alpha)$.*

In comparison with $\text{PDD}(S; k)$, a local cluster around a point p keeps all vectors $\vec{q} - \vec{p}$ for neighbours q without reducing them to distances. Since each cluster $C(S, p; \alpha)$ is not invariant under rotations around p , the isoset $I(S; \alpha)$ is necessarily defined via isometry classes of clusters. In practice, this abstraction is resolved by computing an approximation to the Lipschitz continuous metric between isosets, which needs only representative clusters, not their invariants, see details in [19, sections 4-5].

B More experimental details on the GNoMEs

Table 9 shows the dates and release notes of the four GNoME versions from [39]. These updates withdrew entries as well as added them, but no note is given of withdrawal counts. The only quantitative characterisation of any update is “significantly increasing the number of released crystals”. The overview still describes the dataset as “~381,000 novel structures”, although the latest version has 554,054 entries.

Old GNoME versions are online in the public bucket `gs://gdm-materials-discovery`, which contains `gnome_data/legacy_versions/`, holding subdirectories `11.29/`, `12.1.23/` and `8.21.24/` with the summary tables and, for the latter two, full structure archives. But this directory appears nowhere in the release notes. The earlier releases are therefore public but undiscoverable from the documentation, which is why comparisons of the kind reported here have not been made: reproducing them requires knowledge of a path that the documentation omits.

Table 9 Dates and public notes about the four GNoME versions, quoted from [39].

dates	release notes of GNoME versions
29 November 2023	“Initial dataset release”
1 December 2023	“Re-introduce paper filters to remove un-physical energies; add 2 missing columns”
21 August 2024	“Adjust threshold to 1meV/atom off the hull significantly increasing the number of released crystals, update to use ‘require_bound’ version of MP corrections, re-relax structure where VASP did not update lattice coordinates, provide convex hull entries from recomputations of MP / OQMD”
21 November 2024	“Ensure consistency in formation energies, provide used elemental references + pseudopotential corrections”

B.1 Details of invariant-based comparisons of the GNoMEs

We ran all experiments on modest AMD Ryzen 5 5600X 6-core, 32GB RAM.

While geometrically matching entries up to thresholds $\tau \leq 10^{-5}\text{\AA}$, all groups of matched entries in GNoME 1&2 are *cliques*, i.e., all pairs of members of a group are within τ . At $\tau = 10^{-2}$, only 9,428 groups (82,224 entries) in GNoMEs 1&2 are not cliques and may have the maximum distance among members larger than τ .

Screening. Pairs were screened in three blocks (those in GNoME1 only, in GNoME2 only, and across the two), initially by RMS on ADA[100]. The largest threshold $\tau = 10^{-2}$ returned 134,686,519 pairs, with the GNoME2 only block (added entries) being the densest. Survivors were then further compared in the sequence PDA[100] $\rightarrow$ PDA^{2}[100] $\rightarrow$ isosets, each stage removing pairs whose distance exceeded τ .

Exactly 1,453 screened pairs were trivial (known identical CIFs) and were passed through without comparison. Of the 134,685,066 pairs compared by PDA[100], 8,010,884 (5.9%) survived to PDA^{2}[100] and 1,344,998 (1.0% of all screened) to the isoset stage, of which 242,010 (18.0%) were within a threshold τ at every stage. With the trivial pairs added back, 243,463 pairs had all distances within $\tau = 10^{-2}$.

At $\tau = 10^{-10}$, GNoME1 has 3,248 entries (0.84%) in 1,567 geometric duplicate groups. These groups were previously found by comparing CIFs as texts in [10, Table 2], which amounts to 1,681 redundant copies, a redundancy rate of 0.437%.

Table 10 Pair counts at each stage of the comparison process for the threshold $\tau = 10^{-2}$. Each invariant column gives the number of pairs compared at that stage. Trivial pairs (known identical CIFs) were passed through costing nothing. The *screened* column contains the number of pairs which passed the initial comparison by ADA[100].

block	screened	trivial	pairs compared by invariants			geometric matches
			PDA	PDA ^{2}	isosets	
GNoME1 only	30,732,748	1,224	30,731,524	2,058,581	420,611	84,014
GNoME2 only	40,861,239	0	40,861,239	2,594,823	420,282	76,475
Both	63,092,532	229	63,092,303	3,357,480	504,105	82,974
total	134,686,519	1,453	134,685,066	8,010,884	1,344,998	243,463

GNoME2 contains only 591 such entries in 273 groups, 318 redundant copies and redundancy rate 0.057%, a more than seven-fold drop in rate despite GNoME2 being 44% larger, and the gap survives every threshold up to 10^{-3} in Tables 5 and 6.

For geochemical matches, GNoME1 holds 1,133 (at $\tau = 10^{-10}\text{\AA}$) and 4,172 (at $\tau = 10^{-2}\text{\AA}$) geochemical groups, giving 2,311 and 8,622 entries in total, whereas GNoME2 contains no geochemical matches up to $\tau = 0.01\text{\AA}$.

Table 11 Computational times. Every pair is screened once at the largest threshold $\tau = 10^{-2}\text{\AA}$. Results for smaller thresholds can then be derived in trivial time. The experiment ran on 8 worker processes, so times are in CPU hours summed over workers and are quoted as percentages of that sum. The total wall-clock time taken to complete comparisons was 16 hours, 11 minutes and 54 seconds.

Stage	comparisons	CPU time (hours:min:sec)	% of total
EMD on PDA[100]	134,685,066	30:19:12	26.6
EMD on PDA ^{2} [100]	8,010,884	30:51:52	27.1
EMD on isoset[5]	1,344,998	52:52:50	46.4
total over 8 workers		114:03:45	100.0

