Continuous invariant-based maps of the Cambridge Structural Database

Daniel E Widdowson and Vitaliy A Kurlin*

Materials Innovation Factory and Department of Computer Science, University of Liverpool, Liverpool L69 3BX, United Kingdom

E-mail: vitaliy.kurlin@liverpool.ac.uk

Phone: +44 (151) 7958861

Abstract

The Cambridge Structural Database (CSD) played a key role in the recently established Crystal Isometry Principle (CRISP). The CRISP says that any real periodic crystal is uniquely determined as a rigid structure by the geometry of its atomic centres without atomic types. Ignoring atomic types allows us to study all periodic crystals in a common space whose continuous nature is justified by the continuity of real-valued coordinates of atoms. Our previous work introduced structural descriptors PDD (Pointwise Distance Distributions) that are invariant under isometry defined as a composition of translations, rotations, and reflections. The PDD invariants distinguished all non-duplicate periodic crystals in the CSD. This paper presents the first continuous maps of the CSD and its important subsets in invariant coordinates that have analytic formulae and physical interpretations. Any existing periodic crystal has a uniquely defined location on these geographic-style maps. Any newly discovered periodic crystals will appear on the same maps without disturbing the past materials.

Introduction: strong motivations for continuous maps

Crystallography traditionally classified periodic crystals almost exclusively in a discrete way by symmetries. This was a natural approach in the past when only a few crystal structures were known. The classification of 230 space groups was a great achievement in the 19th century by Fedorov[1] and Schonflies[2] in 1891. Due to the important work of Olga Kennard, who established the Cambridge Structural Database (CSD) in 1965, and her numerous successors,[3] the CSD contains now 1.25+ million known materials including more than 830 thousand periodic crystals with no disorder and full atomic geometry. Many more thousands of crystal structures are computationally predicted even for a fixed chemical composition.[4] These big numbers motivate a finer (stronger) classification into infinitely many classes.

For a simple comparison, triangles can be classified by symmetries as equilateral, isosceles, and generic (nonisosceles). However, geometry did not stop at these three classes and moved on to a much stronger classification, which is now called the side-side-side theorem. This SSS theorem from school geometry says that two triangles can be rigidly matched if and only if they have the same triple of side lengths a, b, c considered up to all permutations.

In Euclidean space $\mathbb{R}^n$, this rigid matching of triangles is called a *congruence* and can be obtained as a restriction of *isometry*, which is any distance-preserving transformation of $\mathbb{R}^n$. Any Euclidean isometry is a composition of translations, rotations, and reflections, as seen in Fig. 1. If we exclude reflections, any composition f of translations and rotations is called a *rigid motion* because f can be included in a continuous family (motion) $f_t : \mathbb{R}^n \to \mathbb{R}^n$, $t \in [0, 1]$, of isometries such that $f_1 = f$ is the given map and f_0 is the identity map.

All rigid shapes of triangles form one infinitely continuous *moduli* space of classes modulo isometry. The SSS theorem can be rephrased so that the moduli space of triangles is continuously parametrised by an ordered triple of interpoint distances a, b, c satisfying $0 < a \le b \le c \le a + b$, where the last inequality guarantees the existence of a triangle, e.g. $(a, b, c) = (3, 4, 5)$ uniquely determines a right-angled triangle with side lengths $3, 4, 5$.

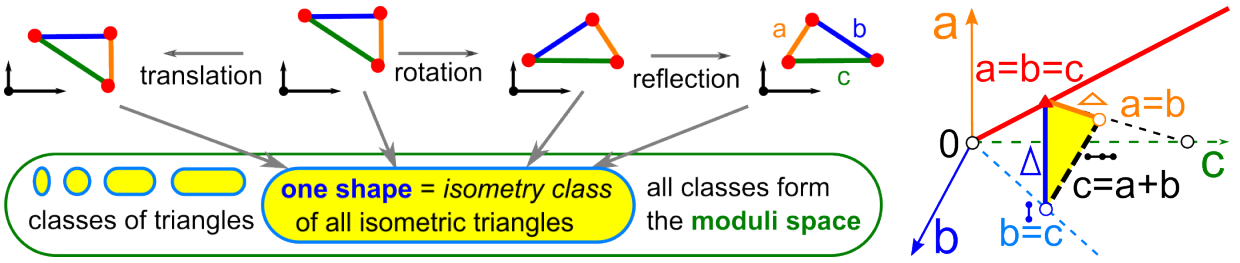

Figure 1: **Left**: one isometry class consists of all triangles isometric to each other. All such classes form the *moduli* space of shapes. **Right**: the moduli space of triangles under isometry is the cone $\{0 < a \le b \le c \le a + b\} \subset \mathbb{R}^3$ parametrised by 3 interpoint distances a, b, c.

Fig. 1 visualises this moduli space as a triangular cone in $\mathbb{R}^3$ with the coordinate axes a, b, c. The red diagonal $\{a = b = c\}$ represents all equilateral triangles. Any point in the yellow section determines a unique triangle under isometry and uniform scaling. The dashed line in the plane $c = a + b$ represents degenerate triangles of 3 points in a straight line.

The moduli space of triangles under rigid motion and uniform scaling is the union of two yellow triangles glued along their common boundaries, where triangles are mirror-symmetric, which gives a topological sphere S^2. This continuous classification of triangles under rigid motion (or slightly weaker isometry, possibly with uniform scaling) is much finer (stronger) than the discrete classification by symmetries with only three classes. The classes of all equilateral and isosceles triangles are low-dimensional subspaces of dimensions 1 (diagonal line) and 2 (union of two boundary sides) in the infinite 3-dimensional cone, respectively.

Because crystal structures are solid (or rigid) materials, their most natural equivalence is *rigid motion*. Indeed, there is little sense in distinguishing crystals related by rigid motion, at least under the same ambient conditions such as temperature and pressure. Hence the crucial question "same or different" [5] has the initial answer *same* (rigidly equivalent) if they are related by rigid motion. The more practical questions are "how to distinguish all different crystals" and "how to continuously quantify their difference". As shown in Fig. 1, such answers for triangles were known already to Euclid in 300 BC. For instance, the distance between any triangles uniquely represented by triples (a, b, c) and (a', b', c') can be quantified in many continuous ways, the simplest being the Euclidean metric $\sqrt{(a' - a)^2 + (b' - b)^2 + (c' - c)^2}$.

Can we rigorously answer the fundamental questions "same or different?" and "if different, how much different?" at least for periodic crystals? The 21st century witnessed an explosive growth of structural databases whose integrity maintenance[6] now requires continuous tools that can quickly detect numerous near-duplicates as motivated below.

Discontinuity challenge of traditional crystallography

This section explains the discontinuity of traditional Definition 1 as in section 8.1.4 of ITA.[7]

Definition 1 (*unit cell*, *lattice*, *motif*, *periodic point set*, *periodic crystal*)**.** Any ordered basis of vectors $v_1, \dots, v_n \in \mathbb{R}^n$ defines the *unit cell* (parallelepiped) $U = \{\sum_{i=1}^{n} t_i v_i \mid 0 \le t_i < 1\} \subset \mathbb{R}^n$ and the *lattice* $\Lambda = \{\sum_{i=1}^{n} c_i v_i \mid c_i \in \mathbb{Z}\}$, as seen in Fig. 2. A *motif* $M \subset U$ is any finite set of points in U. A *periodic point set* $S = M + \Lambda$ is the infinite set of points $p + v$ for all $p \in M$ and $v \in \Lambda$. In $\mathbb{R}^3$, if each point of M is an atom or ion with a chemical element and charge, then the periodic point set S with these atomic attributes is called a *periodic crystal*. ▲

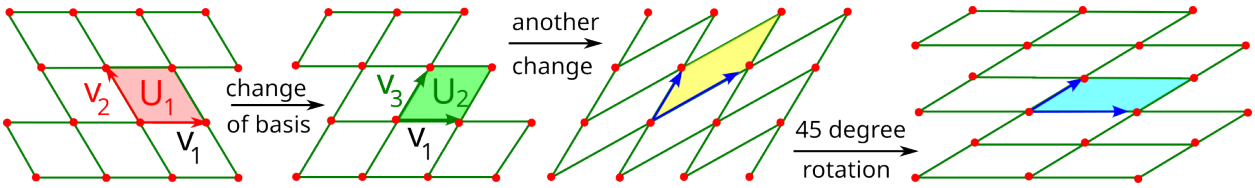

Figure 2: Ambiguity of choosing a basis of a lattice $\Lambda \subset \mathbb{R}^2$ as introduced in Definition 1.

Any lattice can be generated by infinitely many bases. If (v_1, v_2) is one basis of a lattice $\Lambda \subset \mathbb{R}^2$, then (Av_1, Av_2) is another basis of Λ for any 2×2 matrix A with integer elements and determinant 1. This ambiguity can be theoretically resolved by a reduced cell.[8] In dimensions 2 and 3, such a reduced cell can have two types: with all angles between basis vectors acute or with all angles obtuse (non-acute). The hexagonal lattice Λ in Fig. 2 has the obtuse and acute cells U_1, U_2, respectively. While we can choose one of them by allowing

the angle 60° and forbidding 120°, all such choices create the discontinuity under almost any perturbation, which was reported[9] in 1965 and resolved for 2D lattices[10] in 2022.

When a motif M is added to a lattice Λ, the ambiguity of this crystal representation $S = M + \Lambda$ significantly increases. Firstly, one can shift a motif within a fixed unit cell U, which changes all fractional coordinates of atoms in a basis of U, but moves the underlying periodic crystal only by a fixed vector. For highly symmetric crystals, this ambiguity is often resolved by fixing atoms at Wyckoff positions but not for generic crystals with the space group P1. Second, any periodic point set $S = M + \Lambda$ can be obtained from an extended motif M' in a scaled-up cell (defining a sublattice $\Lambda' \subset \Lambda$) so that $S = M' + \Lambda'$. In theory, an extended motif and cell can be scaled down to a *primitive* cell that has a minimum volume. In practice, a tiny atomic displacement can make an extended cell primitive.

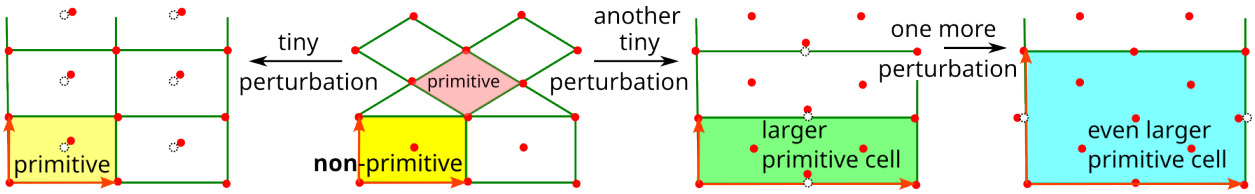

Figure 3: Any primitive or reduced cell arbitrarily extends under almost any perturbation.

So a Crystallographic Information File (CIF) describing a crystal S via a unit cell and motif as in Definition 1 can be considered a single "photograph" of S. The standard settings carefully developed in the International Tables for Crystallography can be informally compared with a standardised passport photo, which sufficed in the 20th century.

In 2024, massive data produced by cheap artificial tools[11] should be validated by rigorous methods[12] such as biometric passports for humans and DNA-style codes for crystals.

A potential attempt to ignore perturbations up to a small threshold $\varepsilon > 0$ by calling any ε-close crystals equivalent (pseudo-symmetric) practically shifts[13] the discontinuity from 0 to ε and theoretically leads to a trivial classification because any crystals can be connected through sufficiently many ε-perturbations. All challenges above for experimental crystals become much worse in simulations because near-duplicates can be easier hidden within big

data. Any iterative optimisation stops at a close approximation to a local minimum, hence many differently looking approximations can accumulate around the same local mininum.

Key concepts for identifying near-duplicate structures

A rigorous way to answer the question "same or different" is to define "same" (equivalent) crystals so that all their physical or chemical properties are equal. We consider only ideal periodic crystals under the standard ambient conditions such as temperature and pressure. Because crystal structures are determined in a rigid form, their strongest equivalence in practice is a *rigid motion*, which is a composition of translations and rotations. Indeed, different rigid structures (even different rigid conformations of the same molecule such as a protein) can have different chemical properties and hence are important to distinguish.

