The Crystal Isometry Principle infers chemistry from geometry

Vitaliy Kurlin's Data Science group: Dan Widdowson, Yury Elkin, Olga Anosova, ...

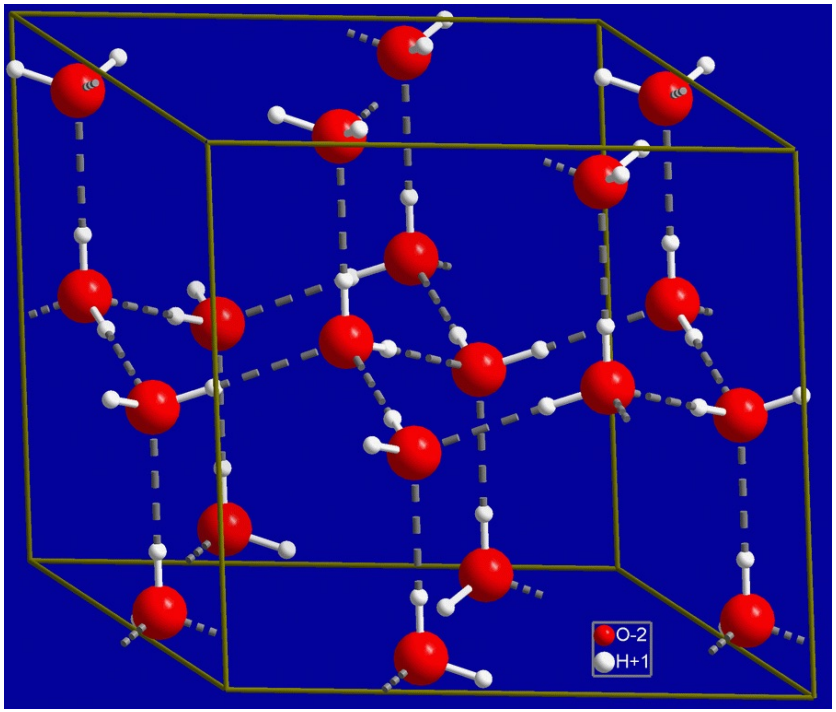
Materials Innovation Factory, Liverpool, UK

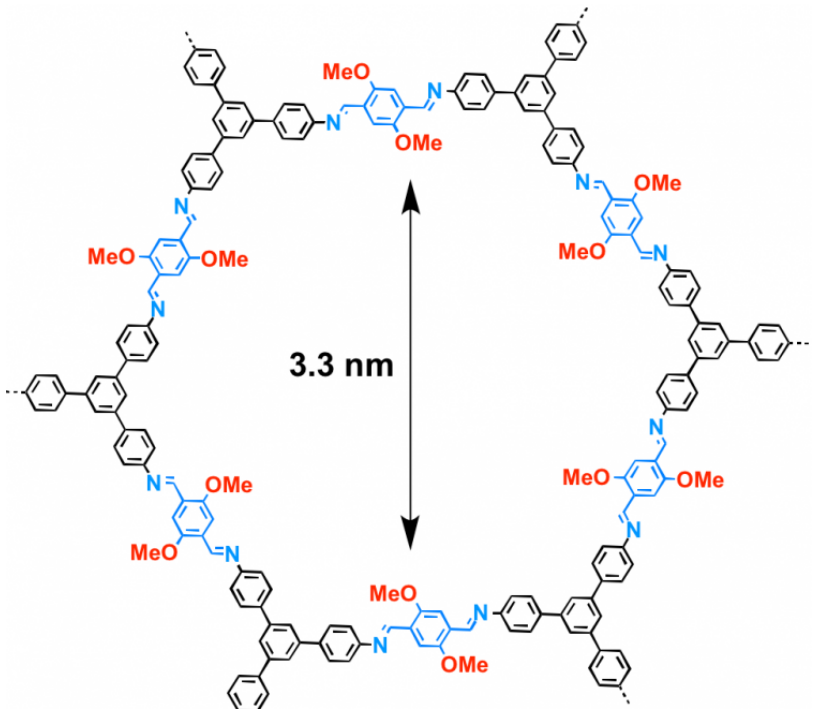
Royal Society APEX fellowship (2023-2025)

THE ROYAL SOCIETY

Objects: all periodic crystals

We study solid crystalline materials at the atomic level. What is a crystal on the left?