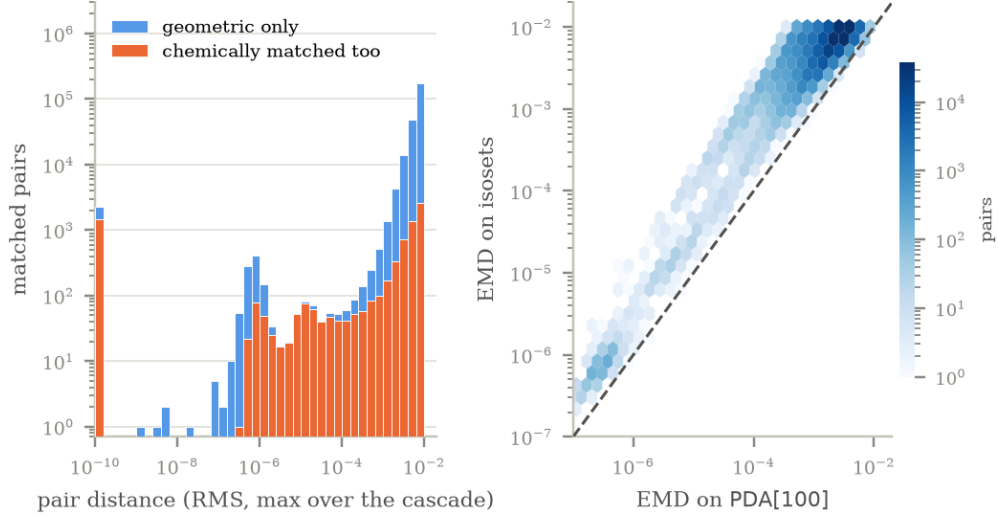


Fig. 9 **Left:** distribution of distances between all pairs compared in GNoME 1&2, taken as the maximum distance over four invariants. **Right:** EMD on PDA[100] vs EMD on isoset[5].

For GNoMEs 1&2, duplication predicts removal very strongly: 95.3% of the GNoME1 entries (2,202 of 2,311) with a geochemical duplicate were dropped at $\tau = 10^{-10}\text{\AA}$, and 73.9% (6,373 of 8,622) at 10^{-2} , against a background removal rate of 21.7% (83,525 of 384,938), but the converse is not true. Only 626 of the 83,525 removed entries (0.75%) had a geometric match that survived into GNoME2 at $\tau = 10^{-10}$,

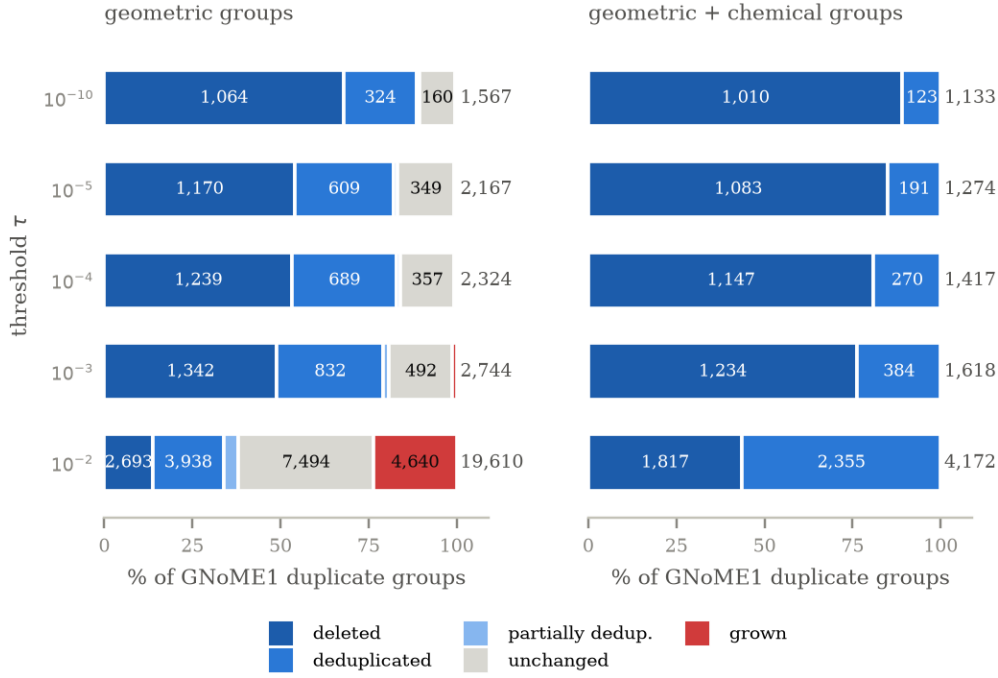


Fig. 10 Split of the fates of the various groups found in GNoME1, as described in Table 8.

and only 367 (0.44%) had a geochemical one; even at $\tau = 10^{-2}$, the figures are 11,547 (13.8%) and 5,169 (6.2%). The bulk of the 83,525 removals were unique structures.

B.2 Details of comparing the GNoMEs against ICSD and MP

This section describes findings from comparisons of both GNoME releases against the Inorganic Crystal Structure Database (ICSD) and Materials Project (MP). Steps taken in carrying out the comparisons are laid out in Section B.3.

We took ICSD and MP version [67] with the parameter “include gnome=False”, and ignored all entries deposited from March 2021, when GNoME started training.

- v2024.11.14 (released 12 Dec 2024): “The current set of 2,047 GNoME-originated materials has been deprecated in preparation for a release of about 120,000 GNoME materials”
- v2024.12.18 (released 20 Dec 2024): “Added 15,483 GNoME-originated materials calculated using r2SCAN ... A further release of an additional 100k GNoME materials is in preparation.”
- v2025.04.10 (released 18 April 2025): “Added 30,00 GNoME-originated materials calculated using r2SCAN”
- v2026.04.13 (released 8th June 2026): “Added 74,052 new GNoME materials computed at the r2SCAN level of theory. This brings the total to 117k GNoME materials.”

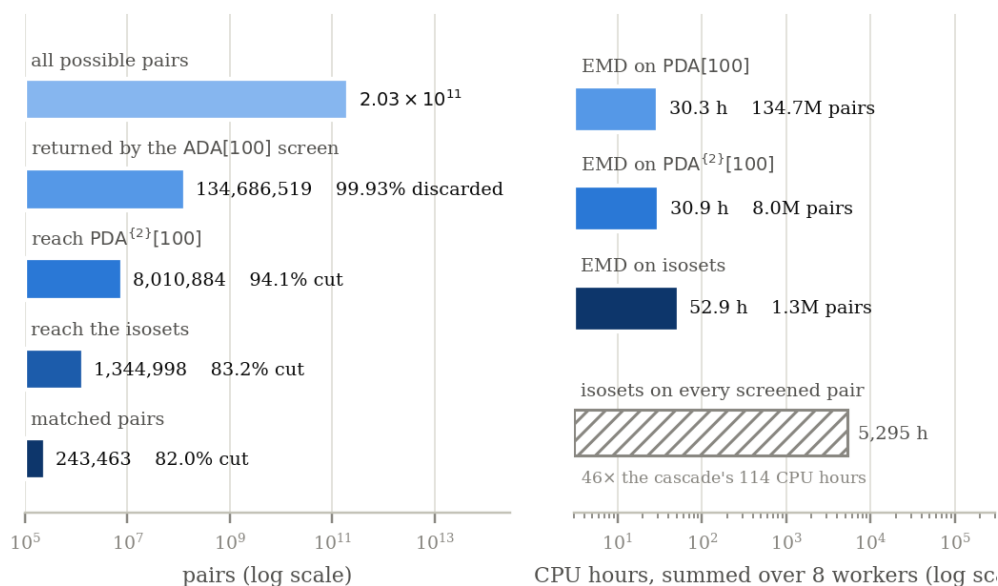


Fig. 11 Left: Number of pairs and percentage cut at each stage of comparisons. Right: Time of each stage (Table 11), and estimated time to compare all pairs passing the ADA[100] screen by isosets.

