The Crystal Geomap visualises materials databases in real time

Daniel Widdowson[1]
Vitaliy Kurlin[1]

[1]Materials Innovation Factory, University of Liverpool

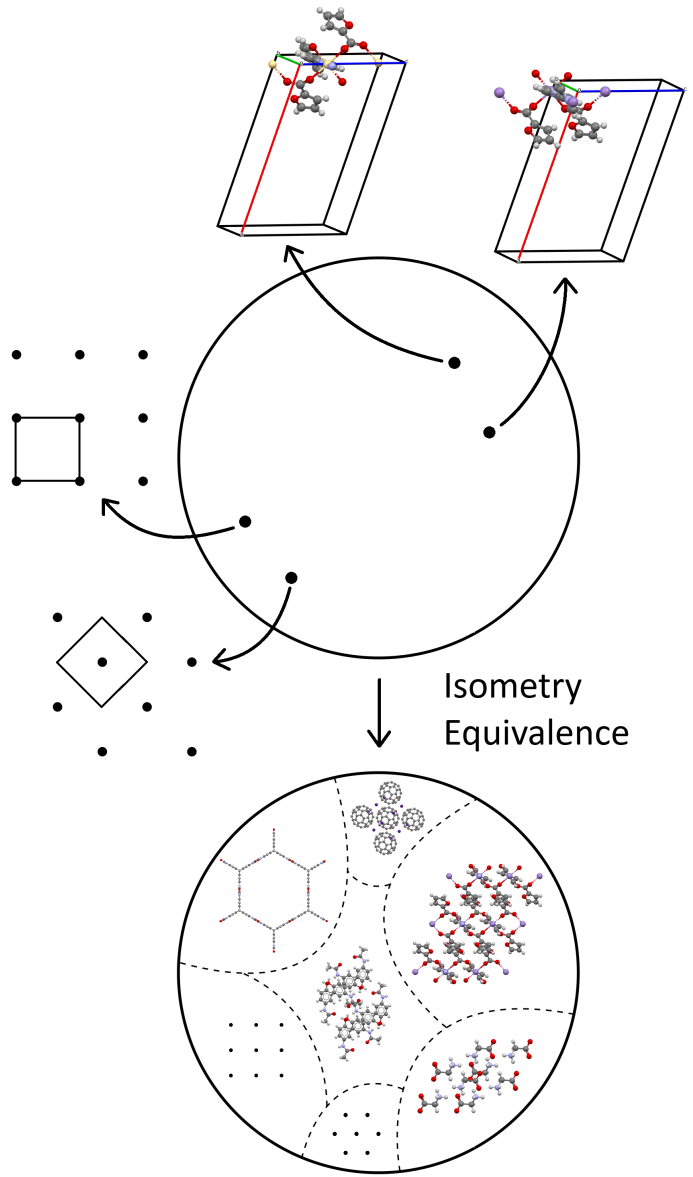
2026

Mathematics for crystals

- IUCr: "A solid is a *crystal* if its atoms, ions and/or molecules form, on average, a long-range ordered arrangement".
- Our approach:
  - Define a *periodic point set* representing (the geometry of) a crystal,
  - Consider them equivalent if they're related by *isometry* making a well-defined *crystal space*

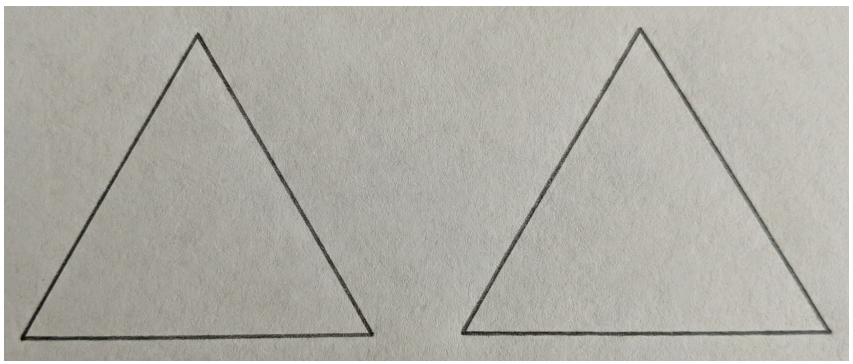
$$\{\Lambda + M : \Lambda \text{ lattice}, M \text{ finite}\} \ / \ \text{isometry}$$

  - Give coordinates for this space ('*isometry invariants*').



