Ultra-fast detection of (near-)duplicate structures across major crystal databases

Daniel Widdowson, Vitaliy Kurlin. Computer Science, *Materials Innovation Factory*, Liverpool

Problem: the incomplete definition of `*isostructural crystals*' allowed anyone to claim a new material by **disguising a known crystal** via almost any perturbation discontinuously changing a reduced cell [1].

Crystal structures are determined in a *rigid form* and hence are indistinguishable under *rigid motion.*

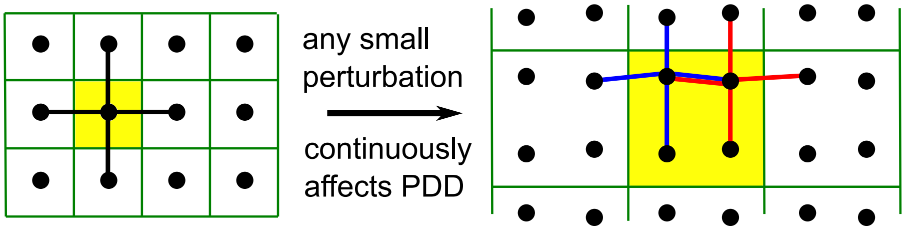
Rigid motion = translations + rotations, ***isometry*** = any rigid motion + reflection. **New definition** [1] :

a ***periodic structure*** is a *class* of periodic sets of atomic centres under *rigid motion* or (weaker) *isometry*

Solution : the invariant descriptor Pointwise Distance Distribution

PDD(S;k) = matrix of weighted rows of distances from an atom of S to k nearest neighbours in the full S, invertible to any generic crystal S.

Earth Mover's Distance (**EMD**) on PDDs satisfies all metric axioms.


any small perturbation

continuously affects PDD

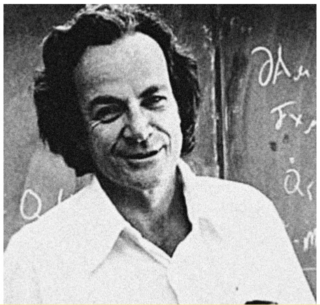
PDD(S;4)=

weight 1	1	1	1	1

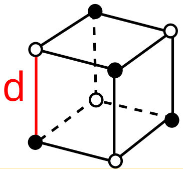
PDD(Q;4)=

weight 0.5	0.8	1.005	1.005	1.2
weight 0.5	1	1	1.005	1.005

EMD = 0.5 (0.2+0.005) = 0.1025 ⩽ 0.2 bound



The **Feynman** lectures, chap. 1 *Atoms in motion* Fig.1-7 is a **hint**!



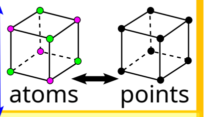
●	○	d (Å)
Na	Cl	2.82
K	Cl	3.14
Ag	Cl	2.77
Mg	O	2.10
Pb	S	2.98
Pb	Se	3.07
Pb	Te	3.17

Crystal Isometry Principle: chemistry ⟷ geometry

all **real periodic crystals** of atoms with *chemical elements*

different crystals ⟷ different structures

geometry *suffices* to reconstruct chemistry



all **periodic structures** of atomic centers *without elements*

isometry classes of point sets outside the crystal space

Scale of duplication : Google's GNoME claimed "2.2M new crystals – equivalent to nearly 800 years' worth of knowledge" [2], made 384K+ CIFs public, of which 43 were synthesised by Berkeley's A-lab [3]

Review [4] of [3] : "none of the 43 materials [of 58 attempted] produced by A-lab were new: the large majority were misclassified, and a smaller number were correctly identified but already known".

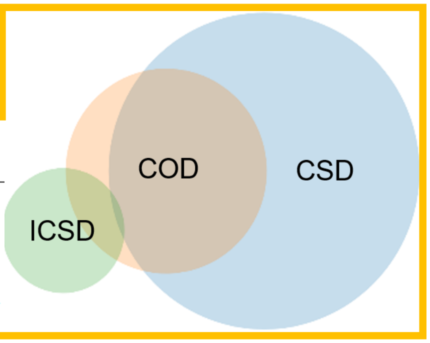
Review [5] of [3] : all 43 materials have near-duplicates in ICSD or Materials Project used for training.