GNoME's geometry is largely the geometry of known crystals. Table 7 gives the distribution of the distance from each entry to its nearest reference structure by geometry alone. The median GNoME2 entry sits 0.034Å from some ICSD structure and 0.03Å from some MP structure. This might be explained by GNoME's generative approach, which takes existing crystals and relabels chemical elements with similar ones.

Chemical matches are rare, and the population that could produce one is small. Only 29,167 GNoME2 entries (5.3%) have any ICSD entry with the same element set, and only 5,201 (0.9%) have one with the same composition. For the Materials Project (MP), the corresponding figures are 33,566 and 3,887. In GNoME2, only 824 entries match an ICSD crystal within 0.05Å by geometry and chemistry together, and only 1 entry matches an MP crystal, totalling 0.15% of GNoME2 (Table 15). The corresponding GNoME1 figures are 1,315 and 155, totalling 0.37% (Table 14).

Duplication of known crystals dropped sharply between releases. Of the 559 GNoME1 entries that were geochemical matches to a known *ordered* crystal within 0.05, 92.7% were withdrawn in GNoME2 (against a background removal rate of 21.7%); of the 155 matching a crystal in the MP 99.4% were withdrawn, leaving one (Table 16).

Removal rates are very different for ordered and disordered entries. Matches to a known *ordered* crystal were removed at a rate of 92.7%, but matches to a *disordered* one at only 40.4%, against the 21.7% background removal rate. This difference leads us to believe that a filter was applied to remove known ICSD crystals from GNoME1, but that it was unable to identify entries matching a disordered ICSD crystal.

Table 12 Curation outcome against duplication status. Each entry is grouped by its multiplicity in its own source release (GNoME2 for added entries, GNoME1 for removed and retained ones). Removal is strongly concentrated on the duplicated entries: at the threshold $\tau = 10^{-10}$, 95.3% of the GNoME1 entries with a geochemical duplicate were dropped, against a background removal rate of 21.7%.

$\tau, \text{\AA}$	status	unique structures	geometric duplicates	geochemical duplicates	total
10^{-10}	added	252,473	168	0	252,641
	removed	80,949	374	2,202	83,525
	retained	300,741	563	109	301,413
10^{-5}	added	252,308	333	0	252,641
	removed	80,414	690	2,421	83,525
	retained	300,028	1,208	177	301,413
10^{-4}	added	252,300	341	0	252,641
	removed	80,190	703	2,632	83,525
	retained	299,932	1,226	255	301,413
10^{-3}	added	251,968	673	0	252,641
	removed	79,740	847	2,938	83,525
	retained	299,424	1,621	368	301,413
10^{-2}	added	197,073	55,568	0	252,641
	removed	67,639	9,513	6,373	83,525
	retained	249,811	49,353	2,249	301,413

Table 13 Entry turnover across the four GNoME releases, computed by matching `MaterialId` between the `stable_materials.summary.csv` of each version. *Withdrawn* counts identifiers present in the earlier release and absent from the later one. The releases are not nested: 10,201 entries that were withdrawn on 8/21/24 were restored on 11/21/24.

date	entries	change from the previous version		
		withdrawn entries	added entries	net change
11/29/23	384,938	—	—	—
12/1/23	384,871	67	0	−67
8/21/24	529,803	92,095	237,027	+144,932
11/21/24	554,054	15,615	39,866	+24,251
11/29/23 → 11/21/24		83,525	252,641	+169,116

GNoME2 still contains 774 such entries against GNoME1’s 879, a fall of 12% where matches to ordered crystals fell by 89%.

B.3 Methods of comparing GNoMEs with ICSD and MP

The comparison of GNoMEs proceeds in the following steps. The query set is the union of GNoME1 and GNoME2 totalling 637,579 structures (retained entries were compared once and reported in both releases). The reference sets were the ICSD and the Materials Project. Counts are of the union of the two releases against each reference set. Justifications for each step and distance are in Table 18.

1. *Candidate lists by ADA*. For each entry and each reference set, three lists ranked by RMS on ADA[100]: the five nearest references overall, the five nearest sharing

Table 14 GNoME1 against ICSD and MP references. Each entry is assigned the first category that applies at a threshold τ : *ordered*, a chemically constrained (geochemical) match to an ordered reference crystal; *disordered* (details in subsection B.3), and *only geometric*, a match by d_{geom} alone.

threshold τ	ordered	disordered	any chemical	only geometric	none
ICSD					
0.005	23	35	58	4,213	380,667
0.01	60	90	150	12,915	371,873
0.02	162	241	403	60,317	324,218
0.03	288	475	763	136,240	247,935
MP (no disordered structures)					
0.005	30	0	30	5,560	379,348
0.01	57	0	57	19,818	365,063
0.02	89	0	89	82,513	302,336
0.03	112	0	112	165,815	219,011

Table 15 GNoME2 against ICSD and MP; columns as in Table 14. Matches to an ordered reference crystal have mostly disappeared, while matches to a disordered one have not.

threshold τ	ordered	disordered	any chemical	only geometric	none
ICSD					
0.005	9	27	36	4,289	549,729
0.01	20	69	89	17,325	536,640
0.02	29	201	230	93,382	460,442
0.03	36	413	449	214,404	339,201
MP (no disordered structures)					
0.005	0	0	0	6,552	547,502
0.01	0	0	0	30,689	523,365
0.02	0	0	0	134,431	419,623
0.03	1	0	1	264,765	289,288

Table 16 What became of the GNoME1 entries that matched a known crystal, at $\tau = 0.05$ and against either reference set. Matches to a known ordered crystal were removed at four times the background removal rate of 21.7%, but only 40.4% of matches to disordered crystals were removed and 226 were added.

fate of crystals	ordered	disordered	any chemical	
retained (301,413)	41	508	549	
removed (83,525)	518	345	863	
added (252,641)	10	266	276	
GNoME1 category at the threshold $\tau = 0.05$			entries	removed (%)
ordered			559	92.7
disordered			853	40.4
any chemical match in ICSD			1,315	58.3
any chemical match in MP			155	99.4
no chemical match			383,526	21.6

Table 17 The closest matches to known crystals surviving in GNoME2, at $\tau = 0.05\text{\AA}$. An asterisk * marks an entry added in GNoME2. Upper panel: geochemical matches to a periodic reference crystal, d_{chem} . Lower panel: matches to a disordered ICSD crystal; the twelve closest are all ordered.