Triangles: same or different?

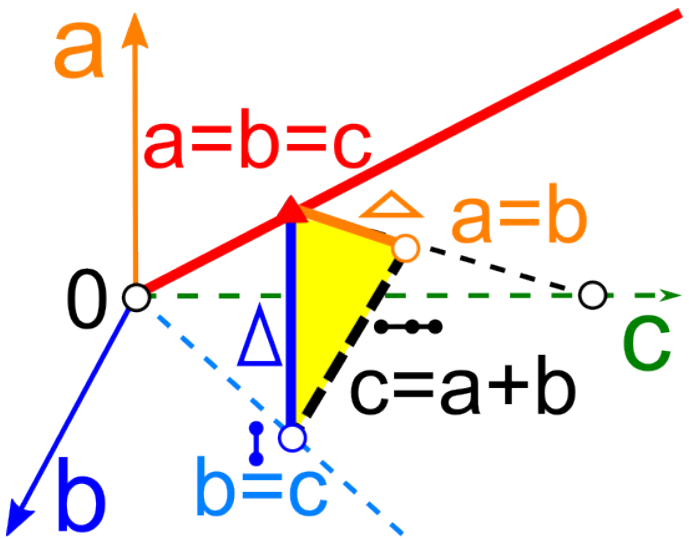
Triangles (3-point sets) are often classified by their symmetries into 3 groups: equilateral, isosceles and scalene. But real-world data is imperfect and can contain noise. What if a measured triangle is *nearly* equilateral? Can we quantify how close it is?



We could say a triangle is *pseudosymmetric* if it's only a small deformation ε away from perfect symmetry. But this just shifts the problem: now near-identical triangles each side of the boundary ε are classified differently.

Triangles: same or different?

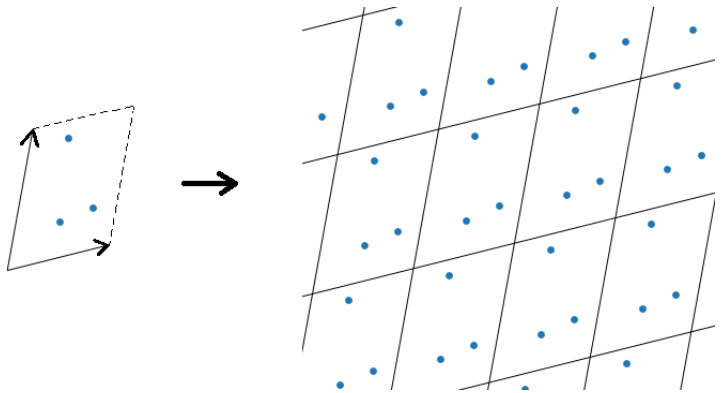
This problem has an elegant solution with the SSS theorem: two triangles are isometric if and only if they have the same 3 side lengths. All triangles live in the 3D cone below, with symmetric triangles in special subspaces of the cone.



We can take a cross-section (yellow) to ignore uniform scaling.

Definition of a periodic point set

A linear basis of $\mathbb{R}^3$ generates a *unit cell* U (parallelepiped on the basis vectors) and a lattice Λ consisting of all linear integer combinations of the basis vectors.