Definition 2 was also proposed in the paper[12] discussing past ambiguities and was motivated by Carolyn Brock's call[14] for "a real-space definition mentioning periodicity". We propose to separate the concepts of a *periodic crystal* (an object fixed in space and represented by a CIF) and a *crystal structure* (a class of all rigidly equivalent crystals). The word *crystal* refers to a 3D object with chemical attributes, while periodic *point sets* and their periodic *structures* (equivalence classes) are based on indistinguishable points in any $\mathbb{R}^n$.

Definition 2 (*periodic structure* and *crystal structure* are classes under rigid motion)**.** In $\mathbb{R}^n$, a *periodic structure* is a class of all periodic point sets that can be exactly matched to each other by rigid motions of $\mathbb{R}^n$. In $\mathbb{R}^3$, a *crystal structure* is a class of all periodic crystals that can be exactly matched to each other (with all atomic attributes) by rigid motions. ▲

The earlier attempt[15] to formalise an equivalence proposed 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. For instance, calcite $CaCO_3$, sodium nitrate $NaNO_3$ and iron borate $FeBO_3$ are isostructural". While the IUCr online dictionary

doesn't have the entry 'structure', the definitions of a crystal structure[16] (in $\mathbb{R}^3$) and a crystal pattern[17] (in any $\mathbb{R}^n$) essentially coincide with a periodic point set and a periodic crystal in Definition 1, respectively. Then any isostructural crystals should coincide as periodic sets of atoms with fixed positions in $\mathbb{R}^3$ without applying any rigid motion. Even if we interpret the crystal structure in the sense of Definition 2, Table 1 summarises how many versions of the above compounds differ by their cell lengths and are not equivalent under rigid motion.

Table 1: Examples of "isostructural" crystals as defined in the IUCr online dictionary. Even for a fixed composition, cell parameters of $CaCO_3$ vary up to 0.95Å (from $c \approx 16.86$Å in FUTWOI01 to $c \approx 17.81$Å in FUTWOI18) among 29 entries in the CSD. All entries in the ICSD are taken for the same space group (R-3cH), room temperature, and standard pressure.

crystal	database references	#	cell lengths $a = b$	cell length c	cell angles
$CaCO_3$	CSD: FUTWOI ...	29	$4.945 \le a \le 4.992$	$16.86 \le c \le 17.81$	90° 90° 120°
$NaNO_3$	ICSD: 14185, ...	10	$5.071 \le a \le 5.107$	$16.82 \le c \le 16.83$	90° 90° 120°
$FeBO_3$	ICSD: 34474, ...	12	$4.621 \le a \le 4.627$	$14.47 \le c \le 14.5$	90° 90° 120°

In Pattern Recognition,[18] the *pattern* means a class under some equivalence. The *structure* also deserves a deeper meaning as in Definition 2 because any rotation of a real crystal changes its CIF (coordinate representation) but preserves the underlying crystal structure.

So crystals are considered *same* (having the same rigid structure) if they can be exactly matched by rigid motion in $\mathbb{R}^3$. The slightly weaker equivalence is *isometry*, which is any distance-preserving transformation. In $\mathbb{R}^n$, any Euclidean isometry is a rigid motion or its composition with any mirror reflection. If we don't distinguish mirror images, any non-mirror-symmetric crystal defines a larger class under isometry than under rigid motion.

If crystals S, Q are *isometric* (matched by an isometry f of $\mathbb{R}^n$), they are rigidly equivalent (matched by rigid motion) or S is rigidly equivalent to the mirror image of Q. One can separate these cases by checking if f preserves the sign of the $n \times n$ determinant consisting of basis vector $v_1, \dots, v_n$ in $\mathbb{R}^n$. So it almost suffices to classify crystals under isometry.

To distinguish *nonisometric* crystals that are not related under isometry, we need an invariant. This concept makes sense for any equivalence, though we consider only isometry.

Definition 3 (*invariant* and *complete invariant*)**.** An isometry *invariant* I is a function on periodic point sets from Definition 1 such that if any S, Q are isometric (denoted by $S \simeq Q$), then $I(S) = I(Q)$. An isometry invariant is called *complete* (or *injective* or *separating*) if the converse holds for any periodic point sets, i.e. if $I(S) = I(Q)$, then $S \simeq Q$. ▲

The centre of mass of a motif is not invariant because shifting a motif within a unit cell moves the center of mass, so isometric crystals can have different values of any non-invariant. Only an invariant descriptor can distinguish crystals under isometry because Definition 3 implies that if $I(S) \neq I(Q)$, then $S \not\simeq Q$. The motif size (number of points in a primitive cell) and the primitive cell volume are isometry invariants but they are incomplete. Indeed, all lattices have motifs of one point and many of them have the same volumes of primitive cells despite being nonisometric. A complete invariant can be considered a DNA-style code or a materials genome that uniquely identifies any periodic crystal under isometry in $\mathbb{R}^3$.

Standard (or conventional) settings[19] for crystal representations were designed to be such a complete invariant, which worked well in the 20th century while structural databases were relatively small. In 2024, near-duplicates can be computer-generated in huge numbers and all represented with very different cells and space groups despite being almost identical. The perturbations of the hexagonal lattice in Fig. 3 can be similarly applied to any periodic crystal. Indeed, arbitrarily extend a given cell of any crystal and slightly shift a single atom within the initial cell, which makes the extended cell primitive. This discontinuity exists even without extensions for only lattices,[10] though examples become more complicated.

A rigorous way to quantify the closeness between near-duplicates is to use a continuous metric, which is a distance function between invariant values of crystals, as defined below.

Definition 4 (distance *metric*)**.** A *metric* on values of an invariant I of periodic point sets (under isometry) is a function d satisfying the following axioms:

(a) *coincidence* : $d(I(S), I(Q)) = 0$ if and only if $I(S) = I(Q)$;

(b) *symmetry* : $d(I(S), I(Q)) = d(I(Q), I(S))$ for any periodic point sets $S, Q \subset \mathbb{R}^n$;

(c) *triangle inequality* : $d(I(S), I(Q)) + d(I(Q), I(T)) \geq d(I(S), I(T))$ for any S, Q, T. ▲

The first coincidence axiom in Definition 4 guarantees that $d = 0$ if and only if $I(S) = I(Q)$. Without this axiom, even the zero function $d = 0$ satisfies all other axioms. If the triangle inequality is allowed to fail with any positive error, one can design a distance d such that the k-means and DBSCAN algorithms output are predetermined[20] and hence are not trustworthy. Hence any clustering should use a distance d satisfying all metric axioms.

For the invariant $I(S)$ equal to the primitive cell volume of S, the simplest metric is the absolute difference $|I(S) - (Q)|$. One can similarly define a distance metric for the complete invariant I consisting of a conventional representation $C(S)$ including atoms at Wyckoff positions in a reduced cell. All these cell-based metrics are discontinuous under almost any perturbation and hence fail to detect the closeness of infinitely many near-duplicates.

A conventional representation $C(S)$ can be used to detect an isometry $S \simeq Q$ and define a discrete metric such as $d(C(S), C(Q)) = \begin{cases} 0 \text{ if } C(S) = C(Q), \\ 1 \text{ if } C(S) \neq C(Q), \end{cases}$ but any such metric is discontinuous. Detecting near-duplicates needs a stronger concept of continuity below.

Definition 5 (*Lipschitz continuity*)**.** An invariant I of periodic point sets is called *Lipschitz continuous* in a metric d if there is a constant $\lambda > 0$ such that if a periodic point set $Q \subset \mathbb{R}^n$ is obtained from S by perturbing every point of S up to any fixed bound $\varepsilon \geq 0$ in the Euclidean distance, then the invariants of S, Q are close in the sense that $d(I(S), I(Q)) \leq \lambda\varepsilon$. ▲

Definitions 3, 4, 5 help state the continuous classification problem in crystallography.

Problem 6 (continuous isometry classification)**.** Find a complete, continuous and quickly computable isometry invariant I of all periodic point sets in $\mathbb{R}^n$. In detail, we need

(a) *completeness* : any periodic point sets are isometric ($S \simeq Q$) if and only if $I(S) = I(Q)$;

(b) *continuity* : I has a Lipschitz continuous metric d in the sense of Definition 5;

(c) *reconstruction* : any $S \subset \mathbb{R}^n$ can be reconstructed (uniquely under isometry) from $I(S)$;

(d) *computability* : for a fixed dimension n, the invariant I, the metric d, and a reconstruction of any $S \subset \mathbb{R}^n$ from $I(S)$ can be obtained in polynomial time of the motif size of S. ▲

The equality $I(S) = I(Q)$ between complete invariants is best checked as $d(I(S), I(Q)) = 0$ due to the coincidence axiom in Definition 4. The reconstruction in condition (c) is stronger than the completeness in (a) because the invariant I might be too complicated and unsuitable for inverse design of crystals. Conditions (a,b,c) in Problem 6 can be easily satisfied by the abstract invariant $I(S) = \{\text{all } Q \text{ isometric to } S\}$. The computability in condition (d) makes Problem 6 practically meaningful because the invariant I can be used as geographic-style coordinates on the space of isometry classes of all periodic crystals similar to a, b, c parametrising the space of triangles in Fig. 1. Problem 6 is even harder for rigid motion.

The progress in continuous classifications of crystals

This section reviews the recent progress in solving Problem 6 for periodic point sets under isometry, which can be replaced with any types of objects under any equivalence. One simple extension is to allow compositions of rigid motion or isometry with uniform scaling in $\mathbb{R}^n$.

The early statement of Problem 6 and a partial solution appeared for lattices[21] in 2020. Now Problem 6 is considered a crystallographic example of the general meta-problem in the new area of Geometric Data Science. The ultimate goal is to continuously parametrise the spaces of equivalence classes of data objects on a geographic-style map to visualise structure-property relations and enable an inverse design of new objects by rational exploration.

If our objects are finite sets of unordered points (say, atoms of a molecule) under isometry, Fig. 1 illustrates the full solution for $m = 3$ points but the problem for $m > 3$ was solved relatively recently: for nonsingular sets[22] in 2004 and completely[23] in 2023. An analogy is the human genome and DNA, whose structure is known[24] and is considered complete in practice, at least for identifying humans in court trials, though identical twins exist. However, a living organism cannot be easily reconstructed from its DNA yet. So an efficient reconstruction of a periodic set in Problem 6(c) is more challenging than completeness.

For all lattices $\Lambda \subset \mathbb{R}^2$, Problem 6 was solved[10] in 2022 (also under rigid motion) by

the complete root invariant RI(Λ) with two slight deviations. First, the extra *realizability* condition explicitly described what values of RI(Λ) can be realised by lattices. Second, if the bound ε on point perturbations is smaller than the minimum quarter-distance between lattice points, one can prove[25] that the perturbed lattice is a translate of the original one, which easily implies condition (b) in Problem 6. If we perturb not points as in Problem 6(b) but coordinates of basis vectors of Λ up to ε, the root invariant RI(Λ) changes up to $\sqrt{6l\varepsilon}$ in the Euclidean metric, where l is the maximum length of basis vectors. The stronger Lipschitz continuity (with $\lambda\varepsilon$ instead of $\sqrt{6l\varepsilon}$) seems unrealistic because the rectangular lattices with the ε-close bases $(l,0),(0,\varepsilon)$ and $(l,0),(0,2\varepsilon)$ can substantially differ even by unit cell areas $l\varepsilon$ and $2l\varepsilon$ whose difference $l\varepsilon$ can be arbitrarily large if l has no upper bound. For all lattices $\Lambda\subset\mathbb{R}^3$, conditions (a,c) in Problem 6 hold for the more complicated invariant[26] whose continuity is being finalised, based on five Voronoi types[27] instead of 14 Bravais classes.[28]

For general periodic point sets from Definition 1, a strong continuous invariant (the density fingerprint[25]) was obtained by extending the point density to k-fold intersections of balls of a variable radius centred at given points. This density fingerprint was proved to be complete for nonsingular periodic point sets (in a general position achieved by almost any perturbation) and Lipschitz continuous in $\mathbb{R}^3$, also computable in polynomial time[29] in dimensions 2 and 3, but its underlying metric so far has only an approximate algorithm. Later the density fingerprint was shown to be incomplete[30,31] even in dimension 1 for singular periodic sequences, which were distinguished by the invariants in Definition 7 below.