GNoME id	composition	reference	distance, $\text{\AA}$
<i>ordered matches</i> (50 in GNoME2, 18 of them to experimental ICSD)			
f7ee908e60*	IrSnTa	ICSD 689927	0.000009
c5f1363c48	GaOSc ₃	ICSD 687606	0.000066
9993fb827a	InOSc ₃	ICSD 687607	0.000154
1cd0016832	AuNNi ₃	ICSD 680114	0.000296
9d280540c9*	NNi ₃ Pd	ICSD 680121	0.000370
12b40befa0*	Co ₃ NPd	ICSD 680008	0.000672
bba5df2ac3*	Co ₃ CuN	ICSD 680017	0.000702
f4e72c3b19	AuCo ₃ N	ICSD 680006	0.000823
b39b607649	H ₆ BaIrK	ICSD 262197	0.003956
cd251c1afc	Ca ₄ P ₂ Se	ICSD 686117	0.005293
8a2692b9bd	CFe ₂ LuSi	mp-1207076	0.022086
<i>matches to a disordered ICSD crystal</i> (774 in GNoME2, 266 of them new)			
d1448232be	Os ₃ V ₁₃	ICSD 647852, Os _{0.4} V _{1.6}	0.000762
6d759f71af	Al ₈ DyEr ₃	ICSD 291821, AlDy _{0.2} Er _{0.3}	0.001673
1d6f38bd7b*	AsSSm ₂	ICSD 611321, As _{0.55} So _{0.45} Sm	0.001980
c5b9a84fae	Al ₄ Co ₃ Fe	ICSD 606563, AlCo _{0.7} Fe _{0.3}	0.001992
4fef57a136	CrFe ₁₅	ICSD 692954, Cr _{0.04} Fe _{1.96}	0.002034
36817ff8d9	Al ₈ DyHo ₃	ICSD 291824, AlDy _{0.2} Ho _{0.3}	0.002052
8d0043ba5a	Be ₅ Co ₃	ICSD 058694, Be _{1.34} Co _{0.66}	0.002201
8666ed5124	Al ₈ Dy ₃ Er	ICSD 291819, AlDy _{0.4} Er _{0.1}	0.002380
67e4bb6449	Al ₈ Dy ₃ Ho	ICSD 291823, AlDy _{0.4} Ho _{0.1}	0.002558
1c3630d10b	Fe ₁₅ V	ICSD 634026, Fe _{1.92} V _{0.078}	0.002579
6fbb16782c	Nb ₅ Re ₃	ICSD 105210, Nb _{1.2} Re _{0.8}	0.003143
a4de687109*	Co ₁₁ O ₁₆ Zn	ICSD 073754, Co _{0.605} OZn _{0.145}	0.003206

its element set, and the twenty nearest sharing its element set and having close enough composition. The third list is the only route by which an ordered supercell of a disordered crystal reaches the comparison, as its ADA is not similar to any one configuration.

2. *PDA comparison of the candidates.* Each structure is compared by its PDA[100]. Every candidate pair within an ADA gate of 0.05, widened to 0.2 for a same-composition candidate, is compared for d_{geom} and d_{chem} . Pairs with the same composition where the reference has partly vacant sites has d_{fw} computed as well.
3. *Reading the EMD plan site by site.* For each of the 3,280 pairs whose reference is disordered, shares the entry’s element set and composition, and has d_{chem} within 0.2, the chemically constrained plan is re-read: for every mixed site of the reference that receives flow, which elements the entry places there and in what shares, and the distance between the closest two rows of different elements sitting on it. Those are the *witness* distances of the following subsection.
4. *Per-entry distances and categories.* A pair is a chemical candidate when the entry shares the reference’s element set and composition. Such a pair scores d_{chem} and,

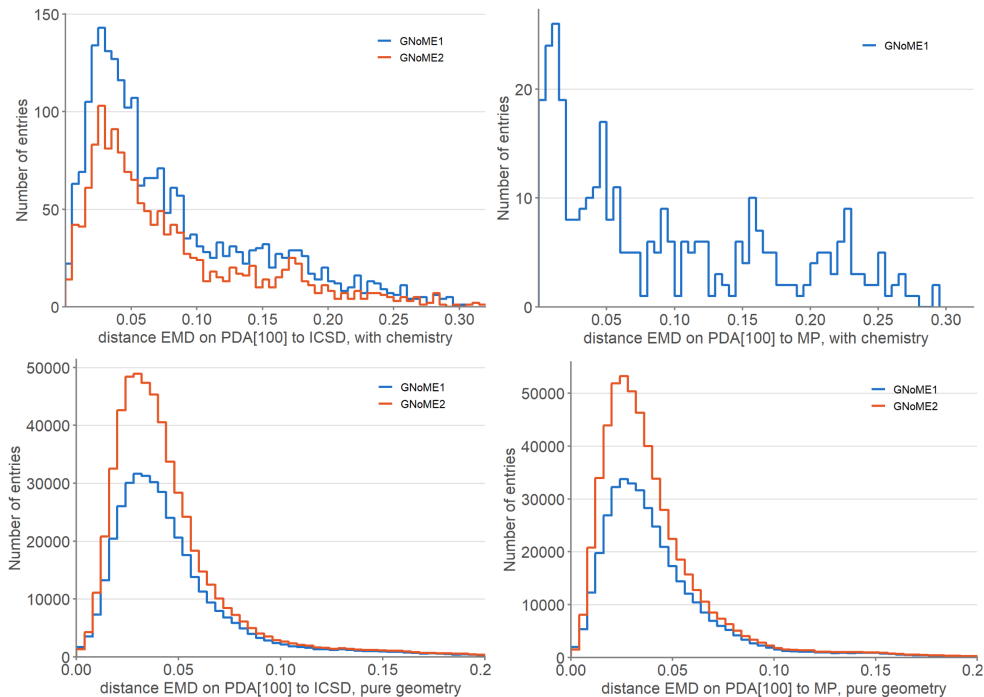


Fig. 12 Histograms of distances based on EMD between invariants PDA[100], for predictions in GNoMEs 1&2 to the nearest neighbours among known crystals in the ICSD and MP. **Left:** ICSD. **Right:** MP. **Top:** distance d_{chem} matches chemistry, see Table 18. The curve on the right is for GNoME1 only, because GNoME2 has no chemical matches in the MP. **Bottom:** distance d_{geom} .

where the reference has partly vacant sites and the surviving framework (after deleting vacancies) retains at least half of its atoms, $\min(d_{\text{fw}}, d_{\text{chem}})$. Each entry then takes the smallest such distance over all of its pairs, separately for ordered and for disordered references, and is assigned the first category that comes within τ : *ordered*, then *disordered*, then *geometry only* at d_{geom} , otherwise none. All disordered matches are combined into one *disordered* category for the results section.