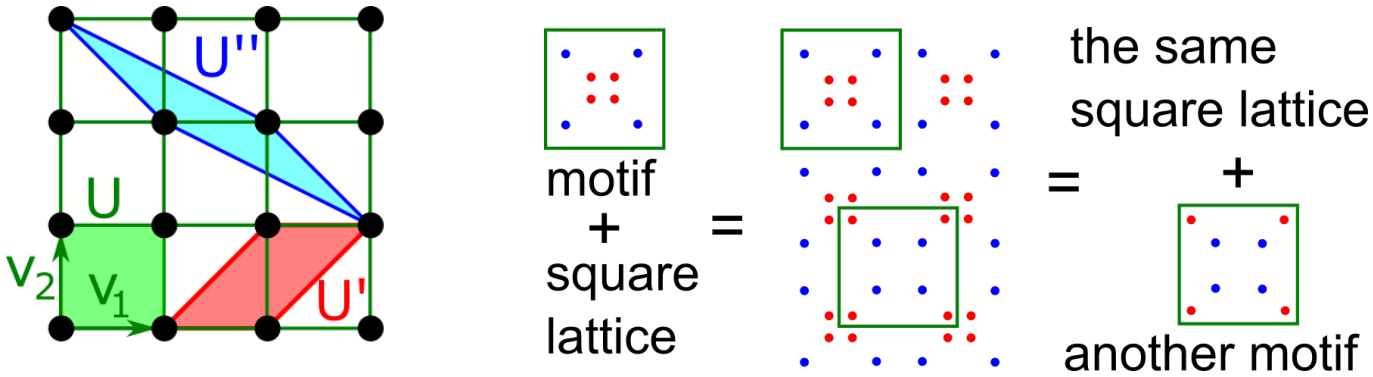


Given a *motif* M (a set of points in U representing atoms), a *periodic point set* is

$$\text{Periodic set } S = M + \Lambda = \{p + l : p \in M, l \in \Lambda\}$$

In practice, S is given by a (basis of a) cell and motif of atoms with coordinates in the basis of U.

Ambiguity of a cell and motif



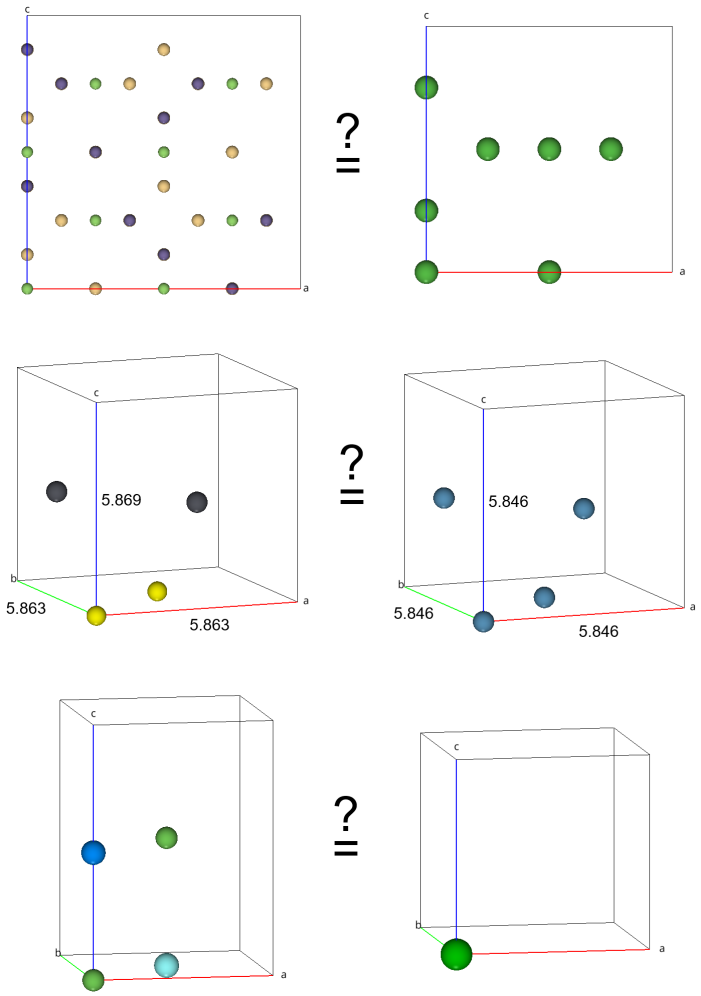
Any periodic set can be defined by many choices of cell and motif, so the collection of all (cell, motif) pairs doesn't form the space we would like.

The collection of all periodic sets also doesn't work because any two periodic sets related by a *rigid motion* (translation + rotation) are different but should be considered equivalent. In this work, we consider *isometries* which are slightly broader, including reflections.