Definition 7 describes much simpler invariants whose slight modifications will be used as geographic-style coordinates on continuous maps of the CSD in the next section.

Definition 7 (distance-based invariants PDD)**.** Let $M\subset U$ be a motif of m points in a (not necessarily primitive) unit cell of any periodic point set $S\subset\mathbb{R}^n$. Fix an integer $k\geq 1$.

(a) For each point $p\in M$, the $m\times k$ matrix $D(S;k)$ has one row of k distances $d_1(p)\leq\cdots\leq d_k(p)$ to the k nearest neighbours of p in the full set S not restricted to any cell or a ball of a cut-off radius. If any $l>1$ rows of $D(S;k)$ coincide, collapse them into a single

row and assign the weight l/m. The resulting matrix with the extra first column of weights is called the *Pointwise Distance Distribution*[32] $\mathrm{PDD}(S;k)$. The *Average Minimum Distance* $\mathrm{AMD}_k(S)$ is the weighted average of the distances to the k-th neighbours, see Fig. 4.

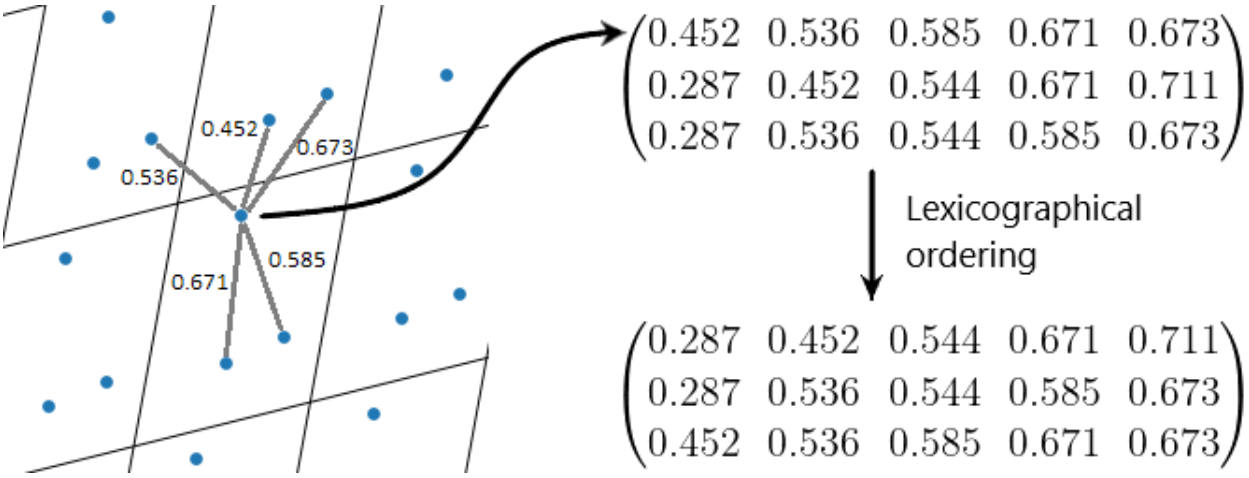

Figure 4: Distances to k nearest neighbours in a periodic point set. The lexicographic order is for convenience. The PDD matrices in Definition 7 are compared with unordered rows.

(b) Let V_n be the volume of the unit ball in $\mathbb{R}^n$. The *Point Packing Coefficient* $\mathrm{PPC}(S) = \sqrt[n]{\dfrac{\mathrm{vol}(U)}{mV_n}}$ is the average volume per point, measured in original units such as angstroms.

(c) The *Average Deviation from Asymptotic* $\mathrm{ADA}_k(S) = \mathrm{AMD}_k - \mathrm{PPC}(S)\sqrt[n]{k}$ has original units. The *Normalised Deviation from Asymptotic* $\mathrm{NDA}_k(S) = \dfrac{\mathrm{ADA}_k(S)}{\mathrm{PPC}(S)}$ is unitless. ▲

Fig. 4 illustrates the PDD computation and highlights the fact that all k neighbours are not restricted to a finite subset whose change may disrupt the output. For any $k \geq 1$, $\mathrm{AMD}_k(S)$ is the weighted average of the $(k+1)$-st column of $\mathrm{PDD}(S;k)$ with the weights from the extra first column, e.g. $\mathrm{AMD}_1(S)$ is the average distance to the first neighbour.

Increasing k only adds more columns of distances to PDD without changing the previous distances. Hence k is considered not a usual parameter that can substantially affect the result but as a degree of approximation similar to the number of decimal places on a calculator.

The Point Packing Coefficient $\mathrm{PPC}(S) = \sqrt[n]{\dfrac{\mathrm{vol}(U)}{mV_n}}$ measures (the n-th root of) the unit cell volume per point normalised by the unit ball volume V_n. Roughly speaking, $\mathrm{PPC}(S)$

inversely proportional to the density of points. Theorem 13 in the AMD paper[33] proved that the curve of values $\mathrm{AMD}_k(S)$ approaches $\mathrm{PPC}(S)\sqrt[n]{k}$ as $k \to +\infty$. Hence the limit behaviour of $\mathrm{AMD}_k(S)$ is largely determined by density and there is no need to substantially increase k because the most descriptive information is contained in smaller atomic environments. This asymptotic motivated the modified invariant $\mathrm{ADA}_k(S)$ in new Definition 7(c). If a periodic point set $S \subset \mathbb{R}^n$ is uniformly scaled by a factor $u > 0$, all interpoint distances and hence $\mathrm{AMD}_k(S)$ and $\mathrm{PPC}(S)$ are multiplied by u, which leaves $\mathrm{NDA}_k(S)$ invariant.

If we collect all distances from $\mathrm{PDD}(S;k)$ into a single distribution, we get a raw version of the Pair Distribution Function (PDF) as a discrete set of all (infinitely many) interatomic distances. This discrete set can discontinuously change under perturbation when two equal distances become slightly different. In the past, this discontinuity was addressed by taking a convolution with a Gaussian kernel, which converts the discrete PDF into a smooth function. The AMD and PDD invariants resolve this discontinuity at the level of discrete invariants by using Earth Mover's Distance without an extra Gaussian deviation parameter.

Writing all distances per point in the Pointwise Distance Distribution makes $\mathrm{PDD}(S;k)$ stronger than PDF, see Example 3.3 in the PDD paper.[32] The parameter-dependent smoothing of the raw PDF was introduced to guarantee the continuity under perturbations as in Fig. 3 when equal distances to neighbours become distinct. For automated comparisons, the smoothed PDF is often uniformly sampled, which creates the counter-intuitive pipeline: a discrete set $S \to$ smoothed PDF $\to$ discretely sampled PDF. To avoid this unnecessary smoothing, PDD matrices can be continuously compared in a parameter-free way.

Consider $\mathrm{PDD}(S;k)$ a discrete probability distribution of unordered rows (vectors in $\mathbb{R}^k$) with weights whose sum is 1. The simplest metric on such distributions is the Earth Mover's Distance (EMD), which came from transportation theory[34] and has been already applied to comparing chemical compositions,[35] see Definition 4.1 in the PDD paper.[32] Briefly, the EMD optimally transforms the rows of one PDD matrix into the rows of another PDD.

Fig. 5 illustrates the EMD computation when we perturb the unit square lattice S to S'

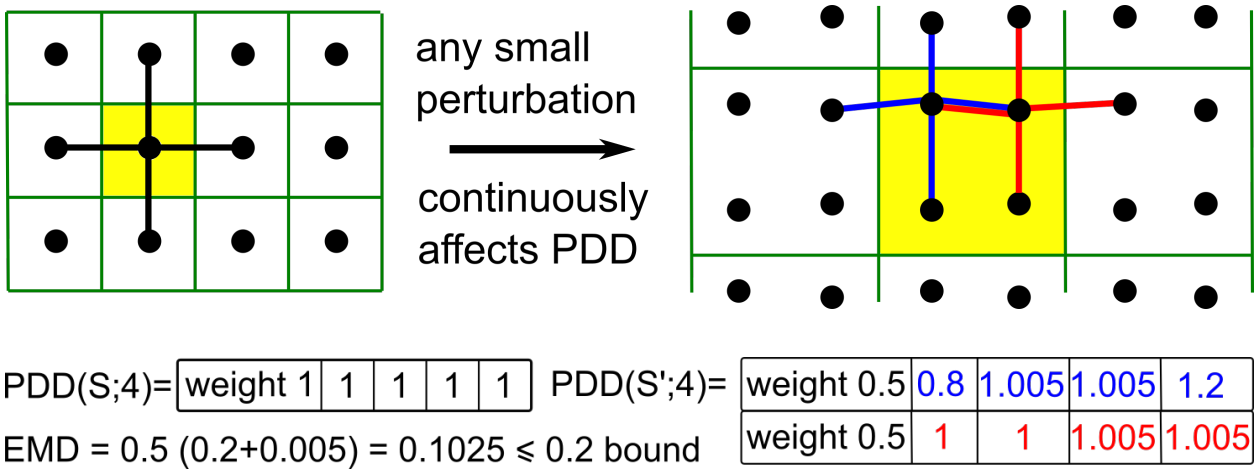

Figure 5: Under a small perturbation by $\varepsilon = 0.1$, the unit cell quadruples but all interpoint distances change by at most $2\varepsilon = 0.2$, which is averaged by EMD to a value under 2ε.

by moving both points in every other pair vertically by 0.1 away from each other. As a result, all interpoint distances remain in the range $[0.8, 1.2]$. The only reasonable way to transform the 1-row PDD$(S;4)$ into the 2-row PDD$(S';4)$ is to split the single row of PDD$(S;4)$ into two halves, which are compared with the two rows of PDD$(S';4)$ by (say) the L_∞ metric measuring the maximum absolute deviation of corresponding coordinates. Then EMD takes the weighted average of the two L_∞ metrics from the row comparisons. Theorem 4.3 in the PDD paper[32] proved the Lipschitz continuity of PDD$(S;k)$ in EMD with constant $\lambda = 2$.

More complicated atom-centred descriptors involving angles and higher order atom interactions use a cut-off radius and an order of points for angles that may not guarantee the invariance under permutations or completeness when near-duplicates can coincide on a large bounded domain as in Fig. 3. Fig. 5 explains how the discontinuity is resolved by using only interpoint distances. If a point p has two or more neighbours at the same distance then, after a small perturbation, a cut-off ball can include any of them, which can discontinuously affect a finite cluster of points but the k smallest distances always change continuously.

The approach through bounded clusters led to the *isoset* invariant,[36] which was proved to be complete for all periodic point sets including singular ones in any Euclidean space $\mathbb{R}^n$. The Lipschitz continuous metric on isosets is approximated with a proved error factor.[37]

The strongest theoretical result is the generic completeness and reconstructability for PDD$(S;k)$ combined with a lattice of S. Theorem 4.4 in the PDD paper[32] proved that any periodic point set $S \subset \mathbb{R}^n$ in a general position (outside a singular subspace of measure 0) can be reconstructed uniquely under isometry from a lattice of S, which can be given by complete invariants[10,26] in dimensions $n \leq 3$, and PDD$(S;k)$, where k should be large enough to include all distances up to $2R(S)$. Here the *covering radius* $R(S)$ is the minimum radius of balls that are centred at all points of S and cover the full ambient space $\mathbb{R}^n$. Theorem 5.1 in the PDD paper[32] proved that, for a fixed dimension n, the computational time of PDD$(S;k)$ depends only near-linearly on the motif size m and the number k of neighbours. The AMD and PDD invariants improved material property predictions[38–40] on some datasets.

The latest implementation of PDD$(S;100)$ can compare all (more than 830 thousand) periodic crystals from the CSD (with no disorder and full geometric data) through more than 345 billion comparisons in under one hour on a modest desktop. This ultra-fast speed allows us to visualise the CSD in the invariant coordinates on a laptop in real time.

The Crystal Isometry Principle inspired by R. Feynman

Definition 2 of a crystal *structure* as an equivalence class under rigid motion or (slightly weaker) isometry implies that all crystal structures can be studied within a common continuous space of periodic structures. Indeed, ignoring all atomic attributes maps any crystal structure to a periodic structure consisting of only zero-sized points at all atomic centres. Any slightly nonisometric crystals as in Table 1 are represented by close points (isometry classes) in the space of all isometry classes of periodic point sets, which is now called the *Crystal Isometry Space* CRIS$(\mathbb{R}^3)$. All periodic sets with at most m points in a unit cell form a $3m$-dimensional subspace CRIS$(\mathbb{R}^3;m)$. Here $3m$ is the number of fractional coordinates of m points, while 6 parameters of a unit cell are counter-balanced by 6-parameter isometries in $\mathbb{R}^3$.