B.4 Details of the most striking examples in GNoMEs 1&2

This section provides more details about invariants that detect outliers in GNoMEs, most of which seem to be physically unrealistic. The two contact bounds d_{min} and d_{max} are not equally decisive. GNoME2’s largest d_{max} is 5.29Å and the next largest is 4.63Å, so the 50 GNoME1 entries above it, running up to 8.92Å for entry eed3027c1e of composition KGd(Cu₂S)₂, are well-separated from GNoME2. The lower contact bound is not like this. GNoME2’s smallest d_{min} is 0.7482Å, and apart from the two Na₄K entries with atoms on the same positions in [19, Fig. 15], the whole “atoms too close” set is 21 hydrides sitting between 0.7460 and 0.7481Å: H₆Fe, H₈Cr, H₈Co, H₈Cu, H₈Be, H₈Ga etc, all of them the H–H separation of a hydrogen molecule (0.74Å). Both

Table 18 Comparison screening steps. Each establishes something the one above it cannot, and the third column gives the case that forces the next. Distances are RMS PDA[100] EMD in Ångströms. The last 3 are combined in results into “disordered”.

Screen	Establishes	Why it is not sufficient alone
d_{geom} <i>geometry only</i>	The two point sets agree as geometry.	It says nothing about chemistry.
d_{chem} , ordered reference <i>ordered</i>	A transport plan exists that moves mass only between the same elements.	Sufficient here, and sharing an element set is not: CoFe ₂ (95369b14fd) aligns with Co ₅ Fe ₁₁ (ICSD 155842) to 0.0053, but 56% of the mass would have to cross Co $\leftrightarrow$ Fe, so $d_{\text{chem}} = \infty$.
d_{chem} , disordered reference	The same, with a mixed reference site accepting flow from any element that occupies it.	Necessary but weak because a mixed site accepts either element. B ₂ FeNb ₂ Ru ₅ (dbf7583a57) is distance 0.0059 from B ₂ Fe _{0.83} Nb _{2.17} Ru ₅ (ICSD 195034), sending only Fe to the (Fe _{0.83} Nb _{0.17}) site.
witness, or every mixed site resolved <i>ordering</i>	Two rows of the entry carrying different elements reach one mixed site, so exchanging their labels leaves the PDA unchanged; the pair has distance $\max(d_{\text{chem}}, \text{witness})$. Failing that, every mixed site takes a single element.	Needed because the shares can be right and the geometry wrong: B ₆ NbNi ₂₀ U ₂ (f82eb9dc20) puts Nb _{0.5} U _{0.5} on the (Nb _{0.534} U _{0.466}) site of ICSD 011919, yet those two rows lie 0.414 apart. The across-sites form is weaker: with the two elements on inequivalent positions there is no label swap to appeal to, only the composition gate.
d_{fw} <i>framework</i>	For a reference with partially occupied sites: the vacancy atom types are deleted from both crystals and the surviving ordered sublattice compared.	It is not a relaxation of d_{chem} but the only finite answer for these, since deleting an atom perturbs every neighbouring row. Guarded by requiring the framework to keep half the reference’s atoms: stripping O and Gd from Gd ₆ O ₁₂ U (ICSD 021945) leaves one U atom of 19.

GNoME 1&2 have many such contacts. GNoME1’s just fall a few thousandths of an Å on the wrong side of GNoME2’s own extremum. All 23 were removed regardless.

All 121 entries are listed in the supplementary file `implausible_structures.csv` with the bound each one violates. Of the 121 implausible removed structures, 44 went two days after the initial release, and the remaining 77 nine months later. The 12/1/23 step is striking in that it withdrew only 67 entries, and 44 of them are entries these bounds independently flag. Both Na₄K entries, the only structures with superimposed atoms, went in the first wave, while 21 of the 23 hydrides survived. The emptiness flags divided more evenly in the sense that it did not specifically target empty cells, and the entries it did remove were the more extreme ones: their median point density is 5.96 nm⁻³ against 12.09 nm⁻³ for those that lasted until 8/21/24.