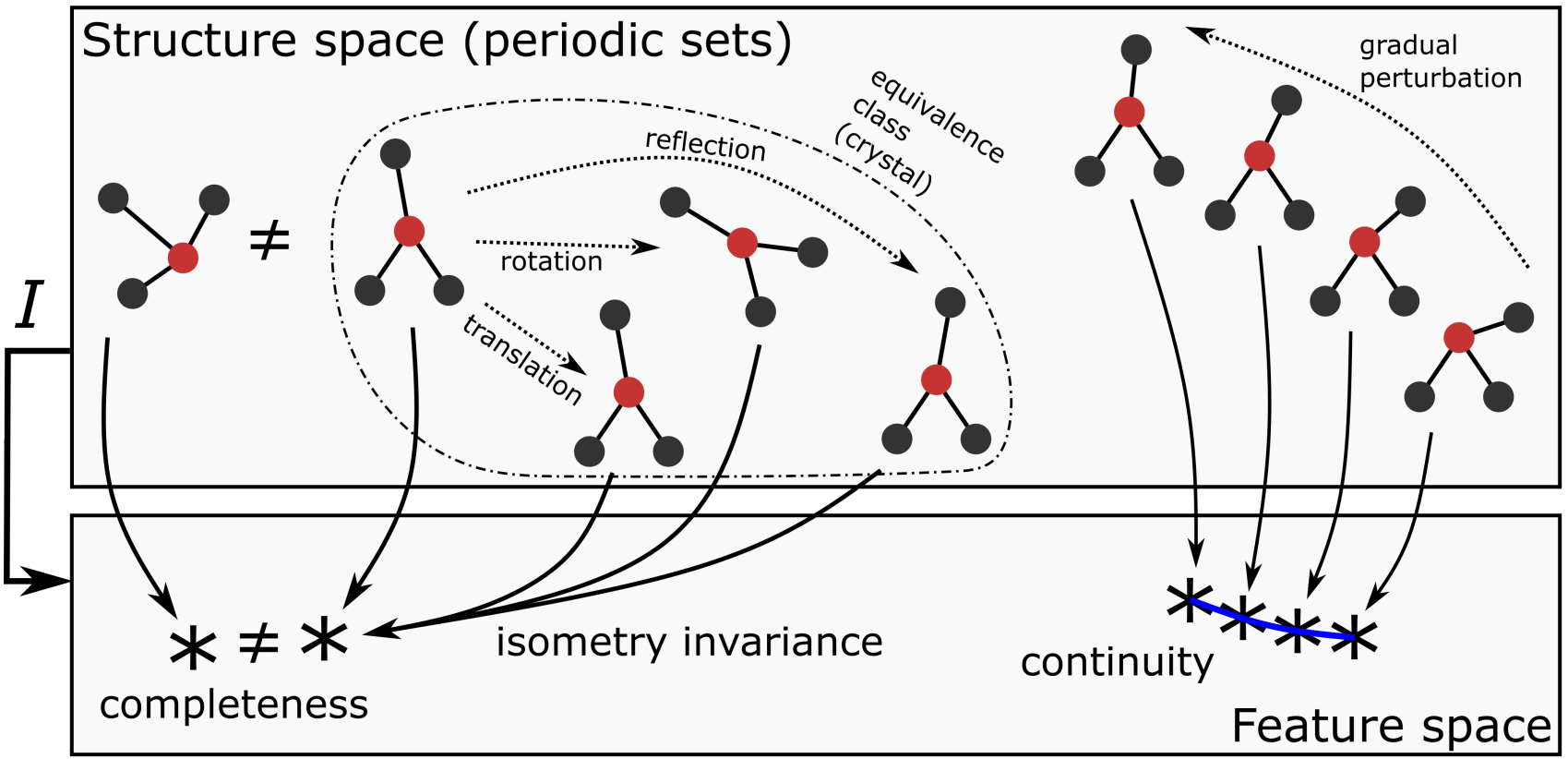
Crystals: same or different?

We want to avoid reliance on the unit cell or motif, which can be different for the same crystal. We also want *continuity*, so if atoms move and break symmetry (even in addition to changing the unit cell), it's always seen as a smooth deformation.

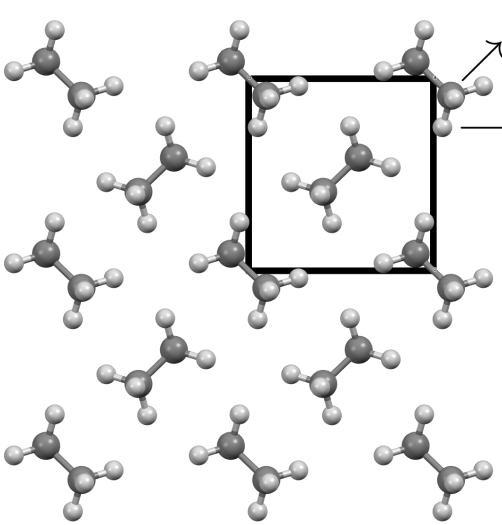
Slightly deforming an equilateral triangle moves it slightly away from the equilateral line in triangle space, that distance can be measured.