Is it possible that we lose some information when ignoring atomic attributes? The first temptation is to keep at least all chemical elements. Traditional chemistry explored this path for centuries by separately studying organic vs inorganic compounds, and smaller subclasses (intermetallic, semiconductors, perovskites) to a level of a single composition.

However, fixing chemical elements breaks the continuous space CRIS($\mathbb{R}^3$) into many thousands of isolated pieces, one for each composition. These disjoint pieces can contain very different structures such as diamond and graphite, which does not help distinguish polymorphs that have the same composition but nonequivalent crystal structures.

The AMD[33] and PDD[32] papers reported several pairs of (near-)duplicates where all numerical parameters in the CIFs were equal with almost all digits but one atom was replaced with a different one. For example, the CSD entries HIFCAB[41] and JEPLIA[42] essentially differ only by replacing Cd with Mn at the same position without any other changes in the unit cell parameters or fractional coordinates. The integrity office at the Cambridge Crystallographic Data Centre (CCDC) checked that their structure factors were also identical and agreed that the found (near-)duplicates need a redetermination with better precision.

The all-vs-all comparisons of only periodic structures (without chemical attributes) for the CSD by PDD(S; 100) implied that if real periodic crystals are not isometric, then their periodic structures are not isometric. So ignoring atomic attributes loses no data, i.e. the **map {real crystal structures} $\to$ {periodic structures} is injective** modulo isometry for the CSD. This conclusion confirms our physical intuition that replacing one atom with a different one should perturb distances to neighbours at least slightly. All-vs-all comparisons are being finalised for other experimental databases such as COD,[43] ICSD,[44] and MP.[45]

The resulting *Crystal Isometry Principle* (CRISP) says that there should be no theoretical obstacle to reconstructing atomic attributes such as chemical elements from a periodic set of only atomic centres if their coordinates are determined with a high enough precision.

The CRISP is inspired by Richard Feynman's visual hint in Fig. 1-7 of his first lecture (atoms and motion) on physics,[46] which showed that 7 cubic crystals differ by their only

geometric parameter (the smallest interatomic distance) given up to 0.01Å. Comparing these 7 numbers was the Eureka moment for the last author in May 2021 and motivated us to complete all-vs-all comparisons of real periodic crystals in the CSD only by their geometry.

The CRISP does not claim that any periodic set of points can be realised as a real crystal because interatomic distances cannot be arbitrary. Hence Problem 6 can be strengthened by adding the realisability condition requiring an explicit parametrisation of all values $I(S)$ that can be realised first by a periodic structure S and then by a real crystal. This realisability has been achieved for lattices in dimensions two[10] and three.[26] In the finite case, any numbers $0 < a \le b \le c$ are realisable as distances between $m = 3$ points in $\mathbb{R}^n$ if and only if $c \le a + b$.

However, six interpoint distances between any four points in $\mathbb{R}^2$ satisfy a complicated polynomial equation saying that the tetrahedron on these four points has volume 0 and hence cannot be easily sampled. For example, six interpoint distances 1 satisfy all triangle inequalities and are realizable by an equidistant tetrahedron but not by four points in $\mathbb{R}^2$.

Because the PDD invariants quickly distinguished all real periodic crystals, any such crystal already has a uniquely defined location in the continuous space CRIS($\mathbb{R}^3$). So the CSD can be considered a very big and important discrete subset of CRIS($\mathbb{R}^3$). Any newly discovered periodic crystal will appear at a new place of CRIS($\mathbb{R}^3$) without disturbing all known ones.

Continuous geographic-style maps of the CSD

This section presents the first maps of the CSD as a subset in the Crystal Isometry Space CRIS($\mathbb{R}^3$) in pairs of invariant coordinates. Big datasets of simulated crystals were often visualised as a structure-energy landscape, which was a scatter plot with two coordinates (density and energy), so the structure was represented by a single invariant. We complement this physical density (g/cm^3) by the Point Packing Coefficient PPC(S) in Definition 7(b).

Fig. 6 shows that the physical density and PPC(S) differ. A periodic crystal S with a

unit cell U containing atoms whose total mass is $\mathrm{mass}(U)$ has the physical density

$$\rho(S) = \frac{\mathrm{mass}(U)}{\mathrm{vol}(U)} = \frac{1}{V_n}\frac{\mathrm{mass}(U)}{m}\frac{mV_n}{\mathrm{vol}(U)} = \frac{\mathrm{ATM}(S)}{V_n}\mathrm{PPC}^n(S), \text{ where } \mathrm{ATM}(S) = \frac{\mathrm{mass}(U)}{m}$$

is the average atomic mass (invariant of a chemical composition) of S. If we fix a chemical composition of S for $n = 3$, so $\mathrm{ATM}(S)$ and $V_3 = \frac{4}{3}\pi$ are fixed, then $\rho(S)$ is inversely proportional to $\mathrm{PPC}^{-3}(S)$, which is confirmed by all carbon allotropes (crystals consisting of pure carbon) whose points $(\rho(S), \mathrm{PPC}(S))$ lie on a cubic hyperbola in Fig. 6.

We zoomed in the central part of all images and excluded outliers beyond the visible ranges, e.g. all crystals with physical densities higher than 4.5 g/cm^3 are removed in Fig. 6.

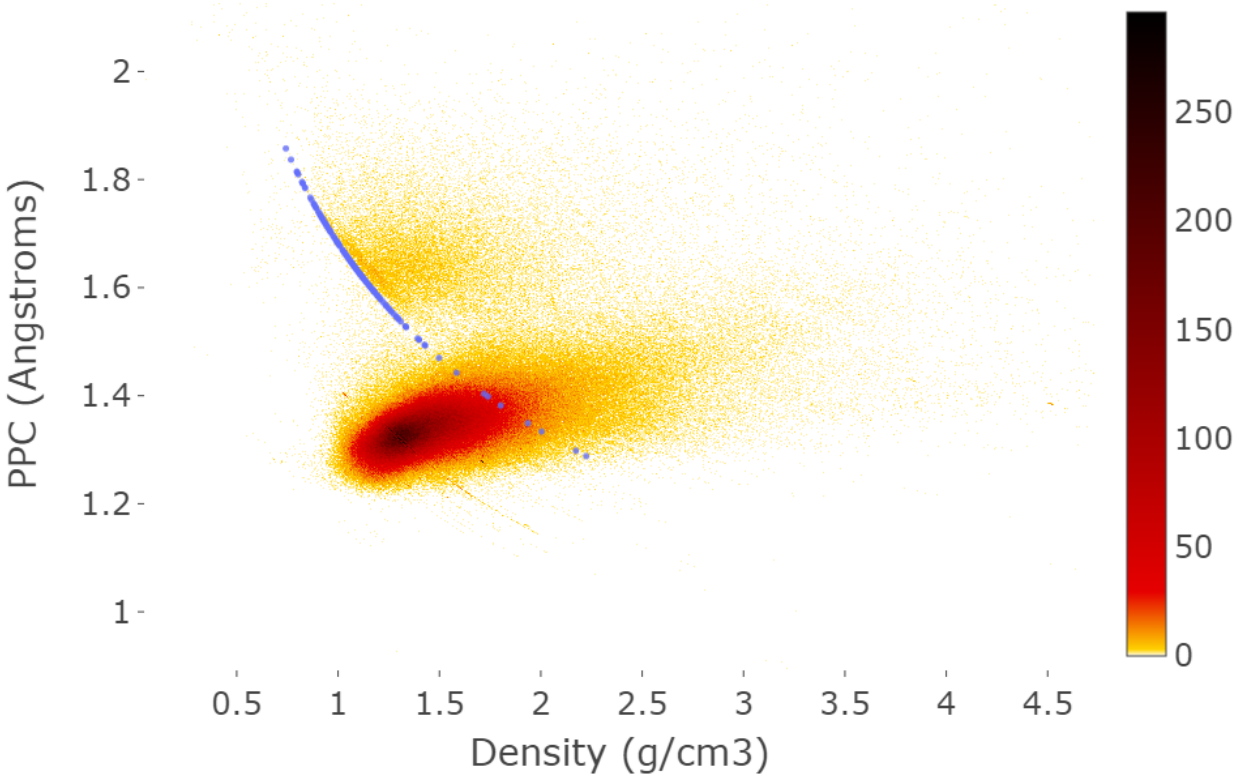

Figure 6: Scatter plot of 208 carbon allotropes over the whole CSD heatmap. The lower dense arc represents compositions H_2CO, $H_5C_2NO_2$, $H_6C_3NO_3$ with close values of the average atomic mass $\mathrm{ATM}(S)$. We zoomed on the densest part in the scatter plot in Fig. 7.

The scatter plot in Fig. 7 further zooms the highest density (black) region from Fig. 6.

All maps were produced by our *Crystal Geomaps* app, which already covers the well-

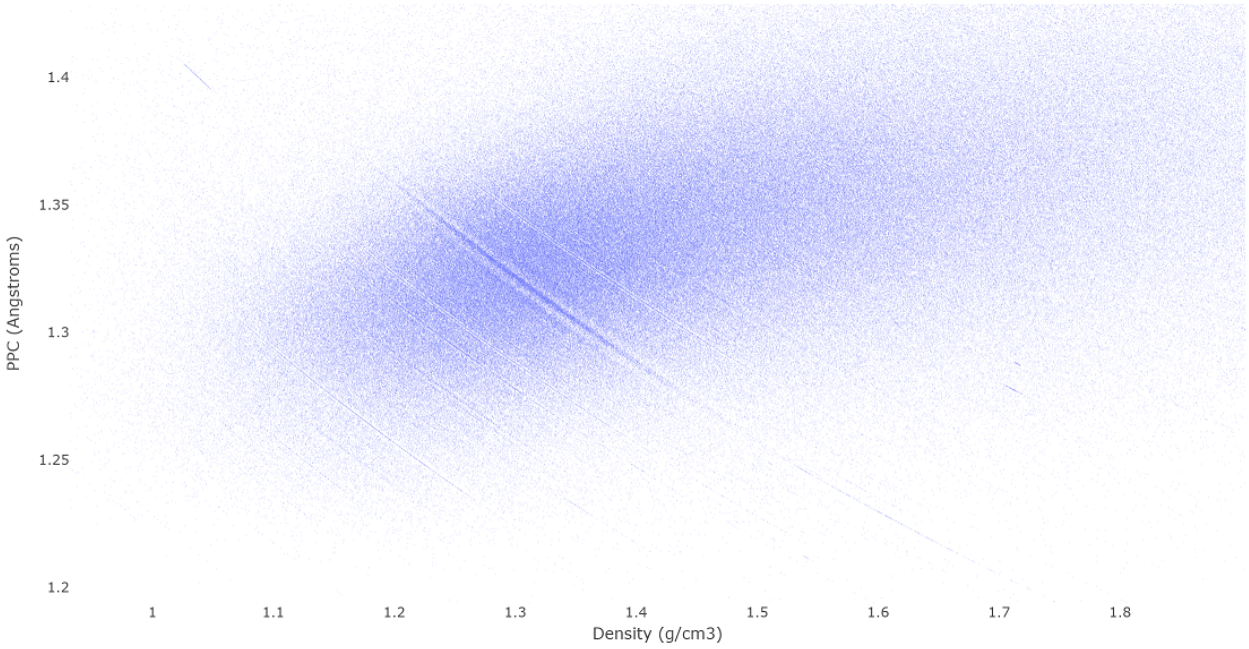

Figure 7: Scatter plot of the densest region in the CSD heatmap from Fig. 6 shows how the physical density $\rho(S) = \dfrac{3}{4\pi}\dfrac{\mathrm{ATM}(S)}{\mathrm{PPC}^3(S)}$ depends on the Point Packing Coefficient $\mathrm{PPC}(S)$.

known databases CSD, COD, ICSD, MP, GNoME, and we can include your data by request. The app is not yet public due to the ongoing IP commercialisation discussions with the CCDC. We are also open to collaboration with other partners in industry and academia.