Review [6] of [2] : "scant evidence for compounds that fulfill the trifecta of novelty, credibility, utility".

Review [1] of [2] : 4 identical CIFs, 43 triples, 1089 pairs, many more thousands of digital duplicates.

Loophole closed : [7-9] used PDD invariants to detect large subsets of near-duplicates that differ up to 0.01A in average atomic deviations for about 2M crystals from 5 databases within 9 min on a modest desktop.

databases	CSD		COD		ICSD		MP		GNoME	
duplicates	count	%	count	%	count	%	count	%	count	%
CSD	**7687**	**0.9**	272649	32.8	4649	0.6	21	0.0	1	0.0
COD	276328	80.3	**19231**	**5.6**	36553	10.6	5239	1.52	2705	0.8
ICSD	4736	4.5	48899	46.5	**35189**	**33.5**	16386	15.6	9123	8.7
MP	64	0.0	11989	7.82	14312	9.3	**19177**	**12.5**	10681	7.0
GNoME	2	0.0	1801	0.5	2459	0.6	3401	0.9	**82859**	**21.5**



[1] **O.Anosova, V.Kurlin, M.Senechal.** The importance of definitions in crystallography. *IUCrJ*, v.11 (4), p.453-463, 2024.

[2] **A.Merchant et al**. Scaling deep learning for materials discovery. *Nature* 624 (7990), 80-85, November 2023.

[3] **N.Szymanski et al**. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* 624 (7990), 86–91.

[4] **J.Leeman et al**. Challenges in High-Throughput Inorganic Materials Prediction. *PRX Energy* 3, 011002, March 2024.

[5] **D.Widdowson, V.Kurlin**. Navigation maps of the materials space for automated self-driving labs, *arxiv:2410.13796*.

[6] **A.Cheetham, R.Seshadri**. Artificial Intelligence Driving Materials Discovery? *Chemistry of Materials* 36, 3490–3495, 2024.

[7] **D.Widdowson et al**. Average Minimum Distances of periodic point sets. *MATCH*, v.87 (3), 529-559, 2022.

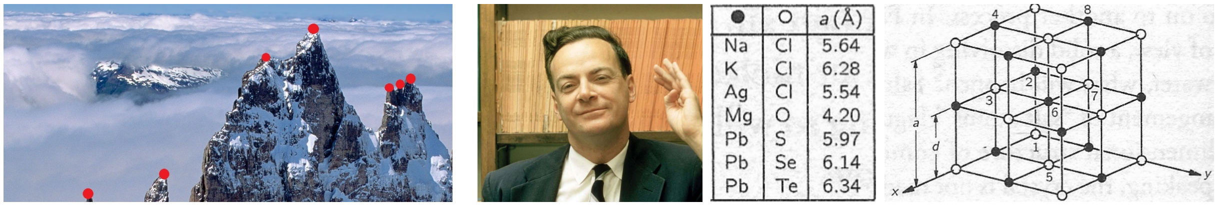
[8] **D.Widdowson, V.Kurlin**. Resolving the data ambiguity for periodic crystals. *NeurIPS*, v.35, 24625-24638, 2022.

[9] **D.Widdowson, V.Kurlin**. Continuous invariant-based maps of the CSD. *Crystal Growth & Design*, v.24, 5627–5636, 2024.

Geometric Data Science (GDS) develops continuous maps of data objects

The vision is to map (continuously parametrize) the space of any data objects considered up to practical equivalences. While Geometric Deep Learning experimentally outputs equivariant descriptors of clouds or graphs, GDS developed analytic, complete and continuous invariants for any finite and generic periodic sets of unordered points in $\mathbb{R}^n$, see the papers in NeurIPS 2022 and CVPR 2023 at http://kurlin.org/research-papers.php#Geometric-Data-Science.