Continuous classification of crystals



Pointwise Distance Distributions (PDD): definition

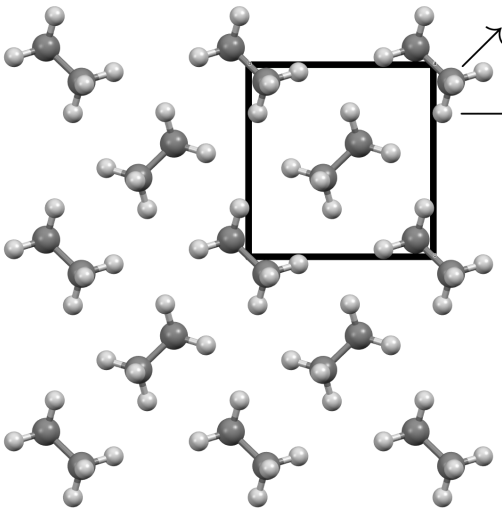


ETHANE01

1. distances to neighbours

The *Pointwise Distance Distribution* (PDD) of a crystal records the interatomic distances for each atom in a way that satisfies most of our conditions.

Pointwise Distance Distributions (PDD): definition



ETHANE01

$1.096, 1.096, 1.096, 1.532, \ldots$

$\longrightarrow 1.096, 1.762, 1.810, 2.160, \ldots$

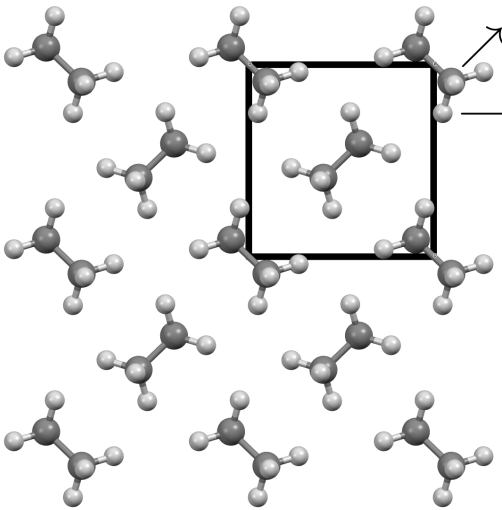
1. distances to neighbours

2. collapse identical lists, give weights

weight	1st	2nd	3rd	4th		10th
0.25	1.096	1.096	1.096	1.532	...	3.253
0.25	1.096	1.762	1.81	2.16	...	2.849
0.25	1.096	1.776	1.81	2.152	...	2.84
0.25	1.096	1.762	1.776	2.189	...	2.849

If any lists are identical (usually due to symmetry), collapse them into one and assign weights proportional to the frequency.

PDD: example computation



ETHANE01

$1.096, 1.096, 1.096, 1.532, \ldots$

$\longrightarrow 1.096, 1.762, 1.810, 2.160, \ldots$

1. distances to neighbours

2. collapse identical lists, give weights

weight	1st	2nd	3rd	4th		10th
0.25	1.096	1.096	1.096	1.532	...	3.253
0.25	1.096	1.762	1.81	2.16	...	2.849
0.25	1.096	1.776	1.81	2.152	...	2.84
0.25	1.096	1.762	1.776	2.189	...	2.849

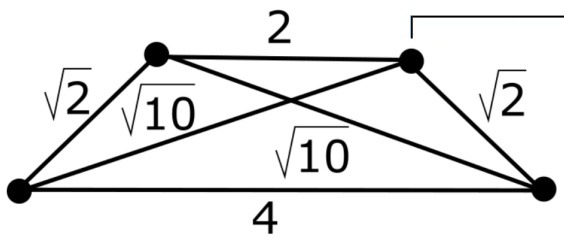
3. order lexicographically

PDD(ETHANE01, 10) =

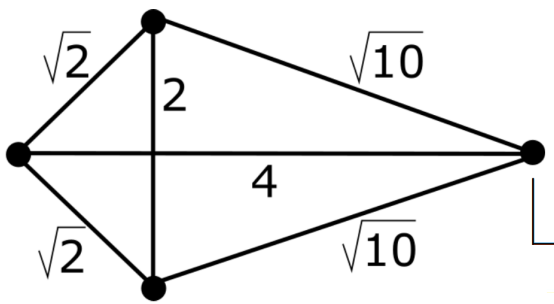
weight	1st	2nd	3rd	4th		10th
0.25	1.096	1.096	1.096	1.532	...	3.253
0.25	1.096	1.762	1.776	2.189	...	2.849
0.25	1.096	1.762	1.81	2.16	...	2.849
0.25	1.096	1.776	1.81	2.152	...	2.84