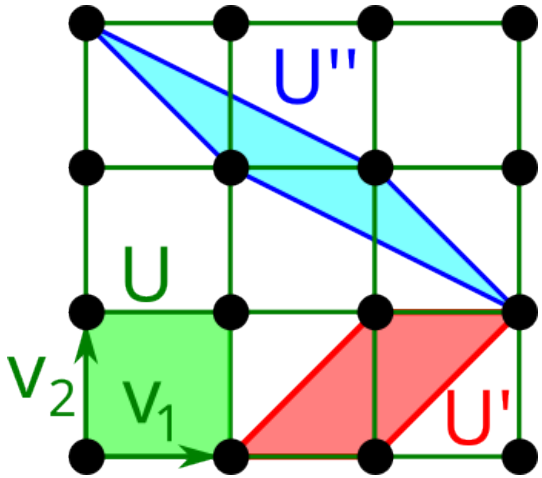


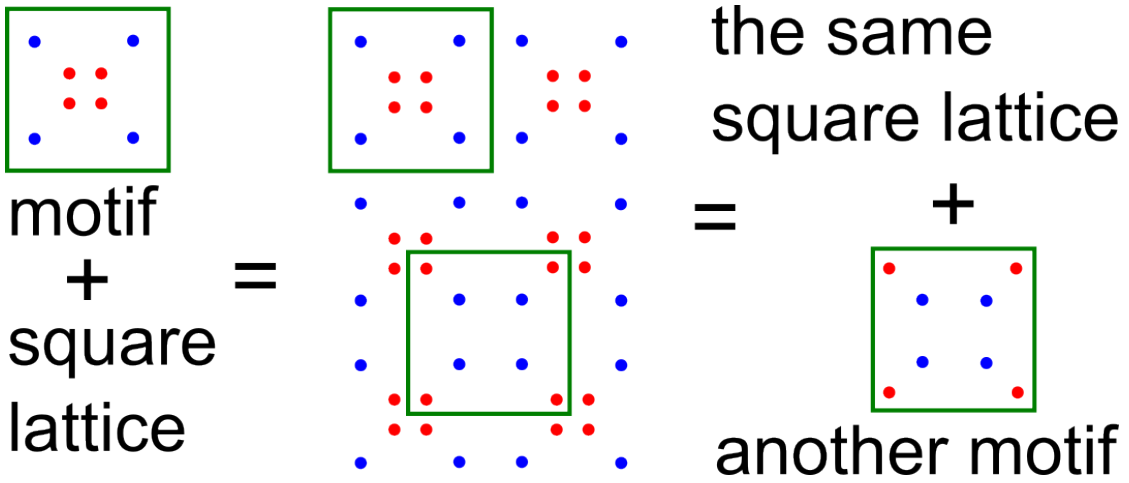
Questions: What is a crystal? What crystals are the same? If different, how much different?

A periodic point set (crystal)

Any basis $v_1, \ldots, v_n$ of $\mathbb{R}^n$ defines the *unit cell* U and generates the lattice $\Lambda = \{\sum_{i=1}^{n} c_i v_i \ : \ c_i \in \mathbb{Z}\}$.

For any finite *motif* $M \subset U$, the *periodic point set* is the sum $S = \Lambda + M = \{v + p \mid v \in \Lambda, p \in M\}$.





Different pairs (basis, motif) give equivalent sets.

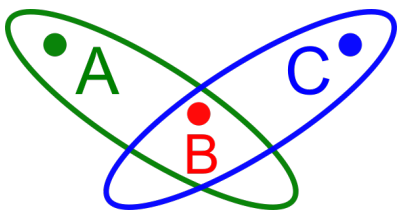
Three axioms of an equivalence

A relation $A \sim B$ between any data objects is called an *equivalence* if the three axioms hold:

(1) *reflexivity*: any object $A \sim A$;

(2) *symmetry*: if $A \sim B$ then $B \sim A$;

(3) *transitivity*: if $A \sim B$ and $B \sim C$, then $A \sim C$.



The transitivity axiom guarantees that all objects are in disjoint classes. Any justified classification needs an equivalence.

Equality is an equivalence: $0.5 = 50\% = \frac{1}{2} = 2 \div 4$

Different equivalence relations

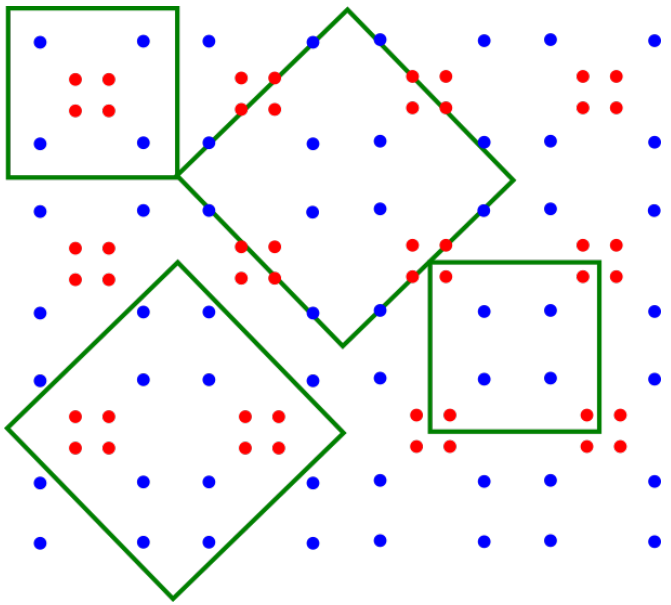
Chemical: crystals $A \sim B$ if A, B have the same composition. Ok, but diamond and graphite with vastly different properties are in the same class.

By property: crystals $A \sim B$ if A, B have the same property. Ok, but crystals that share one property can differ by many other properties.

Space-group types: crystals $A \sim B$ if A, B have isomorphic space groups. Fedorov (1891): 219 or 230 classes. Then NaCl, MgO, TiC, LaN, NaI, RbF, SrS, ... have the same group (225, $Fm\bar{3}m$).

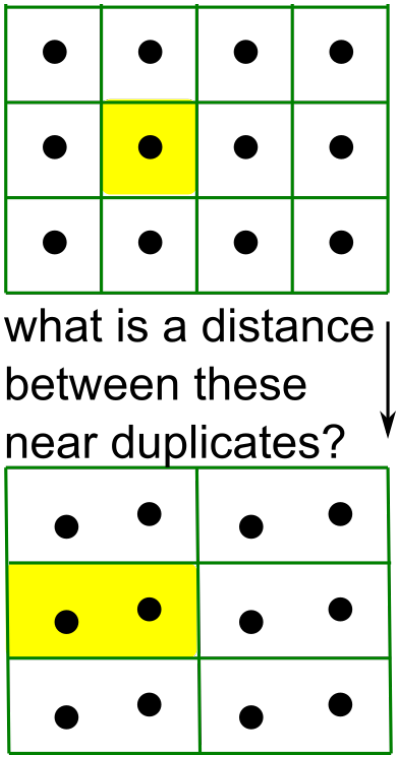
How about standard conventions?

IUCr online dictionary: “crystals are said to be *isostructural* if they have the same structure ... CaCO3, NaNO3, FeBO3 are isostructural”.



All conventional representations in the International Tables of Crystallography are *correct in theory* but are **no longer practical** because all data are noisy and tiny displacements of atoms need very different (standard) settings.