The supporting information of this paper contains more maps produced by the Crystal Geomaps app, which allows one to interactively explore the major databases (CSD, COD, ICSD, MP, GNoME), individual CIFs or user-uploaded datasets of simulated crystals.

Because the average distance AMD_k to the k-th neighbour increases with respect to k, the maps with coordinates x, y from the list $\mathrm{AMD}_1 \le \mathrm{AMD}_2 \le \mathrm{AMD}_3 \le \dots$ are restricted to the half-plane $x \le y$. To avoid this artificial restriction, we subtracted the limit curve $\mathrm{PPC}(S)\sqrt[3]{k}$ from $\mathrm{AMD}_k(S)$ to get the less restrictive invariants $\mathrm{ADA}_k(S)$ in Definition 7(c).

Fig. 10 shows that projections of the CSD to the pairs of invariant coordinates (ρ, PPC) and $(\mathrm{ADA}_1, \mathrm{ADA}_2)$ are very different and hence represent different structural data.

Some intermetallic compounds can have close geometries and might appear close neigh-

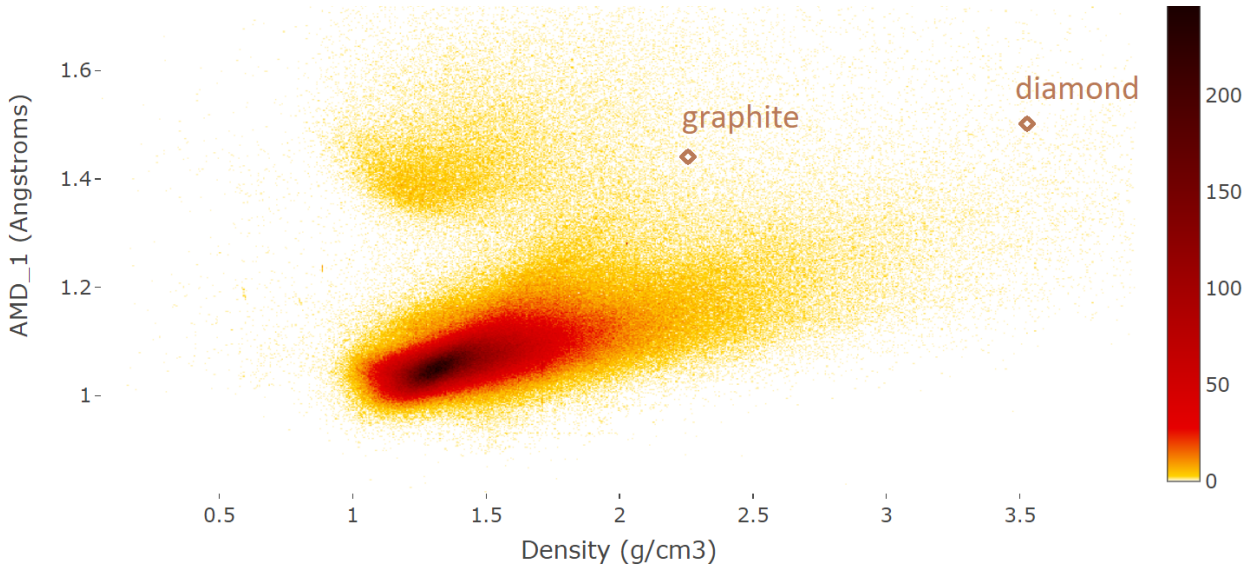

Figure 8: The clusters separated by a gap of $\mathrm{AMD}_1 \in [1.2, 1.3]$ are explained by the presence and absence of hydrogens that make AMD_1 smaller and larger, respectively, see Fig. 9.

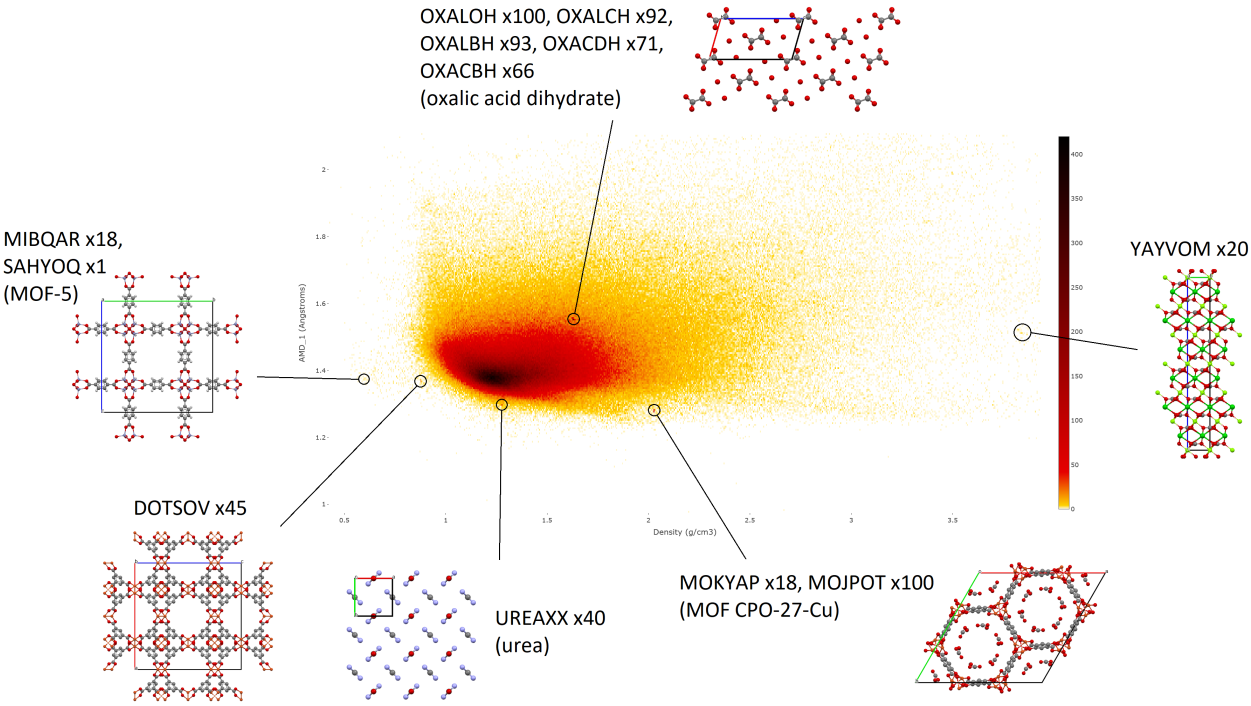

Figure 9: After removing hydrogens, the CSD becomes one large cluster, see Fig. 8 in the same coordinates (ρ, AMD_1). All dark spots represent groups of many (near-)duplicates.

bours in the space $\mathrm{CRIS}(\mathbb{R}^3)$ but we conjecture that all of them can be distinguished if we know the atomic coordinates precisely enough under the same ambient conditions as always.

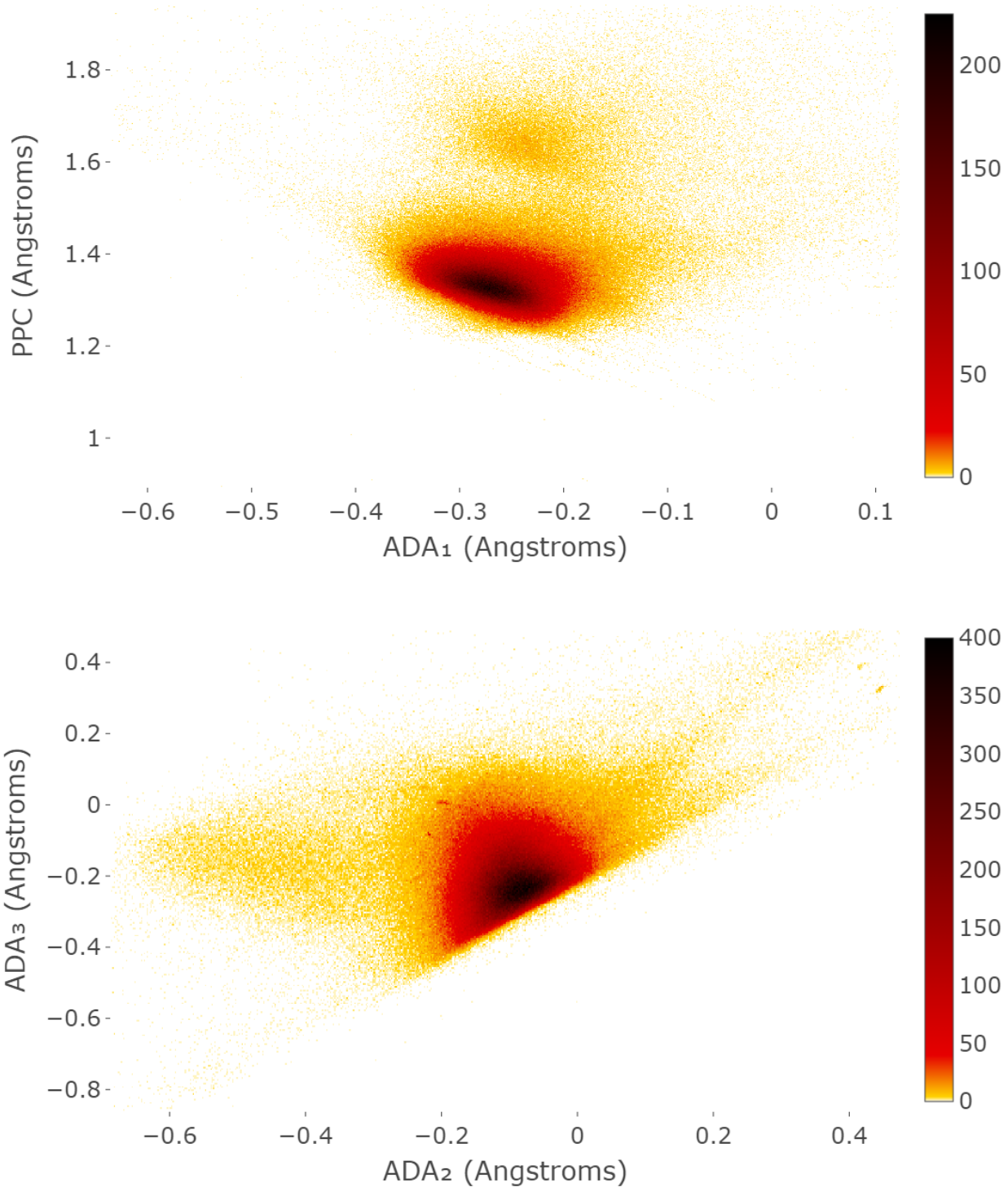

Figure 10: CSD heatmap in the new coordinates PPC, ADA_1, ADA_2 and ADA_3.

Conclusions: explore a continuous universe of crystals

When Olga Kennard established the Cambridge Structural Database (CSD) nearly 60 years ago, the crystallographic world was much smaller both in terms of people and crystals. Today, crystals are determined by many more methods and produced by artificial tools,[47] alongside 'paper mills'[48] claiming new materials without sufficient evidence.[49] The integrity of major databases[50] including the CSD[6] can now be validated by AMD[33] and PDD invariants.[32]

Fig. 3 illustrated how tiny perturbations can disguise any periodic crystal by making any extended cell primitive and pushing any symmetry down to the translation group P1. Because these changes can only slightly affect interatomic distances, then replacing chemical elements with similar ones may not raise any alarm. In November 2023, Google published[11] the GNoME database of 384,398 CIFs that were claimed to be 'stable' materials generated through DFT computations. All-vs-all comparisons will be discussed in another work but anyone right now can check that the four CIFs with IDs 4135ff7bc7, 6370e8cf86, c6afea2d8e, e1ea534c2c are identical texts. The GNoME contains 43 such triples, 1089 pairs, and many more thousands of numerical duplicates[12] that differ only by chemistry, not by geometry.