PDD distinguishes homometric sets

PDD distinguishes two finite point sets with the same inter-point distances:



$$\begin{pmatrix} \sqrt{2} & 2 & \sqrt{10} \\ \sqrt{2} & 2 & \sqrt{10} \\ \sqrt{2} & \sqrt{10} & 4 \\ \sqrt{2} & \sqrt{10} & 4 \end{pmatrix} \longrightarrow \left(\begin{array}{c|ccc} 1/2 & \sqrt{2} & 2 & \sqrt{10} \\ 1/2 & \sqrt{2} & \sqrt{10} & 4 \end{array}\right)$$



$$\begin{pmatrix} \sqrt{2} & \sqrt{2} & 4 \\ \sqrt{2} & 2 & \sqrt{10} \\ \sqrt{2} & 2 & \sqrt{10} \\ \sqrt{10} & \sqrt{10} & 4 \end{pmatrix} \longrightarrow \left(\begin{array}{c|ccc} 1/4 & \sqrt{2} & \sqrt{2} & 4 \\ 1/2 & \sqrt{2} & 2 & \sqrt{10} \\ 1/4 & \sqrt{10} & \sqrt{10} & 4 \end{array}\right)$$

PDD: theoretical properties

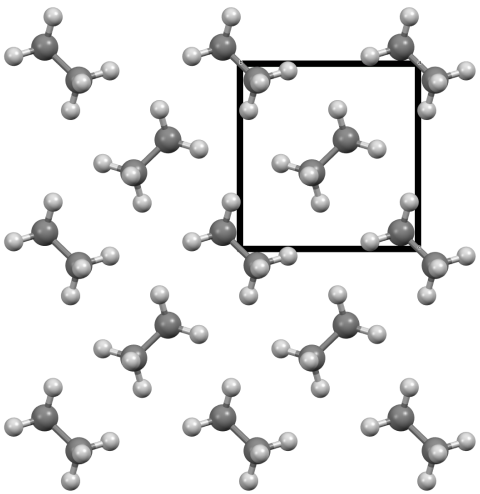
- Since a PDD is just an arrangement of inter-point distances, it's an *invariant* independent of any choice of cell and motif.
- PDDs are continuous in the metric EMD (*Earth mover's distance*), even if symmetries break under perturbation:

$$\text{EMD}(\text{PDD}(S;\, k),\, \text{PDD}(Q;\, k)) \leq 2d_B(S, Q)$$

- PDDs are *computable* in near-linear time $O(\nu(S; n)km \log(km))$, and performs well practically ($\sim$1ms/structure for $k = 100$).
- PDDs are *generically complete*, meaning any generic crystal (without repeated distances except for periodicity) can be reconstructed from a lattice (whose complete isometry invariants we have described in dimensions 2,3) and PDD large enough such that distances in the final column are greater than two times the covering radius of S.

Average minimum distances

Taking a weighted average over the columns of a PDD gives a simpler invariant called *average minimum distance* (AMD):



PDD ⇨

weight	1st	2nd	3rd	4th		10th
0.25	1.096	1.096	1.096	1.532	...	3.253
0.25	1.096	1.762	1.776	2.189	...	2.849
0.25	1.096	1.762	1.81	2.16	...	2.849
0.25	1.096	1.776	1.81	2.152	...	2.84

⇩ weighted average

$$\text{AMD} = \begin{pmatrix} 1.096 & 1.599 & 1.623 & 2.008 & \dots & 2.947 \end{pmatrix}$$

AMDs are faster to compare than PDDs, but were still able to distinguish all 800,000+ real periodic crystals in the world's largest collection of experimental materials CSD (Cambridge Structural Database).