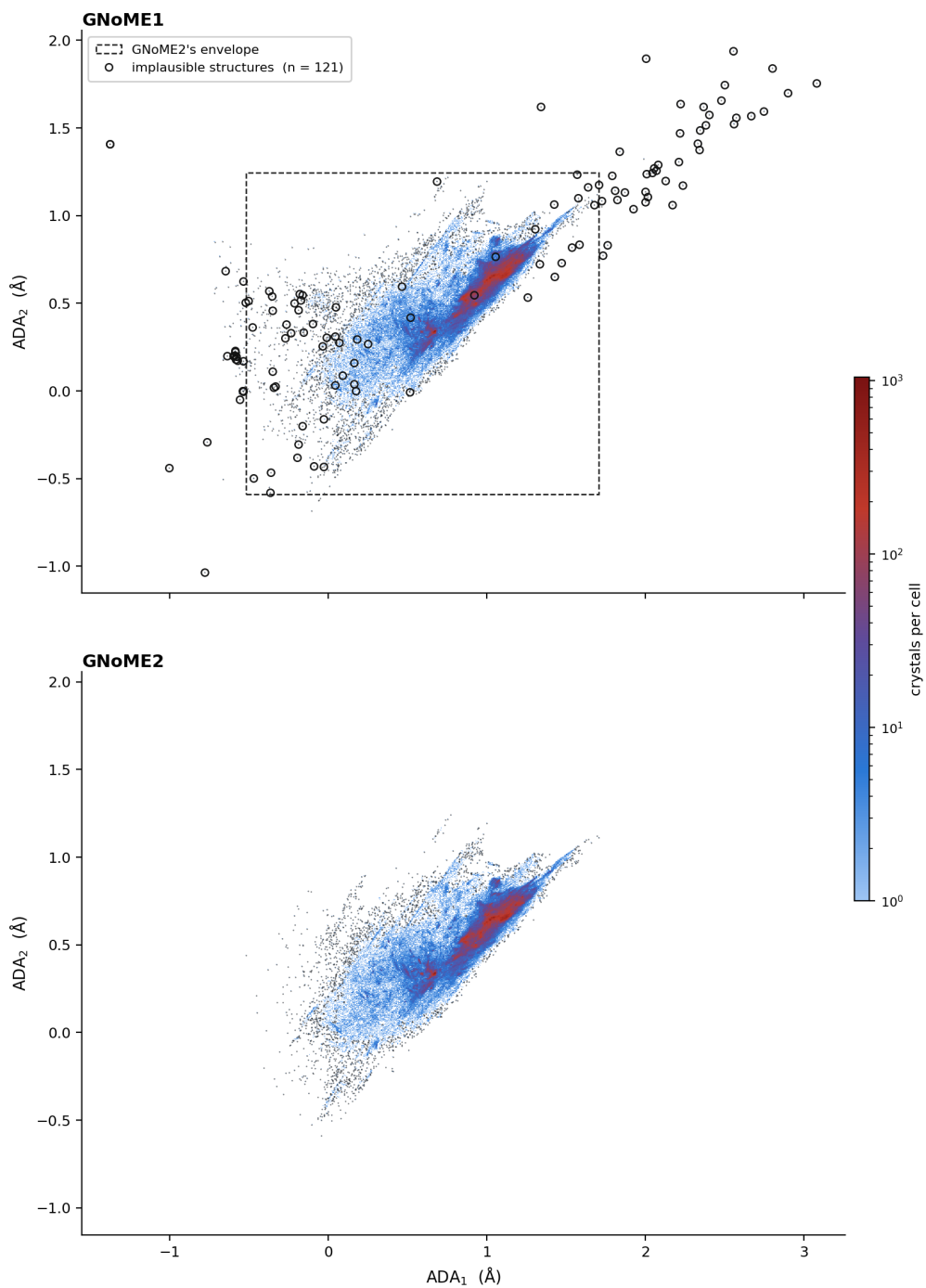


Fig. 13 GNoMEs 1&2 are plotted on the axes ADA_1 vs ADA_2 , with the same scale. The most unusual entries that were removed are small circles sitting far from the main cloud in GNoME1.

Table 19 The ten sparsest and the five most overlapped GNoME1 entries. All were removed from GNoME2. Here ρ is the point density, and $d_{\min}$ and $d_{\max}$ are the smallest and largest distances between any atoms. The median GNoME1 entry has $\rho = 52.2 \text{ nm}^{-3}$. The two Na₄K entries are the same structure twice.

entry	formula	$\rho \text{ (nm}^{-3}\text{)}$	$d_{\min} \text{ (Å)}$	$d_{\max} \text{ (Å)}$	density (g/cm ³)
<i>sparsest</i>					
88e5e5c7a3	S ₇ Gd ₂ Tb ₃	1.29	4.42	8.30	0.18
c247c290e6	Cu ₃ Sn ₂ I ₉ U ₂	1.73	6.90	7.58	0.37
37f9f79871	P ₃ FeCo ₂ U ₂	1.81	7.87	8.48	0.28
ea4f390476	Ge ₂ GdPt ₂ Th	1.94	6.62	7.73	0.50
ce080d08df	OCu ₂ La ₉ Pb ₆	2.31	4.34	8.15	0.56
f4614ac0c2	Ni ₃ In ₃ CeGd ₂	2.32	7.04	8.25	0.42
784a1426b2	Mg ₂ Ga ₃ GdTb ₃	2.38	7.09	8.11	0.39
bd207edbb8	S ₉ Gd ₂ Tb ₂	2.40	3.75	7.39	0.28
ca5372206f	SnSmGdAu ₂	2.49	6.61	7.93	0.68
cb590a0692	Ga ₄ Nd ₄ GdPt ₃	2.66	6.84	7.94	0.59
<i>most overlapped</i>					
0e2d8f26d6	Na ₄ K	21.9	0.000	4.20	0.95
cdc06a1a2a	Na ₄ K	21.9	0.000	4.20	0.95
98cb8c7d70	H ₈ Li ₂ CsIr	71.8	0.746	2.96	3.45
a1590781f0	H ₆ Fe	116.4	0.746	1.74	1.71
ba6e08bb47	H ₈ Cr	116.1	0.746	1.80	1.29

Table 20 GNoME1 entries falling outside the range GNoME2 occupies. Each bound is GNoME2’s own extremum, so no threshold is chosen by hand. The categories overlap, 41 of the 121 entries violate two bounds and 13 violate three, so the counts sum to more than the union. The background removal rate is 21.7% (83,525 of 384,938).

violated bound	criterion	entries in GNoME 1	removed (%)
too sparse	$\rho < 12.61 \text{ nm}^{-3}$	94	100.0
atoms too close	$d_{\min} < 0.7482 \text{ Å}$	23	100.0
atoms too far	$d_{\max} > 5.29 \text{ Å}$	50	100.0
density too low	$< 0.49 \text{ g/cm}^3$	21	100.0
any of the above		121	100.0

GNoME ID	ICSD ID	GNoME composition	ICSD composition	EMD on PDA[100]
cdc06a1a2a	138567	KNa ₄	I3V	0.649084
0e2d8f26d6	138567	KNa ₄	I3V	0.649084
eed3027c1e	281367	Cu ₄ GdKS ₂	Y	0.623388
3e5dc620b0	27030	GdIn ₂ Pt ₄ Rh	S	0.584775
ce080d08df	20564	Cu ₂ La ₉ OPb ₆	CoK	0.565175

Table 21 Top 5 predicted crystals S in GNoME1 that maximise the Local Novelty Distance $\text{LND}(\text{EMD}, \text{PDA}[100], \text{ICSD})(S)$ to their nearest neighbour in the ICSD. The first two predictions cdc06a1a2a and 0e2d8f26d6 include two pairs of repeated atoms, see [19, appendix A, Fig 15]. The 3rd prediction eed3027c1e has the largest $d_{\max} = 8.92 \text{ Å}$ in Fig. 7 (left). The next two predictions in Fig. 15 also have no close atomic contacts.