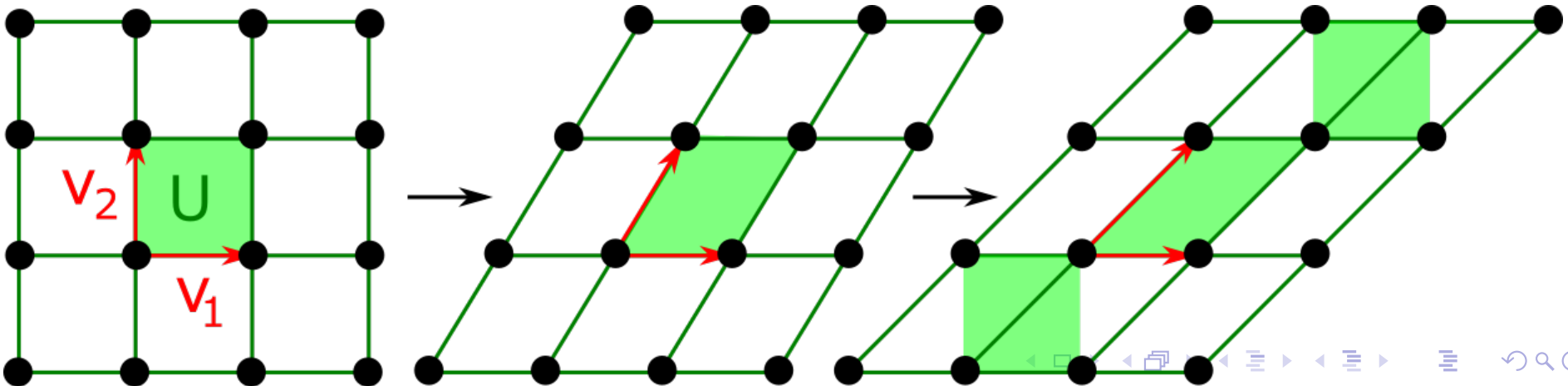
Discontinuity of conventional cells



Any *reduced or conventional* cell is discontinuous under noise and atomic displacements.

All discrete *symmetry-based crystallography* cannot continuously quantify a distance between crystals. RMSD, 1-PXRD and all others are discontinuous or fail the metric axioms.

Any *pseudo-symmetry* (equivalence up to a threshold > 0) leads to a trivial classification.



What is the strongest relation?

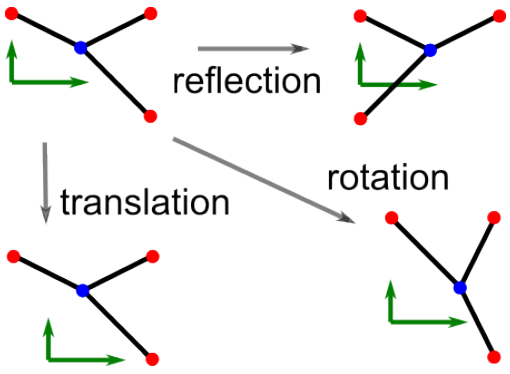
P. Sacchi et al. **Same or different - that is the question**: identification of crystal forms. CrystEngComm, 22(43), 7170-7185 (2020).

IUCr ACTIVITIES

CHANGE TO THE DEFINITION OF "CRYSTAL" IN THE IUCr ONLINE DICTIONARY OF CRYSTALLOGRAPHY

Definitions are not final without equivalence.

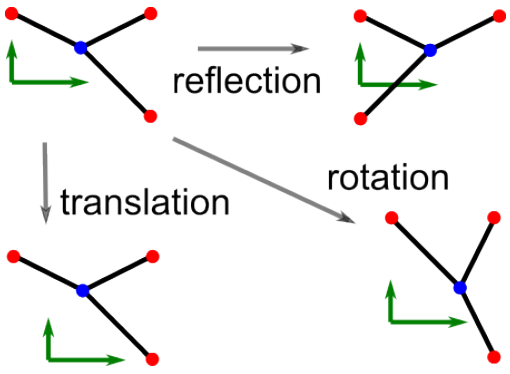
Definition of a crystal structure



Since crystal structures are determined in a *rigid form*, the strongest relation in practice is **rigid motion** = translations + rotations in $\mathbb{R}^3$.

Slightly weaker: **isometry** = rigid motion + reflections = any map preserving distances.

Definition of a crystal structure



Since crystal structures are determined in a *rigid form*, the strongest relation in practice is **rigid motion** = translations + rotations in $\mathbb{R}^3$.

Slightly weaker: **isometry** = rigid motion + reflections = any map preserving distances.

One Crystallographic Information File is not a **periodic structure** $=$ *rigid class* of crystals

$$= \frac{\text{infinitely many periodic crystals (CIFs) in } \mathbb{R}^3}{\textit{equivalence} \text{ under } \textit{rigid motion} \text{ (or isometry)}}$$

Descriptors vs isometry invariants

An **invariant** is a function I : { isometry classes of crystals } $\to$ {*a simpler space*} of numbers, vectors, ..., where comparisons are easier.

Crystals can be distinguished only by *invariants* taking the same value on all equivalent objects.

If $S \simeq Q$ are *isometric*, then $I(S) = I(Q)$; or

if $I(S) \neq I(Q)$, then $S \not\simeq Q$ are not isometric.

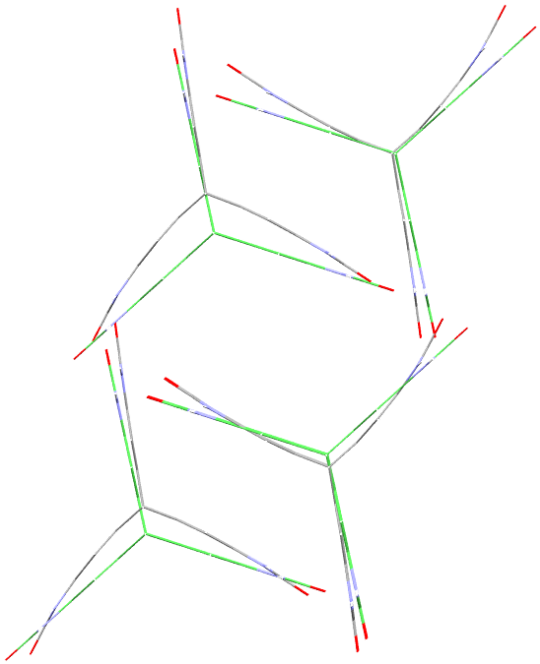
non-invariants
atomic coordinates in a cell basis, *cannot* distinguish crystas

invariants can distinguish some, possibly not all crystals: density, symmetry

complete invariants, e.g. conventional representations distinguish all *in theory*

continuous **fast & reconstructable**

Crystals live in a continuous space



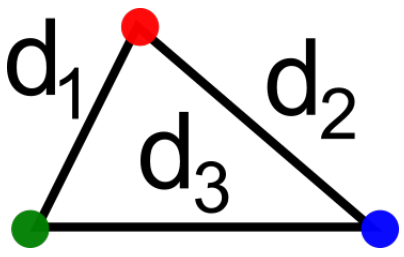
All crystals consist of discretely located atoms, which have *continuous* real-valued coordinates in $\mathbb{R}^3$.