All images in this previous section are similar to usual geographic maps because they are deterministic projections of the infinite-dimensional Crystal Isometry Space CRIS($\mathbb{R}^3$) to pairs of coordinates that are invariant under isometry and rigid motion. All these coordinates have analytic definitions and physically meaningful units such as Angstroms. This interpretability is a key advantage of the new maps in comparison with the past approaches. For example, many algorithms of dimensionality reduction such as t-SNE[51] and UMAP[52] are stochastic so that they can produce different outputs by running at different times and on other computers. Even the deterministic algorithms such as regression[53] and PCA[54] have data-dependent coordinates and can be discontinuously affected by noise. In 2016, mathematicians proved[55] that any function $\mathbb{R}^m \to \mathbb{R}^n$ for all $m > n$ (reducing the dimension from m to n) is either discontinuous (makes close points distant) or collapses an unbounded region

to one point (loses an infinite amount of data) similar to the projection $(x, y) \mapsto x$.

This 'no-free-lunch' result implies that a similarity analysis for any high-dimensional data can be justified only by distance metrics in the original high-dimensional space, e.g. by Earth Mover's Distance (EMD) on invariants $\mathrm{PDD}(S;k)$ with k up to 100, while low-dimensional projections help visualise data but cannot confirm that any given objects are close.

All crystals whose simplest invariants fall into a single pixel in the maps of Figures 8 and 9, can be visualised with further invariant coordinates $\mathrm{ADA}_2(S), \mathrm{ADA}_3(S)$ and so on. This gradual expansion (or zooming in) guarantees that all crystals eventually become distinct because the vector $\mathrm{AMD}(S;100)$ distinguished all periodic crystals in the CSD.

Because all maps used only two of many available invariants, we cannot claim that crystals at close positions such as $(\mathrm{PPC}(S), \mathrm{ADA}_1(S))$ are always similar. However, these invariants quickly filter out dissimilar crystals so that if S, Q have distant values of $\mathrm{ADA}_k(S)$, then S, Q can not be made identical by a small perturbation of atoms. Due to the Lipschitz continuity, any value of Earth Mover's Distance $\delta = \mathrm{EMD}(\mathrm{PDD}(S;k), \mathrm{PDD}(Q;k)) > 0$ means that we need to shift all atoms by at least by $\delta/2$ on average to fully match the crystals S, Q.

In conclusion, we motivated studying periodic crystals under much stronger equivalences (rigid motion and isometry) that distinguish many crystals that were previously considered similar or isostructural as in Table 1. In the 20th century, the symmetries importantly helped to determine many 3-dimensional structures from their diffraction patterns, especially for highly symmetric crystals. In 2024, the ultrafast PDD invariants allowed us to move beyond the 230 space groups in $\mathbb{R}^3$ toward the continuous universe containing all known periodic crystals (already visible 'stars' on our maps) and also all not yet synthesised ones.

Visualising important properties such as energy as mountainous landscapes on these maps will help distinguish between shallow local minima and more stable materials in deeper 'wells' surrounded by energy barriers. Now all crystal structures at least in the CSD are fully discriminated by a series of Lipschitz continuous PDD invariants. This solution of the discriminative problem justifies a generative approach to explore new PDD matrices, which

are always guaranteed to be realisable by at most one real crystal whose properties are unique. Further work will continuously quantify the novelty of any newly discovered crystal by a distance metric to its closest structural neighbour across all experimental databases.

The Crystal Isometry Principle and underlying invariants were presented at the IUCr congresses 2021 and 2023, European Crystallographic Meeting 2022, British Crystallographic Association meetings 2022-2024, MACSMIN 2021-2023 (Mathematics and Computer Science for Materials Innovation), SIAM Mathematical Aspects of Materials Science conferences 2021 and 2024, and many MIF++ seminars at the Materials Innovation Factory in Liverpool, UK.

This work was supported by the second author's Royal Academy of Engineering Fellowship 'Data Science for Next Generation Engineering of Solid Crystalline Materials' at the Cambridge Crystallographic Data Centre (IF2122/186), the EPSRC New Horizons grant 'Inverse design of periodic crystals' (EP/X018474/1), and the Royal Society APEX fellowship 'New geometric methods for mapping the space of periodic crystals' (APX/R1/231152).

The supplementary information describes the functionality of the Crystal Geomaps app and continuous maps of some CSD subsets of crystals with specific chemical compositions.

References

(1) Fedorov, E. The symmetry of regular systems of figures. *Proceedings of the Imperial St. Petersburg Mineralogical Society* **1891**, *28(2)*, 1–146.

(2) Schönflies, A. Crystal Systems and Crystal Structure. **1891**,

(3) Ward, S. C.; Sadiq, G. Introduction to the Cambridge Structural Database – a wealth of knowledge gained from a million structures. *CrystEngComm* **2020**, *22*, 7143–7144.

(4) Pulido, A.; Chen, L.; Kaczorowski, T.; Holden, D.; Little, M. A.; Chong, S. Y.; Slater, B. J.; McMahon, D. P.; Bonillo, B.; Stackhouse, C. J.; others Functional materials discovery using energy–structure–function maps. *Nature* **2017**, *543*, 657–664.

(5) Sacchi, P.; Lusi, M.; Cruz-Cabeza, A. J.; Nauha, E.; Bernstein, J. Same or different – that is the question: identification of crystal forms from crystal structure data. *CrystEngComm* **2020**, *22*, 7170–7185.

(6) Francis, M. CCDC blog. https://prewww.ccdc.cam.ac.uk/discover/blog/new-and-notable-structures-added-to-the-csd-additional-improvements-and-data-integr

(7) Hahn, T. *International tables for crystallography*; 2005; Vol. A.

(8) Niggli, P. *Krystallographische und strukturtheoretische Grundbegriffe*; Akademische verlagsgesellschaft mbh, 1928; Vol. 1.

(9) Lawton, S.; Jacobson, R. *The reduced cell and its crystallographic applications*; 1965.

(10) Kurlin, V. Mathematics of 2-dimensional lattices. *Foundations of Computational Mathematics* **2022**, 1–59.

(11) Merchant, A.; Batzner, S.; Schoenholz, S. S.; Aykol, M.; Cheon, G.; Cubuk, E. D. Scaling deep learning for materials discovery. *Nature* **2023**, 80–85.

(12) Anosova, O.; Kurlin, V.; Senechal, M. The importance of definitions in crystallography. *IUCrJ* **2024**, *11*.

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

(14) Brock, C. P. Change to the definition of "crystal" in the IUCr Online Dictionary of Crystallography. https://www.iucr.org/news/newsletter/etc/articles?issue=151351&result_138339_result_page=17, 2021.

(15) Chapuis, G. Isostructural crystals in the IUCr Online Dictionary of Crystallography. https://dictionary.iucr.org/Isostructural_crystals, 2024.

(16) Chapuis, G. Crystal structure in the IUCr Online Dictionary of Crystallography. `https://dictionary.iucr.org/crystal_structure`, 2024.

(17) Chapuis, G. Crystal pattern in the IUCr Online Dictionary of Crystallography. `https://dictionary.iucr.org/crystal_pattern`, 2024.

(18) Pattern recognition. `https://en.wikipedia.org/wiki/Pattern_recognition`, 2024.

(19) Parthé, E.; Gelato, L.; Chabot, B.; Penzo, M.; Cenzual, K.; Gladyshevskii, R. *TYPIX standardized data and crystal chemical characterization of inorganic structure types*; Springer Science & Business Media, 2013.

(20) Rass, S.; König, S.; Ahmad, S.; Goman, M. Metricizing Euclidean Space towards Desired Distance Relations in Point Clouds. *arXiv:2211.03674* **2022**,

(21) Mosca, M. M.; Kurlin, V. Voronoi-based similarity distances between arbitrary crystal lattices. *Crystal Research and Technology* **2020**, *55*, 1900197.

(22) Boutin, M.; Kemper, G. On reconstructing n-point configurations from the distribution of distances or areas. *Adv. Appl. Math.* **2004**, *32*, 709–735.

(23) Widdowson, D.; Kurlin, V. Recognizing Rigid Patterns of Unlabeled Point Clouds by Complete and Continuous Isometry Invariants With No False Negatives and No False Positives. *Proceedings of Computer Vision and Pattern Recognition* **2023**, 1275–1284.

(24) Watson, J. D.; Crick, F. H. The structure of DNA. Cold Spring Harbor symposia on quantitative biology. 1953; pp 123–131.

(25) Edelsbrunner, H.; Heiss, T.; Kurlin, V.; Smith, P.; Wintraecken, M. The Density Fingerprint of a Periodic Point Set. Proceedings of SoCG. 2021; pp 32:1–32:16.

(26) Kurlin, V. A complete isometry classification of 3D lattices. *arxiv:2201.10543* **2022**,

(27) Voronoi, G. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. *J. Reine Angew. Math* **1908**, 97–178.

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

(29) Smith, P.; Kurlin, V. A practical algorithm for degree-k Voronoi domains of three-dimensional periodic point sets. LNCS (Proceedings of ISVC). 2022; pp 377–391.

(30) Anosova, O.; Kurlin, V. Density functions of periodic sequences. LNCS (Proceedings of DGMM). 2022; pp 395–408.

(31) Anosova, O.; Kurlin, V. Density functions of periodic sequences of continuous events. *Journal of Mathematical Imaging and Vision* **2023**, *65*, 689–701.

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

(33) Widdowson, D.; Mosca, M. M.; Pulido, A.; Cooper, A. I.; Kurlin, V. Average Minimum Distances of periodic point sets - foundational invariants for mapping all periodic crystals. *MATCH Comm. in Math. and in Computer Chemistry* **2022**, *87*, 529–559.

(34) Kantorovich, L. V. Mathematical methods of organizing and planning production. *Management science* **1960**, *6*, 366–422.

(35) 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* **2020**, *32*, 10610–10620.

(36) Anosova, O.; Kurlin, V. An isometry classification of periodic point sets. LNCS (Proceedings of DGMM). 2021; pp 229–241.

(37) Anosova, O.; Kurlin, V. Algorithms for continuous metrics on periodic crystals. *arxiv:2205.15298* **2022**,

(38) Ropers, J.; Mosca, M. M.; Anosova, O. D.; Kurlin, V. A.; Cooper, A. I. Fast predictions of lattice energies by continuous isometry invariants of crystal structures. Intern. Conference on Data Analytics and Management in Data Intensive Domains. 2022; pp 178–192.

(39) Balasingham, J.; Zamaraev, V.; Kurlin, V. Material Property Prediction using Graphs based on Generically Complete Isometry Invariants. *Integrating Materials and Manufacturing Innovation* **2024**, 1–14.

(40) Balasingham, J.; Zamaraev, V.; Kurlin, V. Accelerating Material Property Prediction using Generically Complete Isometry Invariants. *Scientific Reports* **2024**, *14*, 10132.

(41) Man-Sheng, C.; Chun-Hua, Z.; Dai-Zhi, K.; Yong-Lan, F.; Yi-Fang, D. Poly [(μ2-4,4'-bipyridine-κ2N:N')(μ2-furan-2-carboxylato-κ2O: O') cadmium (II)]. *Acta Crystallographica Section E: Structure Reports Online* **2007**, *63*, m1290–m1291.

(42) Man-Sheng, C.; Yi-Fang, D.; Dai-Zhi, K.; Chun-Hua, Z.; Yong-Lan, F.; Yun-Lin, P. Synthesis, crystal structure and quantum chemistry of a 2D coordination polymer [Mn (FA)(2)(4, 4'-bipy)](n). *Chinese Journal of Inorganic Chemistry* **2006**, *22*, 1715–1718.

(43) Gražulis, S.; Daškevič, A.; Merkys, A.; Chateigner, D.; Lutterotti, L.; Quiros, M.; Serebryanaya, N. R.; Moeck, P.; Downs, R. T.; Le Bail, A. Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic acids research* **2012**, *40*, D420–D427.

(44) Zagorac, D.; Müller, H.; Ruehl, S.; Zagorac, J.; Rehme, S. Recent developments in the Inorganic Crystal Structure Database: theoretical crystal structure data and related features. *Journal of applied crystallography* **2019**, *52*, 918–925.

(45) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; others Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL materials* **2013**, *1*.

(46) Feynman, R. *The Feynman lectures on physics*; 1971; Vol. 1.

(47) Cheetham, A. K.; Seshadri, R. Artificial Intelligence Driving Materials Discovery? Perspective on the Article: Scaling Deep Learning for Materials Discovery. *Chemistry of Materials* **2024**, *36*, 3490–3495.

(48) 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.