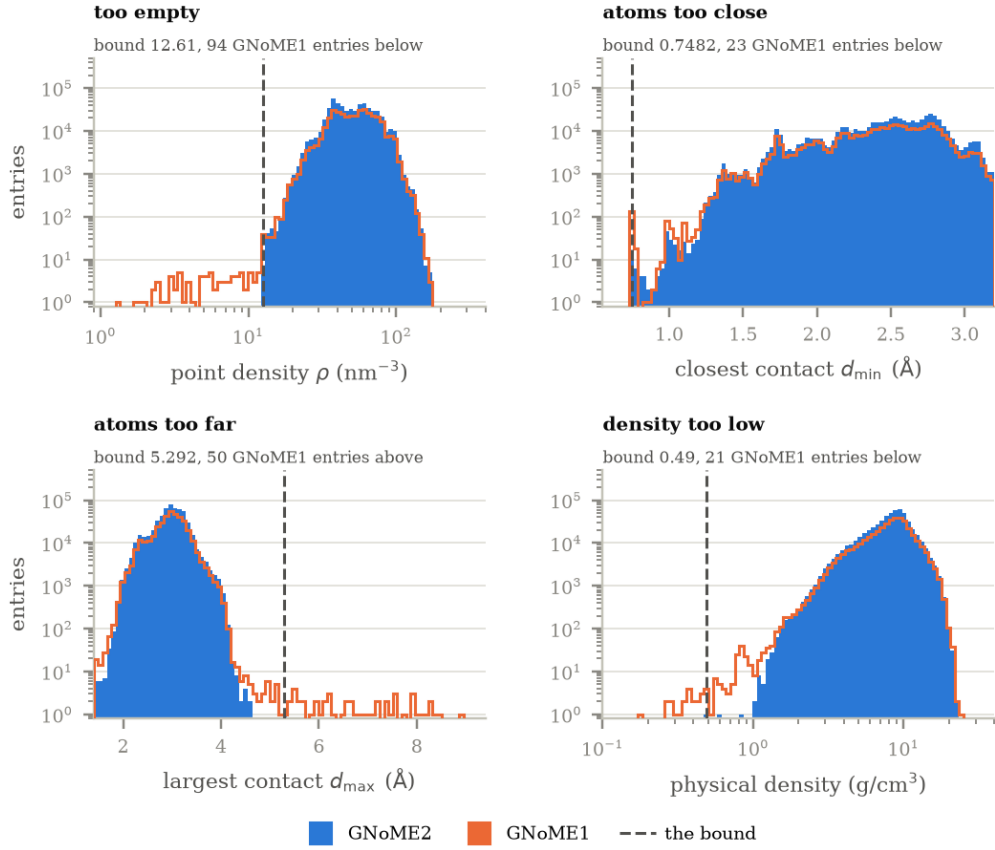


Fig. 14 The distribution of entries in GNoME1 and GNoME2 by the four values used to determine implausible structures in Table 20.

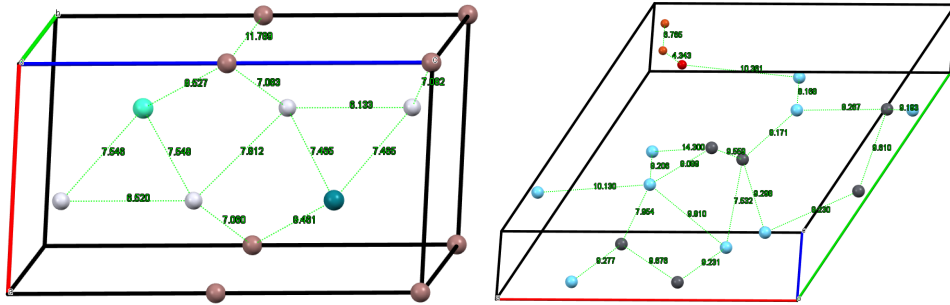


Fig. 15 Unrealistic predictions in GNoME1 with large values of LND (EMD, PDA[100], ICSD) in Table 21 and without close atomic contacts. **Left:** entry 3e5dc620b0. **Right:** entry ce080d08df.

C More details of software for crystal disguising

The second author developed a web application in Python called “Disguise a Crystal”. The underlying white-box algorithm can generate 2.2 million new-looking crystals in

about 24 hours on a modest laptop with the power of school-level geometry rather than AI. With the press of a single button, anyone can disguise an existing crystal so that its Crystallographic Information File (CIF) looks completely different. After “disguising” the CIF, most if not all the numbers in the file change and it is difficult to manually spot that the crystal it represents is either an exact or near-duplicate.

Filters that do not use Lipschitz continuous metrics cannot recognise the disguised CIF as a near-duplicate reliably. For example, Mercury’s Crystal Packing Similarity tool in the 2024.2 CSD Release often crashes when attempting to compare the original file and its disguise, presumably due to the exponentially many bijections between chemically identical atoms. This problem for continuously comparing finite atomic configurations under rigid motion has been resolved in polynomial time in [68].

The “Disguise a Crystal” program motivates the necessity for rigorous and computationally fast methods for crystal comparison. It is important to analyse new creations to see which are the same and which are different. It is not enough to generate millions of files using AI and claim that they are crystals that are new or in fact real.

The simplest trick is to randomly permute (lines of) atoms in a given CIF, which already creates exponentially many different representations of crystals with indistinguishable atoms. Any attempt to order atoms by coordinates breaks down continuity under small perturbations. For example, a lexicographic order of triples (x, y, z) requires that $(0, 0.3, 0.6) < (0, 0.3, 0.9)$, but a slight change in the y coordinate gives the opposite order: $(0, 0.29, 0.9) < (0, 0.3, 0.6)$.

Even if all given atoms are unique, an initial cell can be transformed by a random 3×3 matrix that has integer coefficients and determinant 1. The resulting skewed cell can be arbitrarily extended by any integer factors along the new basis vectors. This extension also enlarges a motif of atoms, which complicates any further processing without changing the underlying structure (a class under rigid motion). In theory, the two transformations above can be neutralised by reducing the cell to (say) Niggli’s form. In practice, any slight perturbation of at least one atom in the new cell (and all its repeated copies by lattice translations) makes the resulting extended cell primitive, and any reduction of the new cell has the same arbitrarily increased volume.

Even worse, if any chemical elements are replaced without changing their geometric positions or by only tiny perturbations, the new CIF can look completely different from the original. If we extend a given unit cell by a factor of 1 or 2 along each of the three basis vectors ($2^3 - 1 = 7$ non-trivial extensions) and then independently shift 10 times one of (say) 10 atoms in the extended cells, we can get 7×10^{10} (70 billion) numerically different CIFs, even without extra disguises such as permuting atomic rows and changing chemical elements.

Replacing a chemical element with its neighbour in the same group of Mendeleev’s table might only slightly affect structure factors if such a crystal can really be synthesised. The resulting crystal can look novel and realistic despite being a disguised near-duplicate of a previously known material.