A small perturbation produces a slightly different crystal not rigidly equivalent to the original structure.

If we restrict comparisons only to a fixed space group, we cut the continuous space into disjoint pieces (230 in 3D), so many near-duplicates fall on different side of boundaries, which is tragic!

Importance of metric axioms

A **metric** $d(S, Q)$ is a function on pairs (say, invariants of crystals) satisfying three axioms:



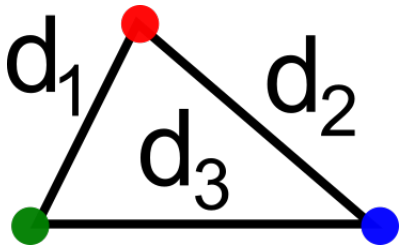
$d(I(S), I(Q)) = 0 \Leftrightarrow I(S) = I(Q)$,

symmetry $d(I(S), I(Q)) = d(I(Q), I(S))$,

triangle inequality $d_1 + d_2 \geq d_3$.

Importance of metric axioms

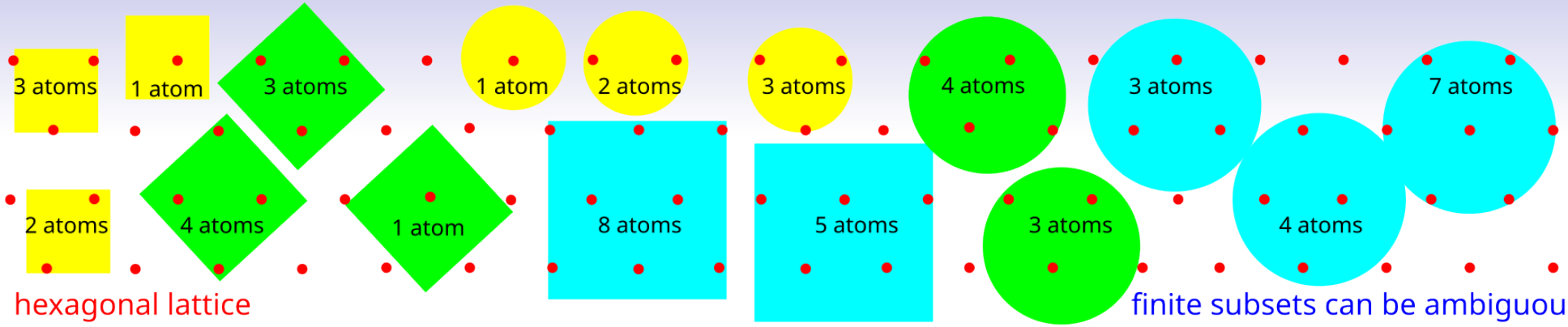
A **metric** $d(S, Q)$ is a function on pairs (say, invariants of crystals) satisfying three axioms:

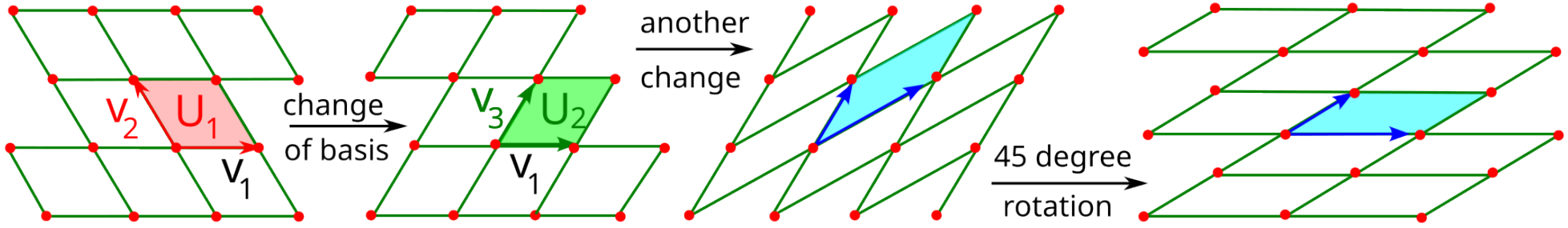


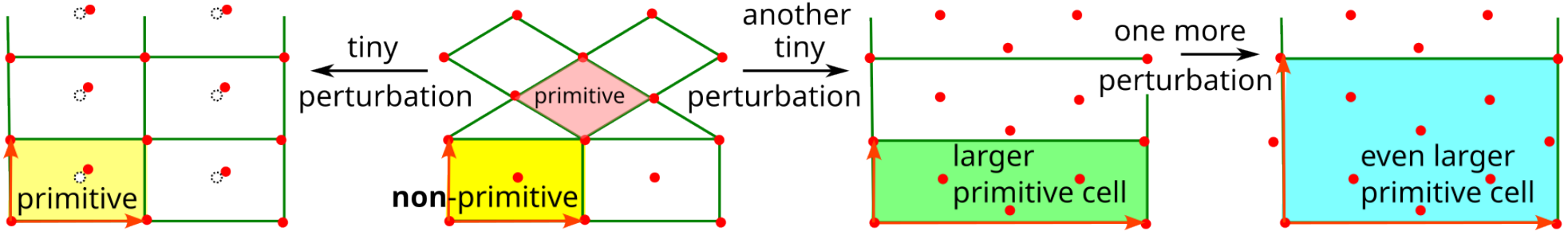
$d(I(S), I(Q)) = 0 \Leftrightarrow I(S) = I(Q)$,
symmetry $d(I(S), I(Q)) = d(I(Q), I(S))$,
triangle inequality $d_1 + d_2 \geq d_3$.