Practical and theoretical results

- Distances as given by AMD and PDD are related:

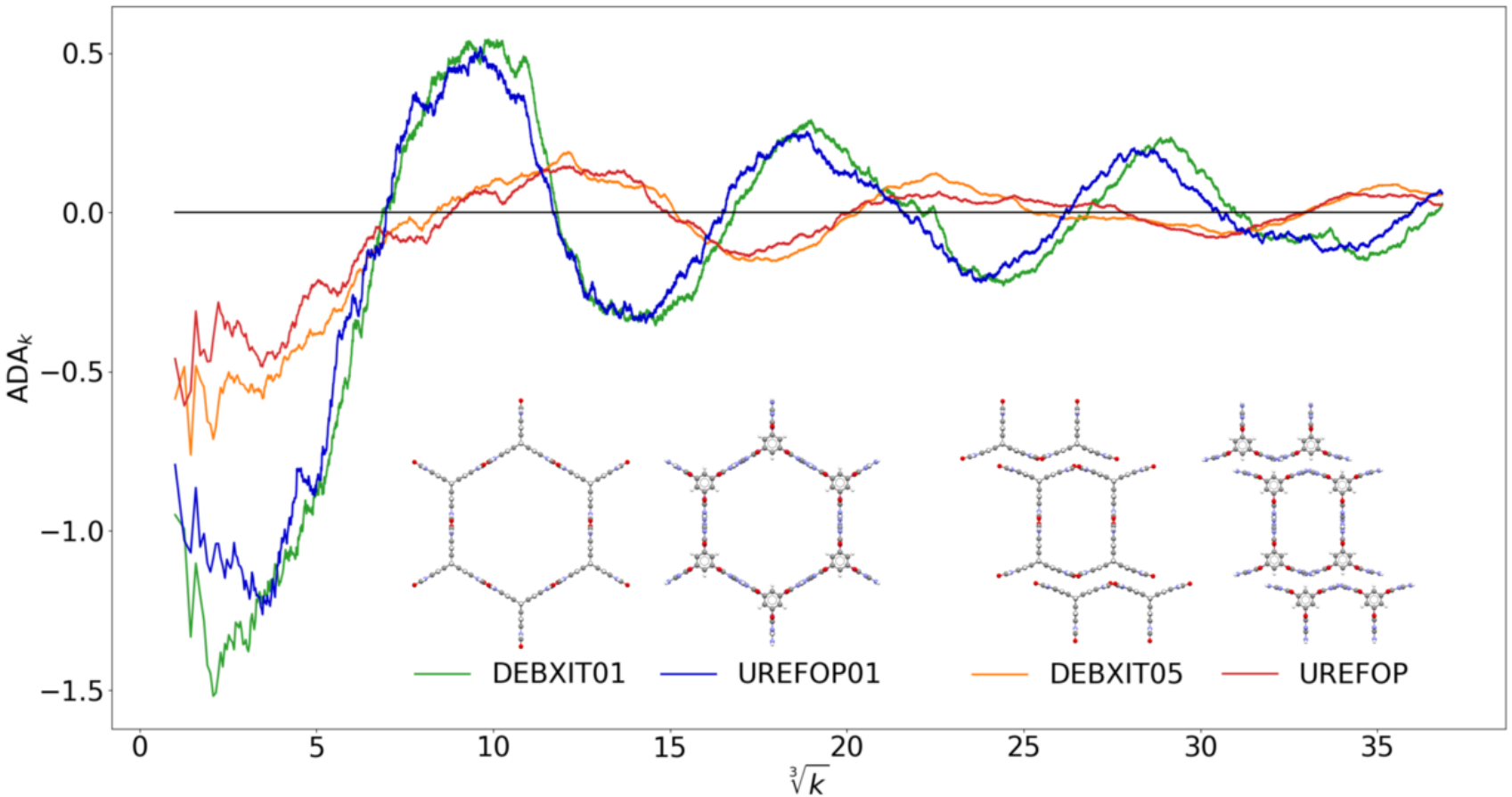
$$||\mathrm{AMD}(S;k) - \mathrm{AMD}(Q;k)|| \leq \mathrm{EMD}_{||\cdot||}(\mathrm{PDD}(S,k), \mathrm{PDD}(Q,k))$$

- AMD and PDD values converge to $\sqrt[n]{\frac{\mathrm{Vol}(U)}{mV_n}}\sqrt[n]{k}$. The difference between AMD/PDD and this asymptote is called ADA/PDA.
- Comparing all 1,000,000+ entries from various crystal databases (on a modest desktop within an hour) produced lists of cross-references between the databases and detected several entries with copied geometry or other errors, some of which are under investigation.