(49) Leeman, J.; Liu, Y.; Stiles, J.; Lee, S. B.; Bhatt, P.; Schoop, L. M.; Palgrave, R. G. Challenges in High-Throughput Inorganic Materials Prediction and Autonomous Synthesis. *PRX Energy* **2024**, *3*, 011002.

(50) Chawla, D. S. Crystallography databases hunt for fraudulent structures. `https://cen.acs.org/research-integrity/Crystallography-databases-hunt-fraudulent-structures/102/i8`, 2024.

(51) Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **2008**, *9*.

(52) McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426* **2018**,

(53) Anscombe, F. Graphs in statistical analysis. *The Amer. Statistician* **1973**, *27*, 17–21.

(54) Elhaik, E. Principal component analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. *Scientific Reports* **2022**, *12*, 14683.

(55) Landweber, P. S.; Lazar, E. A.; Patel, N. On fiber diameters of continuous maps. *The American Mathematical Monthly* **2016**, *123*, 392–397.

Supporting information with continuous maps of some CSD subsets

Fig. 11 shows two parts of the interactive menu in the Crystal Geomaps app. Hovering the mouse over any pixel in a heatmap shows the coordinates x, y and the number of crystals at this position (x, y). Hovering the mouse over any dot (x, y) in a scatter plot shows the database reference, composition, and space group number of the crystal at (x, y).

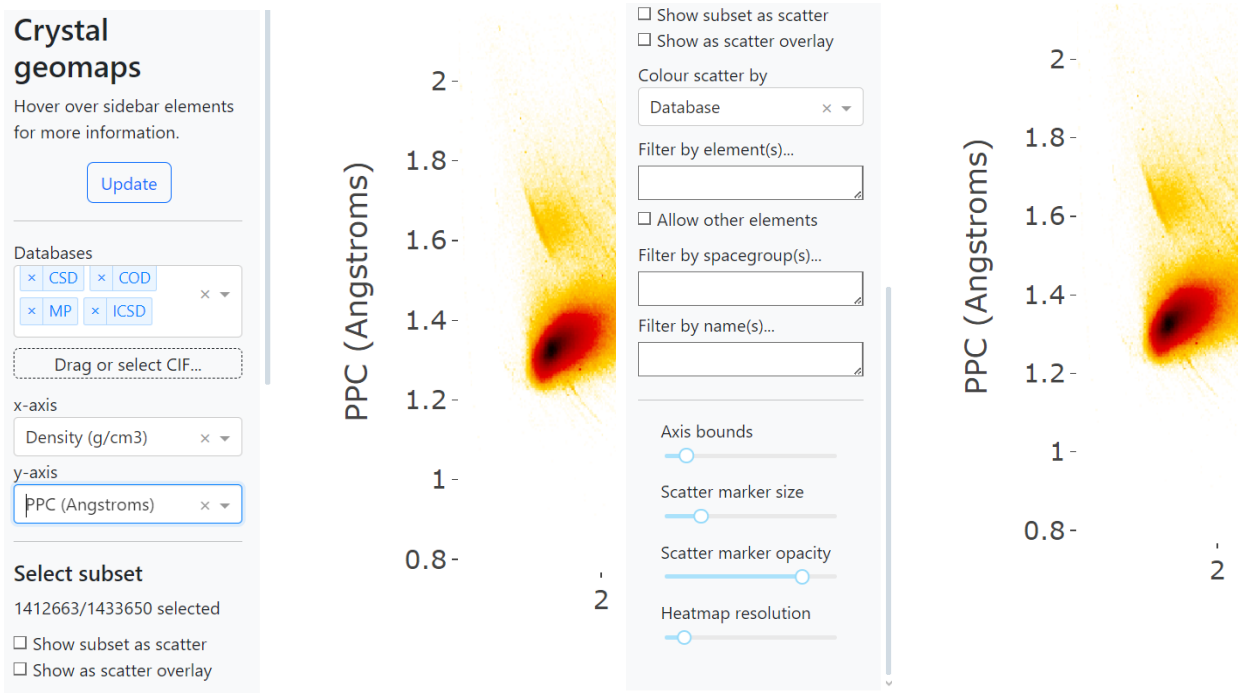

Figure 11: The images on the left and right show two parts of a user menu with choices for coordinates (density, PPC, ADA_k, NDA_k) and display options. A scatter plot represents any crystal from a selected subset by an individual dot and can also be plotted over a density heatmap of chosen databases. A subset can be selected by chemical elements, space groups, e.g. 195-230 for the cubic crystal system, or database reference names.

Fig. 11 implies that the total number of ideal periodic crystals (with no disorder) in the four major databases is more than 1,433,000, though many of them are (near-)duplicates often deposited from the same publication. These overlaps between different databases will be analysed in another work. The displayed number of crystals is slightly smaller because some outliers are outside the visible ranges that can be adjusted by the slider "Axis bounds". Other sliders control sizes of dots and pixels. In any scatter plot, all crystal dots can be coloured by a database name, 7 crystal systems, 230 space groups and invariant values.

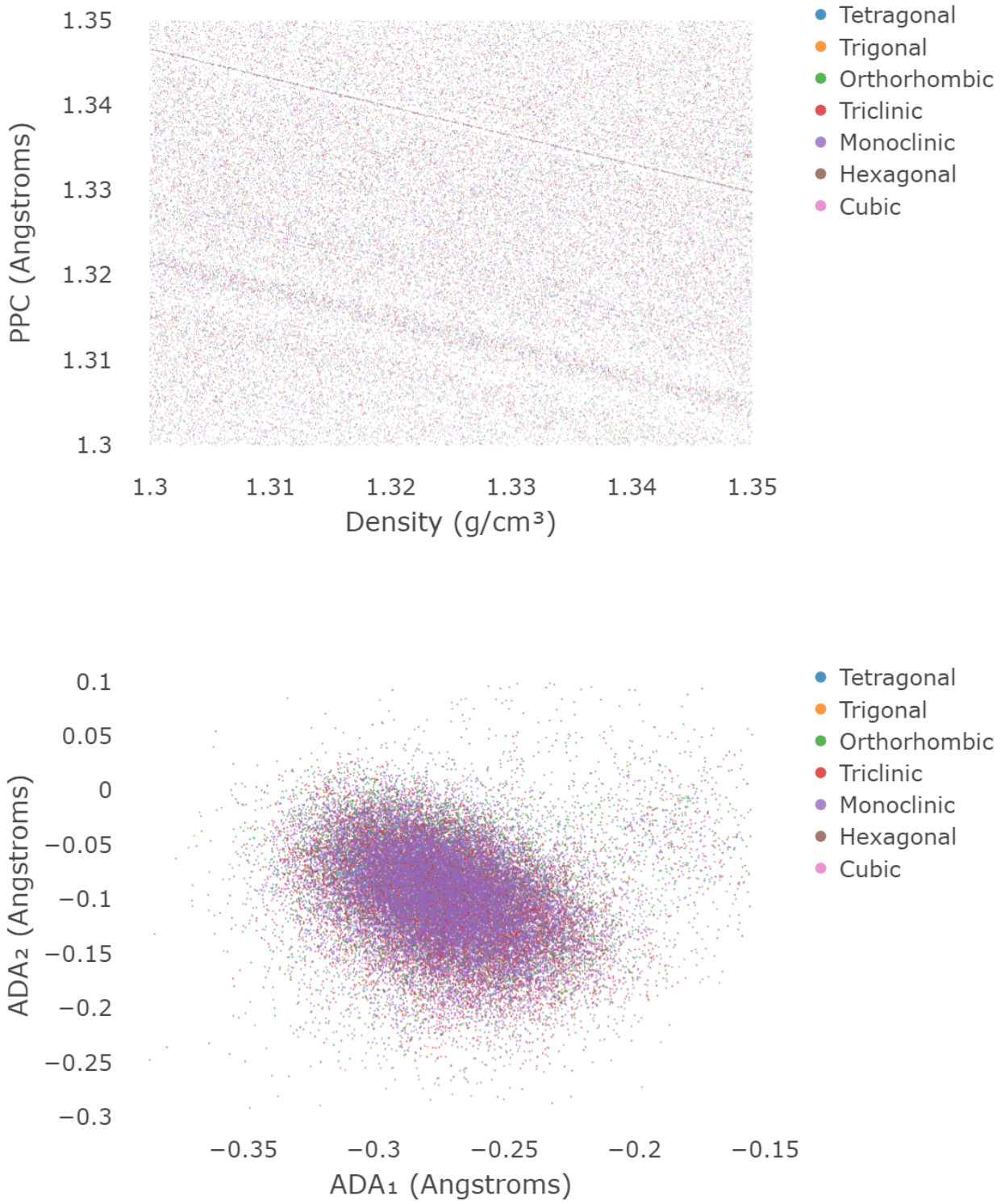

Figure 12: **Top**: the densest rectangular region of 44102 crystals from the CSD in the coordinates (ρ, PPC). **Bottom**: the same subset is differently projected to $(\mathrm{ADA}_1, \mathrm{ADA}_2)$.

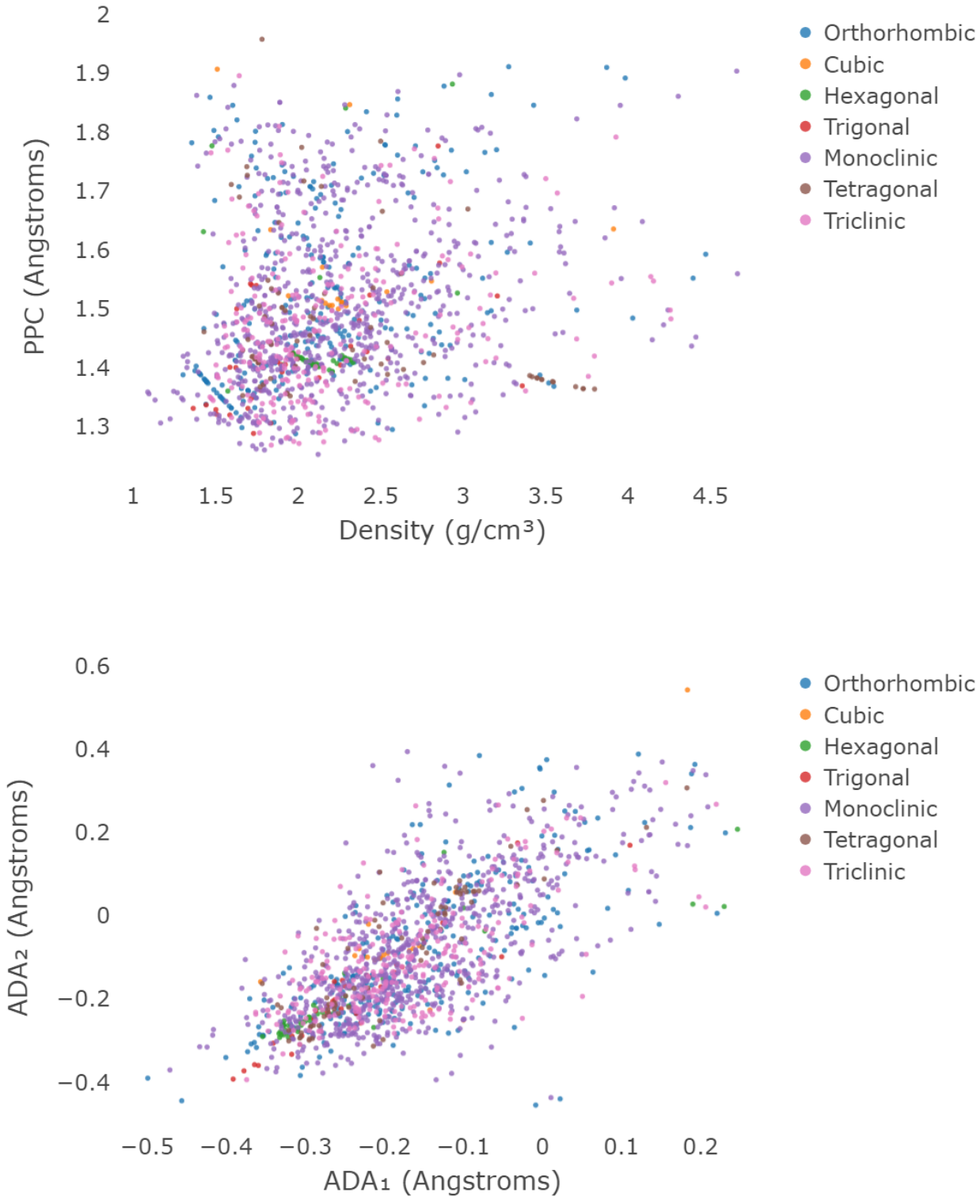

Figure 13: 1453 crystals containing carbon and sulfur, all coloured by their crystal systems.