The discrete distance is a discontinuous metric: $d(S, Q) = 0$ for equivalent $S \simeq Q$, else $d = 1$.

Rass et al (2024): if we allow the $\triangle$ inequality to *fail with any small error*, the results of k-means clustering and DBSCAN can be pre-determined.







any periodic crystal ■ has infinitely many **near-duplicates with larger motifs**

■ the *smallest subspace* of crystals with m atoms in a fixed cell

a *larger subspace* of crystals with 2m atoms in a primitive cell

an even larger subspace of crystals with 3m atoms in a primitive cell

infinitely many layers in the continuous space of periodic crystals

Isometry classification problem

Find an easy continuous and complete isometry invariant I for discrete sets of *unordered points*.

Invariance: if point sets $S \simeq Q$ are isometric, then $I(S) = I(Q)$, so I should be well-defined on isometry classes or I has *no false negatives*.

Completeness: if $I(S) = I(Q)$, then $S \simeq Q$ are isometric, hence I has *no false positives*.

Continuity: find a *metric* d and a constant λ such that if any point of S is perturbed within its ε-neighborhood, then $I(S)$ changes by max $\lambda\varepsilon$.

Harder practical requirements

Reconstruction (inverse design): any $S \subset \mathbb{R}^n$ can be reconstructed from its invariant $I(S)$.

Harder practical requirements

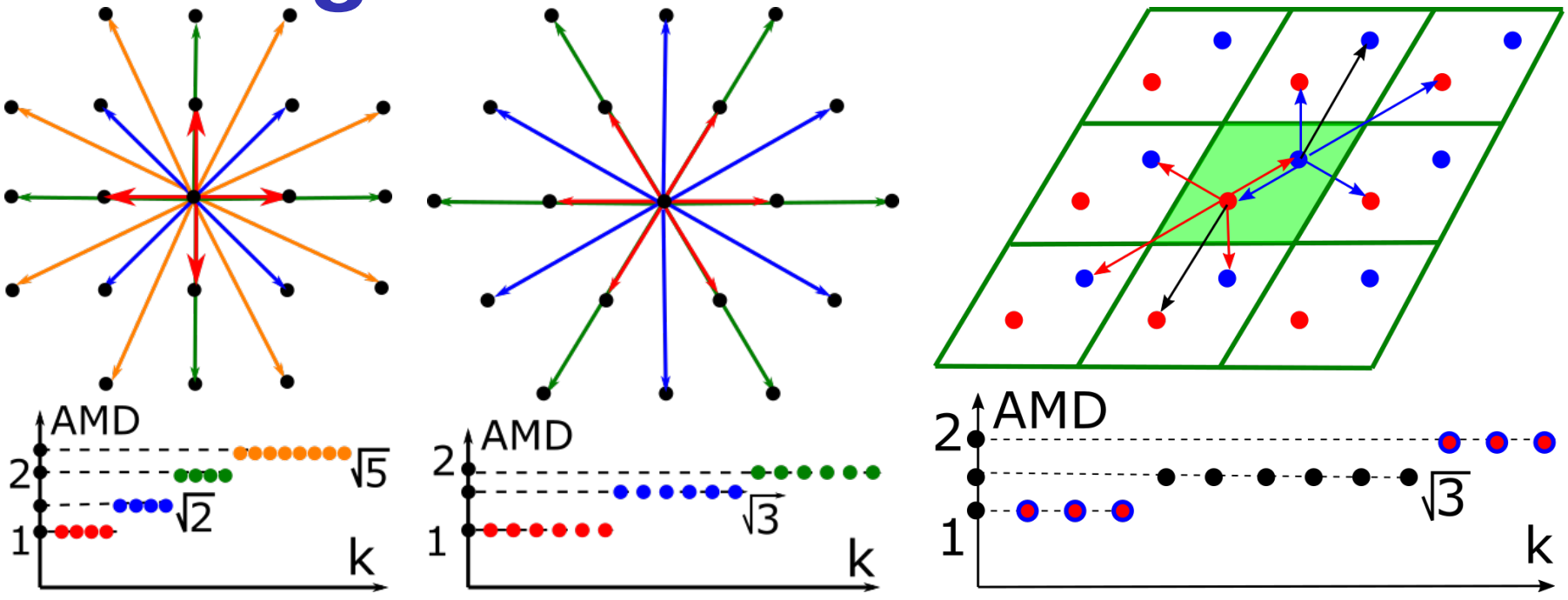
Reconstruction (inverse design): any $S \subset \mathbb{R}^n$ can be reconstructed from its invariant $I(S)$.

Computability: I, d, and reconstruction of S from $I(S)$ can be obtained in polynomial time in the motif size (number of atoms in a unit cell), hence *no infinite/exponential size* invariants.

If all conditions hold, I is *universal* for all types of periodic crystals, independent of symmetry.

If I is simple enough, I defines geographic-style coordinates on the space of all periodic crystals.

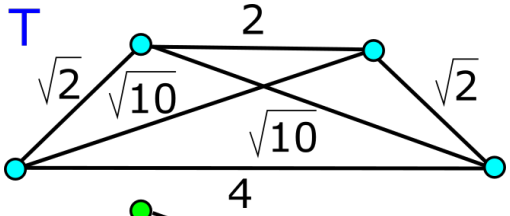
Average Minimum Distance: AMD



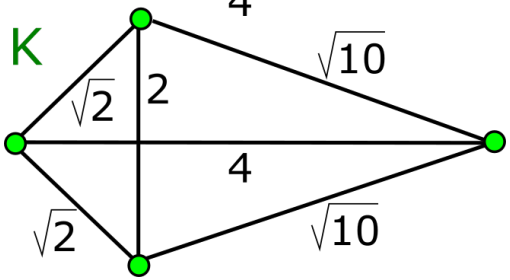
For a finite or periodic set $S \subset \mathbb{R}^n$, let d_{ik} be the distance from a point p_i in a motif, $i = 1, \ldots, m$, to its k-th nearest neighbor in S. For $k \geq 1$, *Average Minimum Distance* $\mathrm{AMD}_k = \frac{1}{m} \sum_{i=1}^{m} d_{ik}$.