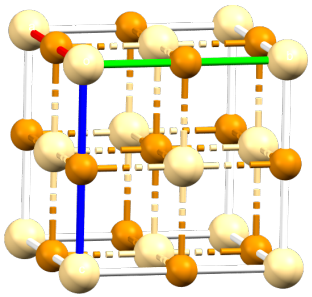
Published	C10	0.2690(9)	0.1893(2)	0.4328(4)
	C11	−0.0101(3)	−0.04935(6)	0.3551(1)
CSD PINHUP	C10 C	0.2690(9)	0.1893(2)	0.4328(4)
	Cl1 Cl	-0.0101(3)	-0.04935(6)	0.3551(1)

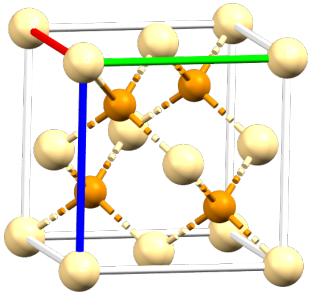
What does a crystal's 'coordinate' look like?

The plot below uses ADA = AMD minus its asymptotic.

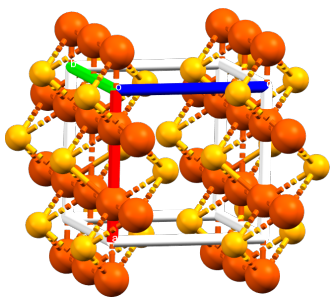


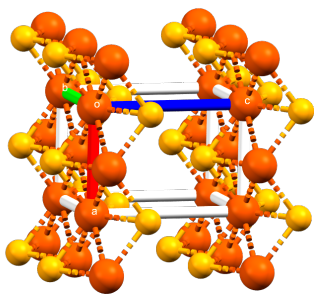
Investigating the COD's duplicate list



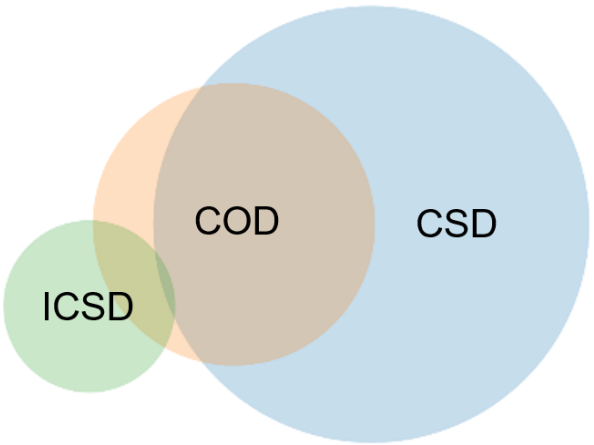


Above: COD 1010536 & 1010539 (CdTe) in 'duplicates.lst' with the same unit cell + composition. Below: COD 163560 & 977900 (FeSe), the same structure with different space groups, unit cells and motifs, *not* found in duplicates.lst.





Duplicates in and between databases



Left: accurate Venn diagram between 3 databases. Below: table of numbers of duplicates within 5 datasets. All invariants were computed in <30 mins; comparing them all took <10 mins.

Database	Entries	# Unique	% Redundancy	Largest group
CSD	1,268,165	1,265,499	0.21	37
COD	514,876	506,549	1.62	6
ICSD	300,019	296,624	1.13	18
MP	178,627	178,624	0.00	2
GNoME	384,938	383,737	0.31	5

References & Links

[1] Average minimum distances of periodic point sets – foundational invariants for mapping periodic crystals. MATCH Communications in Mathematical and in Computer Chemistry, 87(3), 529-559 (2022).

[2] Resolving the data ambiguity for periodic crystals. Neural Information Processing Systems (2022).

[3] Analogy Powered by Prediction and Structural Invariants: Computationally Led Discovery of a Mesoporous Hydrogen-Bonded Organic Cage Crystal. J. Am. Chem. Soc. 144, 22, 9893–9901 (2022).

[4] Continuous Invariant-Based Maps of the Cambridge Structural Database, arXiv:2108.04798.

[5] Pointwise distance distributions of periodic sets, arXiv:2108.04798.

Daniel Widdowson - sgdwiddo@liverpool.ac.uk
Vitaliy Kurlin - vkurlin@liverpool.ac.uk
Andrew I Cooper - aicooper@liverpool.ac.uk

Try AMD or PDD: github.com/dwiddo/average-minimum-distance