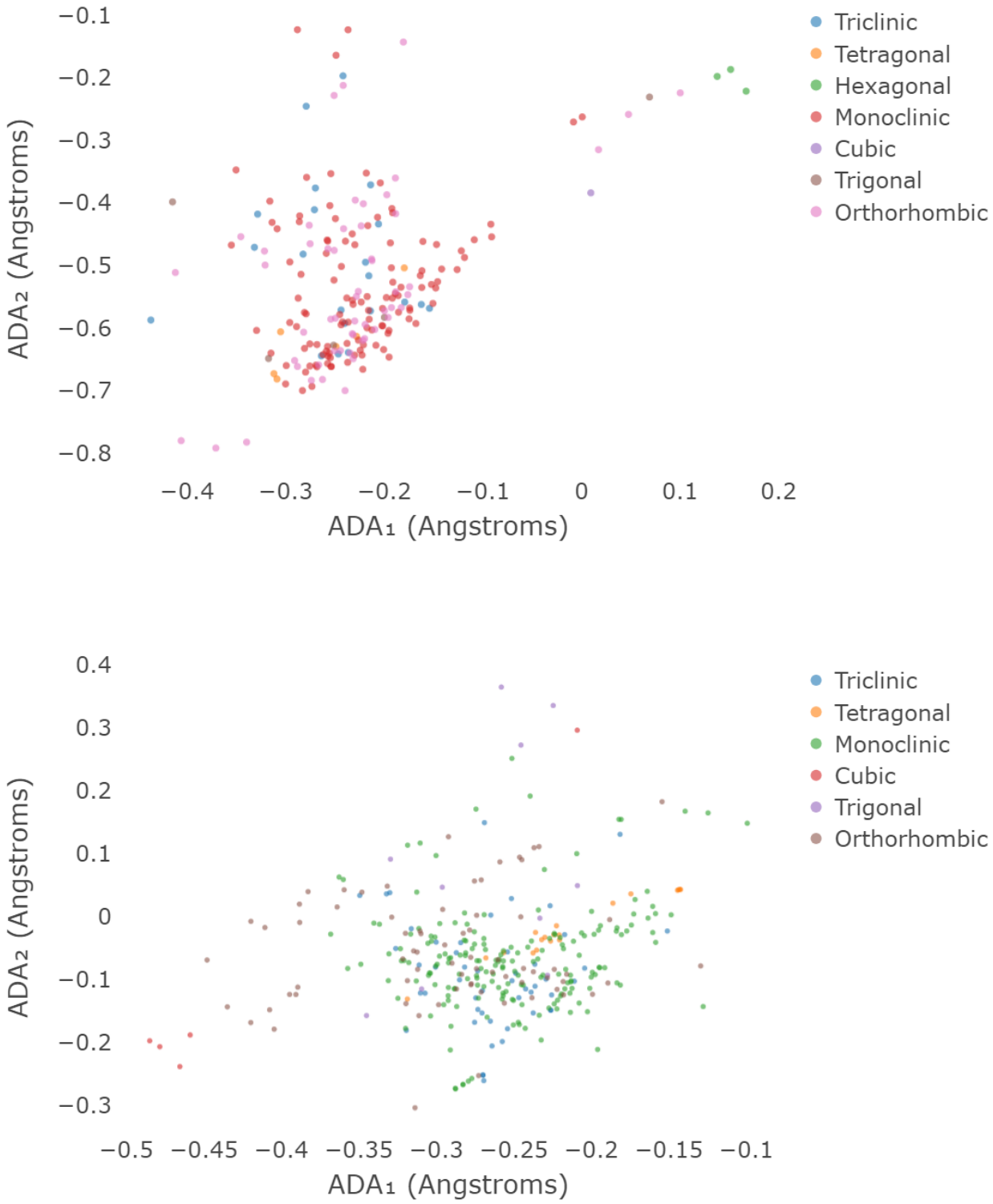

Figure 14: **Top**: the scatter plot of 208 carbon allotropes in the CSD. **Bottom**: the scatter plot of 345 carbohydrates in the CSD, all coloured by their crystal systems.

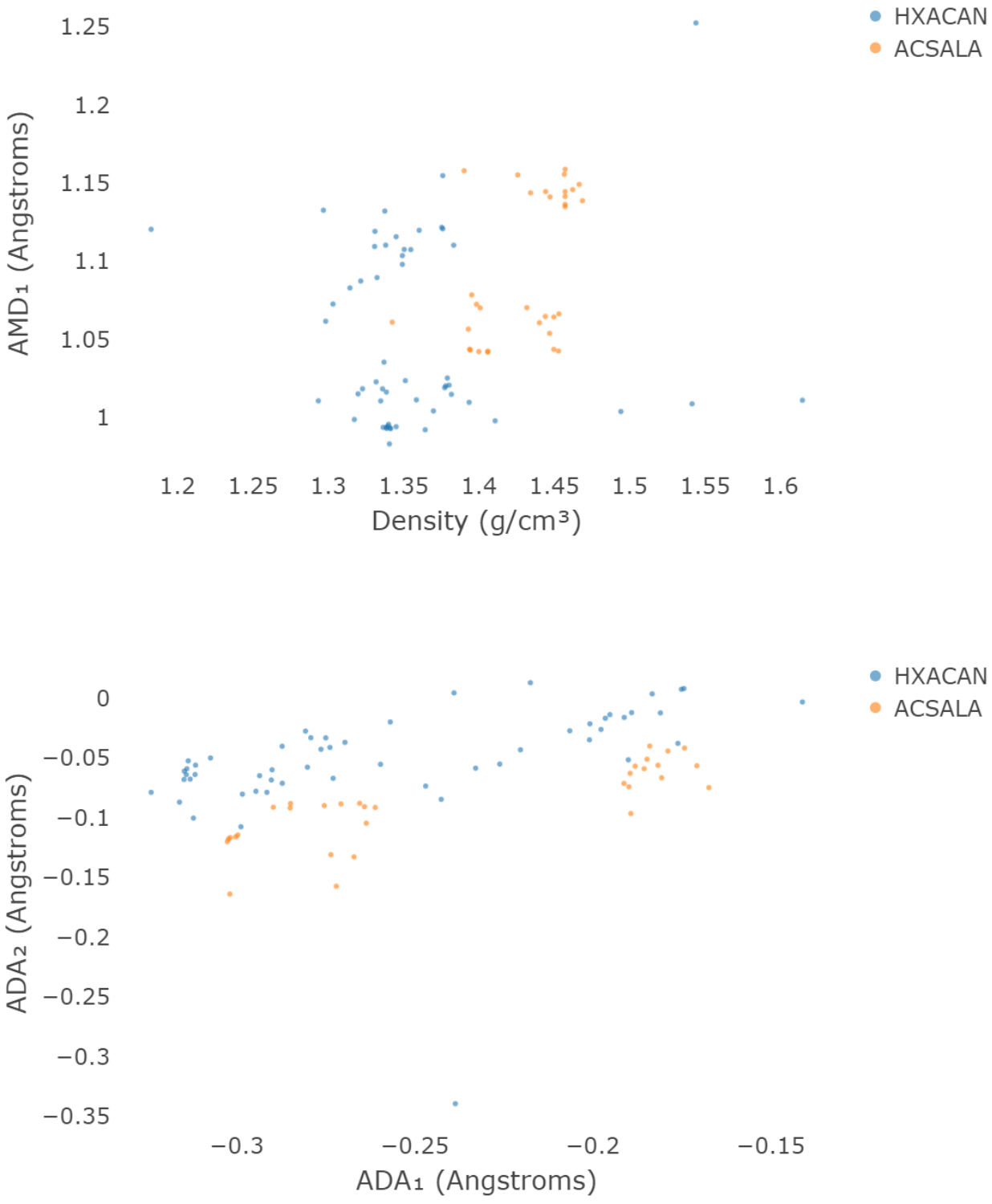

Figure 15: 86 crystals of ACSALA (aspirin) and HXACAN (paracetamol) in the CSD.

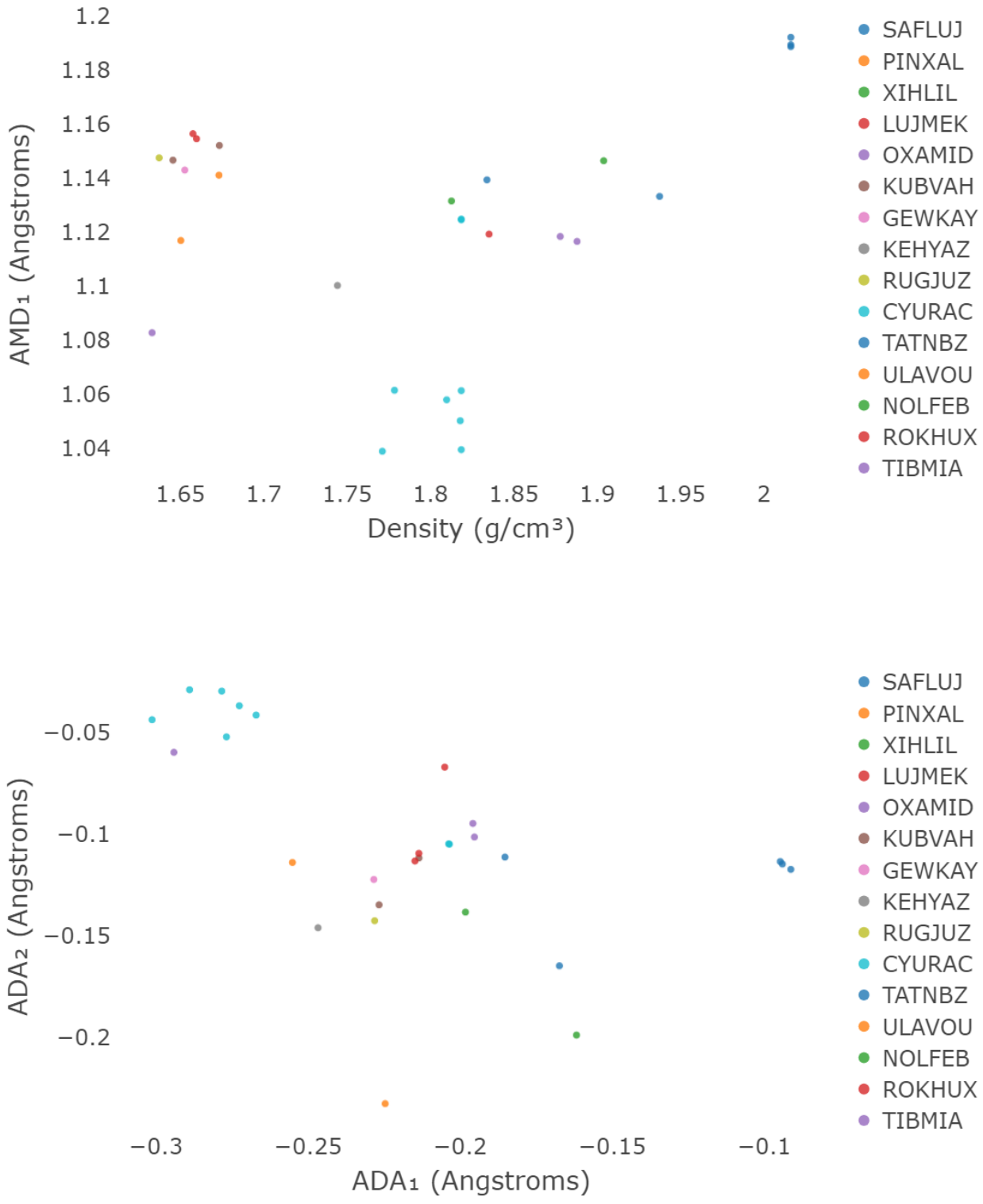

Figure 16: 28 crystals containing only C, O, H, N, all coloured by CSD refcode families.