The key obstacle for periodic crystals was the ambiguity of conventional data based on minimal or reduced cells that are discontinuous under atomic displacements. Without continuously quantifying the crystal similarity, the brute-force Crystal Structure Prediction produces millions of nearly identical approximations to numerous local energy minima, see red peaks in Fig. 1.



●	○	a (Å)
Na	Cl	5.64
K	Cl	6.28
Ag	Cl	5.54
Mg	O	4.20
Pb	S	5.97
Pb	Se	6.14
Pb	Te	6.34

Figure 1: **Left**: energy landscapes show crystals as isolated peaks of height$=-$energy. To see beyond the 'fog', we need a map parametrized by invariant coordinates with a continuous metric. **Right**: R. Feynman's first lecture showed that 7 cubic crystals differ by side lengths, while our invariants distinguished all 850K+ periodic crystals in the CSD. These crystals have unique positions in a common *Crystal Isometry Space* whose one 2D projection is in Fig. 2.

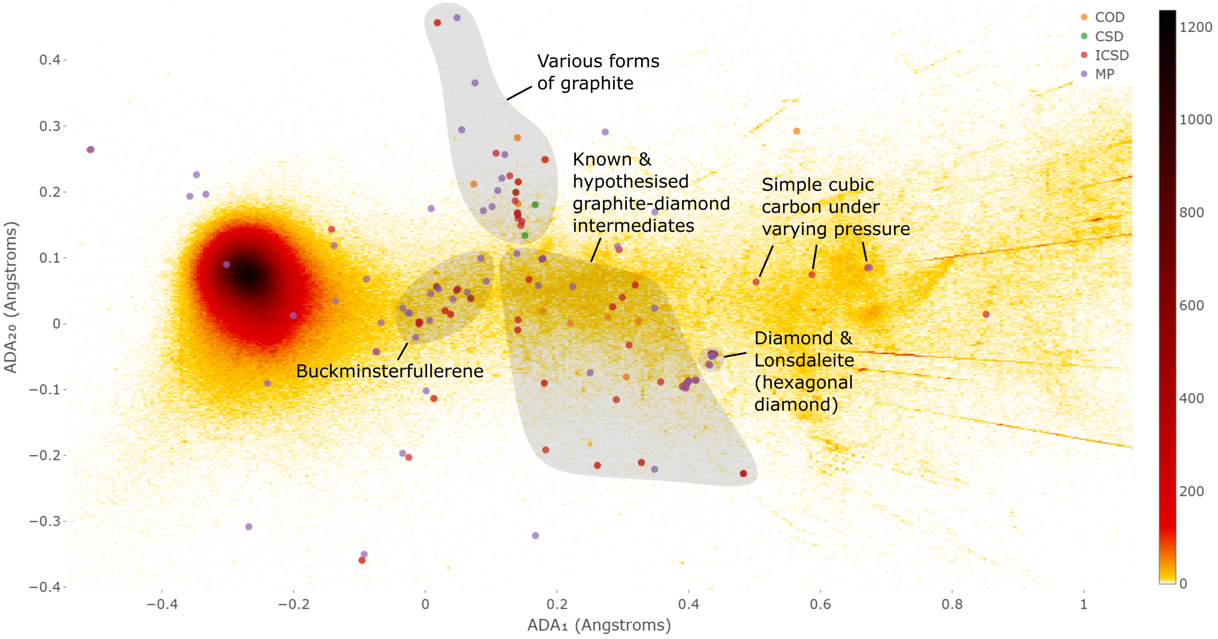


Figure 2: Carbon allotropes on a continuous map of periodic crystals in the Cambridge Structural Database (CSD), Crystallography Open Database (COD), Inorganic Crystal Structure Database (ICSD), and Materials Project (MP). The colour indicates the number of crystals whose AMD_k (average distance to the k-th atomic neighbour) are discretized to each pixel.

The crystal space can be visualized in other explicit coordinates from AMD vectors and PDD matrices. The density of S has been extracted from the asymptotic of $\mathrm{AMD}_k(S)$ as $k \to +\infty$.