C.1 How to Disguise your own Crystal

You can run the program on your own computer by downloading the files from the GitHub Repository [47]. The 'readMe' file guides you on how to download and run the program for Windows and Mac.

Disguise a Crystal

This is a web app that can generate a new-looking Crystallographic Information File (CIF) based on any input CIF!

Instructions:

- 1) Use the on/off toggles to select which transformations you would like to apply to your input CIF.
- 2) Scroll below the blue table and input your CIF in the "Input" column on the left-hand side.
- 3) Press the "Generate new CIF" button in the "Output" column.

You can download the new CIF using the "Download CIF" button.

On/Off	Transformation	Advanced Information & Modifications									
☒	Extend unit cell	Modify max extension scale factor: 2 - + Scale factors used in generation (will display once file is generated): cell_length_a cell_length_b cell_length_c									
☒	Shift atoms slightly	Modify the upper bound of shift: (angstroms) 0.01 - + Toggle to shift all atoms: ☒ Or pick number of shifts: 0 - +									
☒	Shuffle atom order	☐ Randomly shuffles the order of the atoms in the motif, then updates the labels.									
☒	Transform the basis	Modify the level of basis transformation: 1 - + The matrix actually applied to the RHS of basis: <table border="1"><tr><td>1</td><td>0</td><td>0</td></tr><tr><td>0</td><td>1</td><td>0</td></tr><tr><td>0</td><td>0</td><td>1</td></tr></table>	1	0	0	0	1	0	0	0	1
1	0	0									
0	1	0									
0	0	1									

Input:

Drag and Drop / Select Files

☐ Or use an example crystal instead

☒ Or use a different example crystal (NaCl)

File Name: example_NaCl.cif

CIF contents:

```
# generated using pymatgen
data_NaCl
_symmetry_space_group_name_H-M 'P 1'
_cell_length_a 5.58812644
_cell_length_b 5.58812644
_cell_length_c 5.58812644
_cell_angle_alpha 90.00000000
_cell_angle_beta 90.00000000
_cell_angle_gamma 90.00000000
_symmetry_Int_Tables_number 1
_chemical_formula_structural NaCl
_chemical_formula_sum 'Na4 Cl4'
_cell_volume 174.50130186
_cell_formula_units_Z 4
loop_
_symmetry_equiv_pos_site_id
```

Output:

Generate new CIF

Download CIF

Rename block in CIF (optional)

```
data_disguised
_symmetry_space_group_name_H-M 'P 1'
_cell_length_a 46.08087112
_cell_length_b 15.8056084
_cell_length_c 27.37611679
_cell_angle_alpha 106.77865488
_cell_angle_beta 153.01593954
_cell_angle_gamma 46.68614334
_chemical_formula_sum 'Na32 Cl32'
_cell_volume 1396.01041831

loop_
_symmetry_equiv_pos_site_id
_symmetry_equiv_pos_as_xyz
1 'x, y, z'

loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
```

Fig. 16 Screenshots of the user interface and example inputs.

First, select which settings you would like to run the program with and which transformations you would like to apply to disguise the crystal using the interactive buttons and input fields in the blue table in the User Interface.

Next, upload your own CIF to the Input column below the table or use the examples provided and once you press the “Generate CIF” button, the output CIF will be displayed and be available to download.

C.2 Algorithm Details

The algorithm inputs a correctly formatted CIF (Crystallographic Information File) [69] with unit cell information and coordinate data listed in the first block. The crystal in the CIF may be in any space-group but should have no disorder which will be ignored by the program. The algorithm consists of 4 independent transformations:

1. *Extension of the Unit Cell.* The program randomly picks 3 scale factors up to the upper bound (2 by default) and extends each unit cell length correspondingly. All atoms in the original motif are duplicated accordingly to fill the new, larger unit cell.
2. *Tiny perturbations of atom coordinates.* Atoms are converted to Euclidean Coordinates, and are shifted by a vector which is sampled uniformly from a 3D ball using the algorithm from [70, section 3.29]. The default maximum shift is 0.01 Angstroms but this can be changed in the “Advanced Information & Modifications” section.
3. *Change of basis.* The unit cell information is orthogonalised i.e. converted into 3 basis vectors, b_1, b_2, b_3 in the Euclidean Space, $\mathbb{R}^3$, that define the unit cell. The basis vectors are then transformed by applying a generated 3 by 3 matrix on the right. The generated matrix has determinant 1 and is formed by first randomly generating $a, b, c, d, e, f \in \mathbb{Z}$ in the range $(-k, k)$ where k is the level of basis transformation that can be modified in settings, and then forming the matrix:

$$\begin{pmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} 1 & 0 & 0 \\ d & 1 & 0 \\ e & f & 1 \end{pmatrix} = \begin{pmatrix} 1 + ad + be & a + bf & b \\ d + ce & 1 + cf & c \\ e & f & 1 \end{pmatrix}$$

The new basis is:

$$\begin{pmatrix} | & | & | \\ b_1 & b_2 & b_3 \\ | & | & | \end{pmatrix} \begin{pmatrix} 1 + ad + be & a + bf & b \\ d + ce & 1 + cf & c \\ e & f & 1 \end{pmatrix}$$

The new fractional coordinates are found by applying the inverse of the generated matrix on the left of the original fractional coordinate. If the original fractional

coordinate (x, y, z) , the new coordinate vector is:

$$\begin{pmatrix} 1 + ad + be & a + bf & b \\ d + ce & 1 + cf & c \\ e & f & 1 \end{pmatrix}^{-1} \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

This way, the euclidean coordinates remain the same while the unit cell lengths and angles, and fractional coordinates are different.

4. *Shuffling of the atomic order.* The atoms in the unit cell are arbitrarily ordered and hence can be randomly shuffled. Atomic labels are updated so that the labels are still listed chronologically in the final file.

The output of the program is a correctly formatted CIF that can be processed by current crystal software that handle CIFs.

If the default transformations were applied, most if not all the numbers in the output file will be different from the numbers in the input file. However, the crystal structure represented will be either the same geometrically or very slightly different.

If no perturbations were applied, then the Earth Mover's Distance (EMD) on the Pointwise Distance Distributions (PDDs) of the input and output files will be 0, or less than 10^{-4} Angstroms, which is due to floating point calculations and is insignificant [23]. If perturbations up to 0.01 Angstroms were applied to all atoms, the maximum EMD is 0.02 Angstroms.

All the elements are kept the same, so if the input CIF represents a real crystal, the output CIF is guaranteed to represent a real crystal too.

For full transparency, comments on lines starting with # are added to the bottom of the input CIF that list the settings of the program when the file was generated and also the transformation matrix and the extension that took place.