Stronger invariants (finite case)

For each of m points in a finite set S, we write distances to k nearest neighbours in increasing order in the $m \times k$ matrix, so unordered points of S are mapped 1-1 to unordered rows.



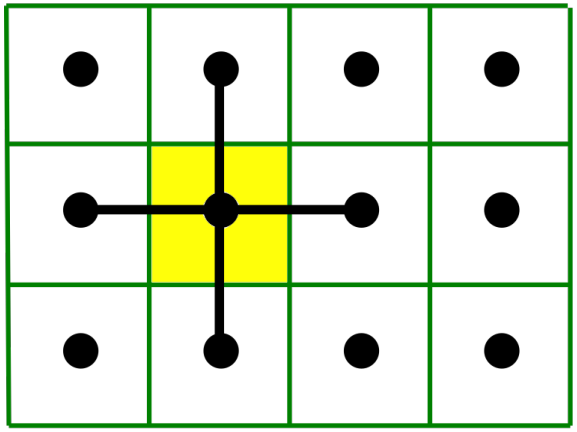
$$\mathrm{PDD}(T;3) = \left(\begin{array}{c|ccc} 1/2 & \sqrt{2} & 2 & \sqrt{10} \\ 1/2 & \sqrt{2} & \sqrt{10} & 4 \end{array}\right) \neq$$



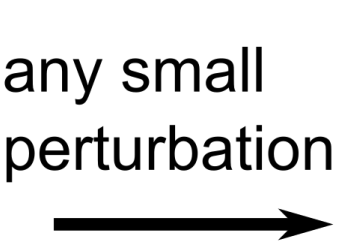
$$\mathrm{PDD}(K;3) = \left(\begin{array}{c|ccc} 1/4 & \sqrt{2} & \sqrt{2} & 4 \\ 1/2 & \sqrt{2} & 2 & \sqrt{10} \\ 1/4 & \sqrt{10} & \sqrt{10} & 4 \end{array}\right).$$

Collapse any l identical rows in the matrix and assign the weight l/m in the 1st extra column.

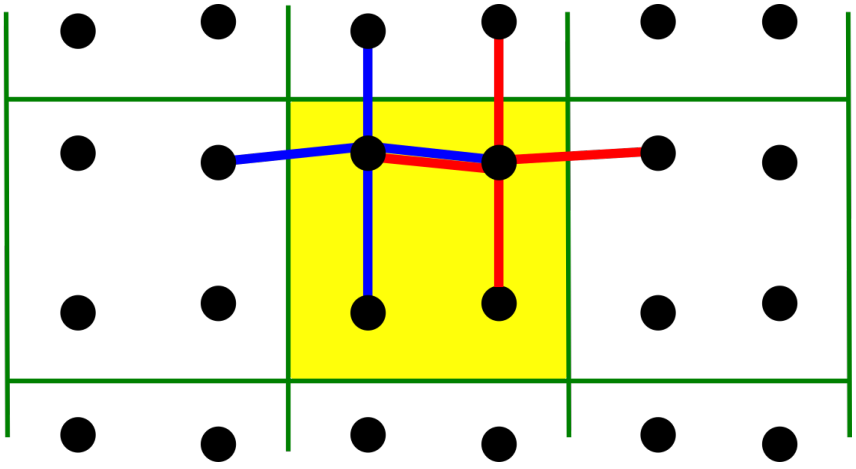
Earth Mover's Distance (EMD)



any small perturbation



continuously affects PDD



Thm (continuity). If we perturb all points of a set S within their ε-neighbourhoods, the perturbed set S' has $\mathrm{EMD}(\mathrm{PDD}(S;k),\mathrm{PDD}(S';k)) \leq 2\varepsilon$.

PDD(S;4)=

weight 1	1	1	1	1

PDD(S';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

EMD minimises a cost of matching weighted rows.

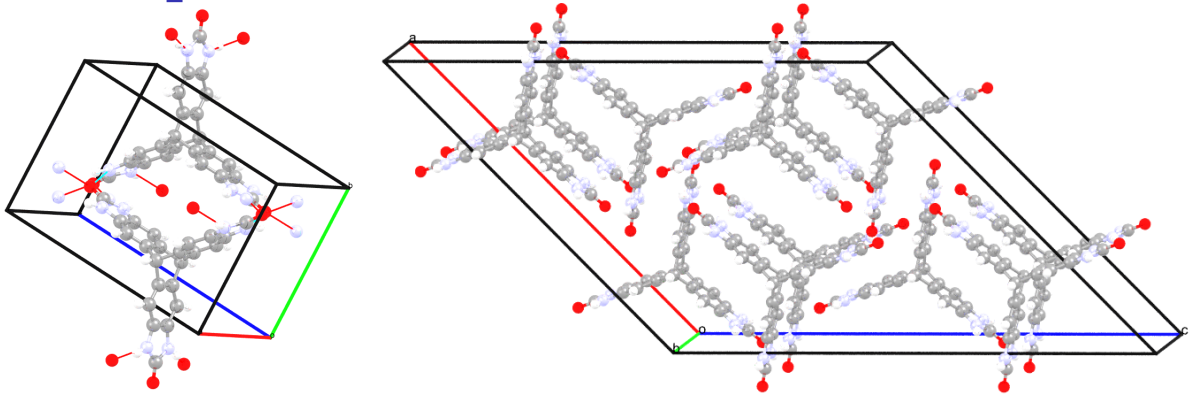
Key results from NeurIPS 2022

Increasing a number k of neighbors only adds more columns, k is a *degree of approximation*.

Thm (strength). Any *generic* periodic point set S (with distinct inter-point distances ignoring periodicity) can be uniquely reconstructed from a lattice of S and $\mathrm{PDD}(S;k)$ with distances up to a double covering radius of S in any $\mathbb{R}^n$.

Thm (time). For any finite or periodic set S with m motif points in $\mathbb{R}^n$, $\mathrm{PDD}(S;k)$ is computable in *near-linear* time $O(km\log(m)\log^2 k)$ for fixed n.

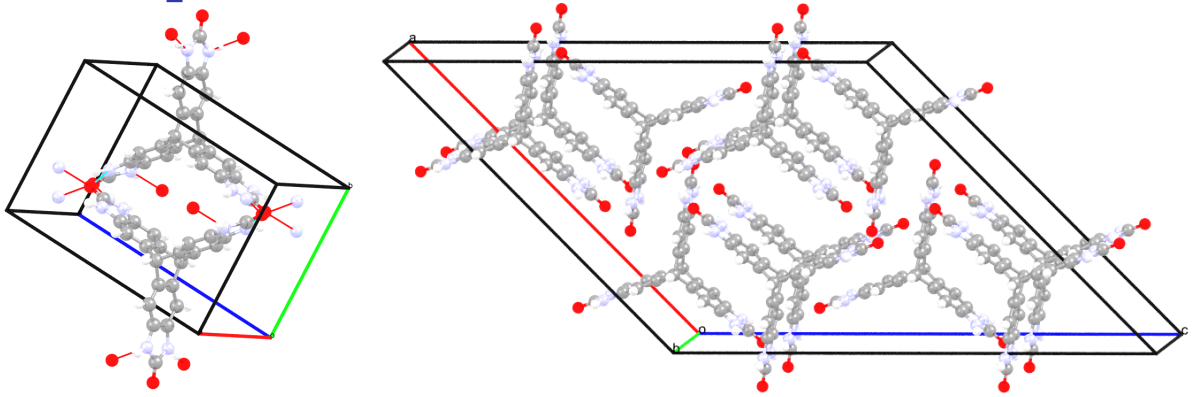
5 pairs of ‘needles in a haystack’



T2-14 vs T2-15 crashed Platon comparisons.

200B+ pairwise comparisons of PDD invariants over *two days on a modest desktop* for 660K+ periodic crystals in the Cambridge Structural Database (CSD)

5 pairs of ‘needles in a haystack’



T2-14 vs T2-15 crashed Platon comparisons.

200B+ pairwise comparisons of PDD invariants over *two days on a modest desktop* for 660K+ periodic crystals in the Cambridge Structural Database (CSD) detected *five isometric pairs* with different chemistry, which seems physically impossible, under investigation by 5 journals:

HIFCAB vs JEPLIA (one atom Cd $\leftrightarrow$ Mn), ...

Detecting (near-)duplicates

CSD Mercury's RMSD (on 15 molecules) was estimated to require *1000+ years* for all pairwise comparisons on the same desktop computer.

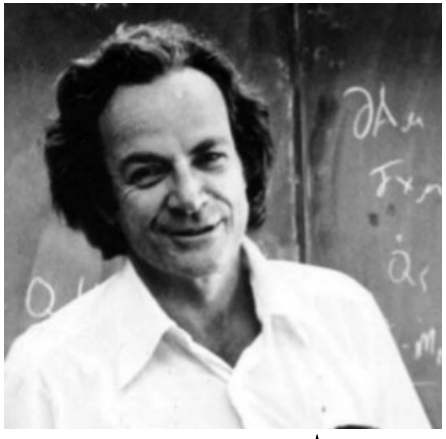
All energy minimization can output many approximations to the *same local minimum*.

Loophole: take the CIF (and structure factors) of a real crystal, change (or double) a unit cell, perturb atoms (to get a new motif in a larger primitive cell), swap atoms, and claim as *new*.

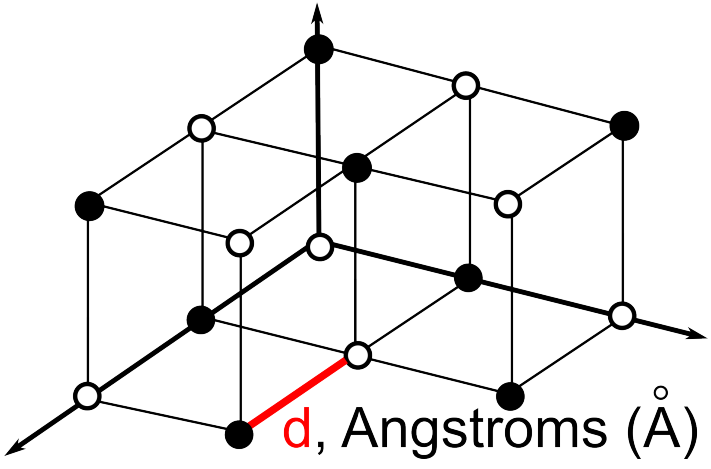
Olga's talk on Google's GNoME: Tuesday 4pm.

CRISP: Crystal Isometry Principle

Map: {**any crystal**} $\rightarrow$ {**set of atomic centres**} sends *different crystals* to *non-isometric sets*, checked for all periodic crystals in the CSD, so chemistry *reduces* to geometry.

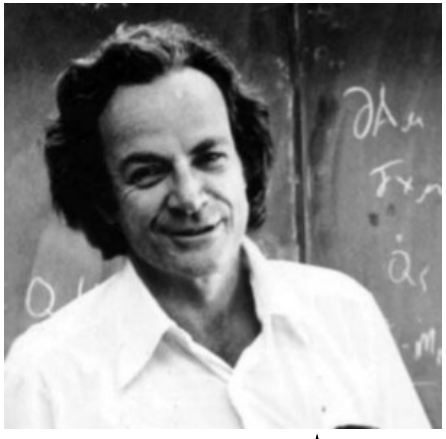


●	○	d (Å)
Na	Cl	2.82
K	Cl	3.14
Ag	Cl	2.77
Mg	O	2.10
Pb	S	2.98
Pb	Se	3.07
Pb	Te	3.17

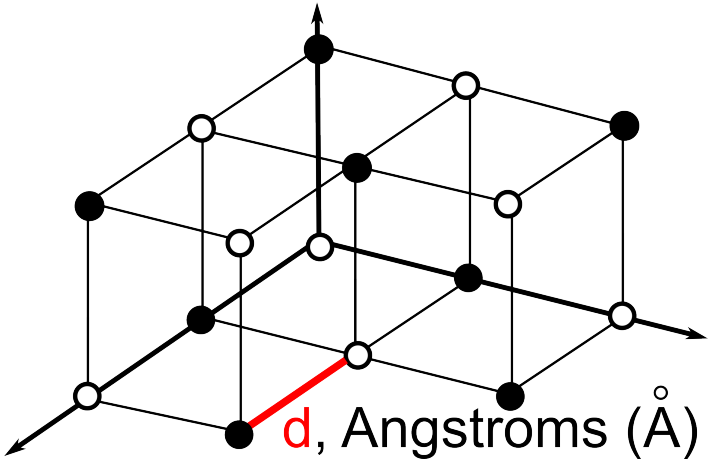


CRISP: Crystal Isometry Principle

Map: {**any crystal**} $\rightarrow$ {**set of atomic centres**} sends *different crystals* to *non-isometric sets*, checked for all periodic crystals in the CSD, so chemistry *reduces* to geometry. All known and *undiscovered* periodic crystals live in the *Crystal Isometry Space* (**CRIS**) of isometry classes of periodic sets. All real crystals are 'visible stars' in this *continuous* **crystal universe**.



●	○	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



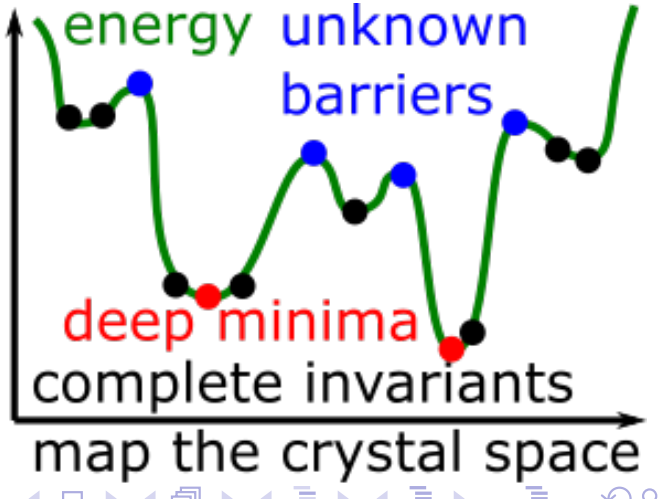
Vision of the crystal universe

In the past: only peaks of height $= -$energy, no locations.

currently available CSP landscapes consist of only isolated dots without any metric information

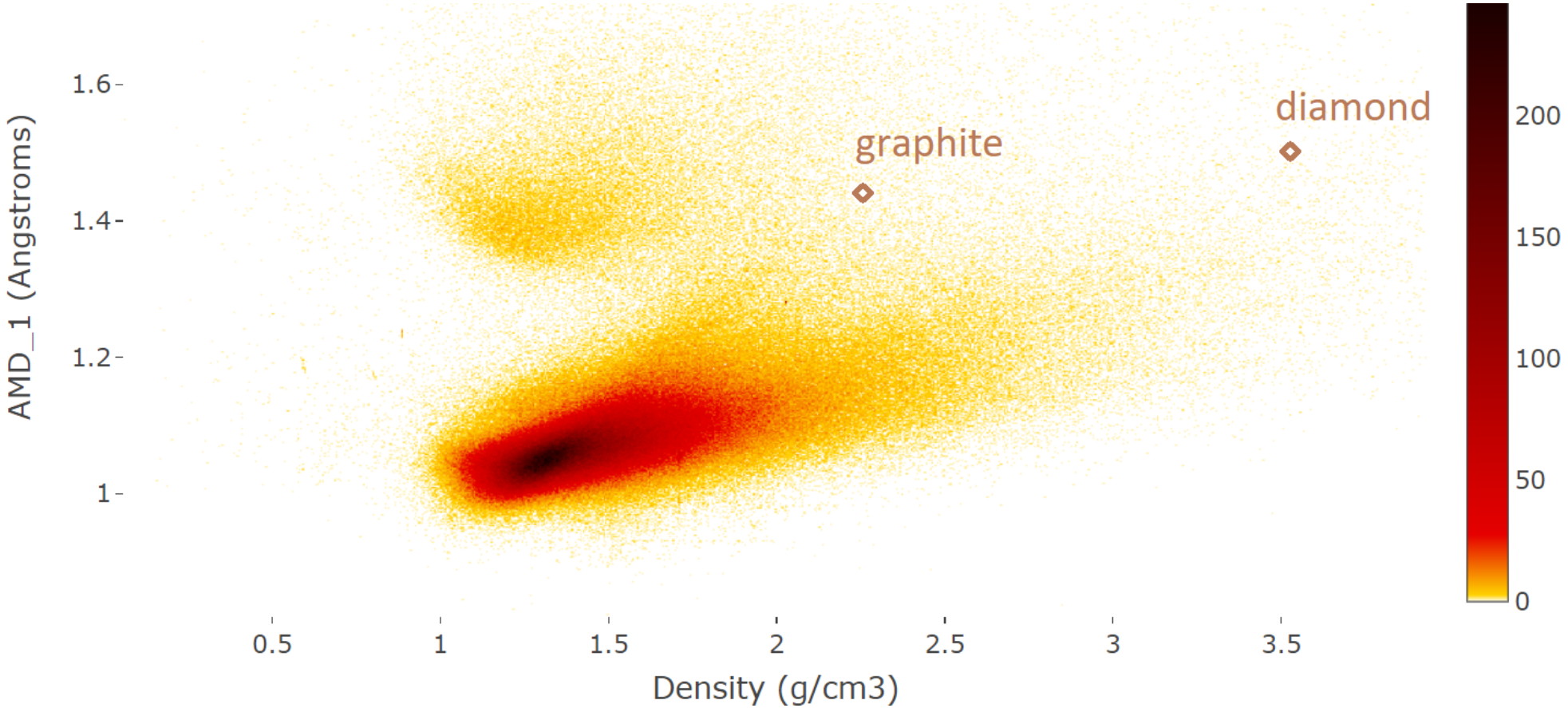
transition paths between minima

join the dots by distances



CSD in meaningful coordinates

$\mathrm{AMD}_1 =$ average distance to 1st atomic neighbour.

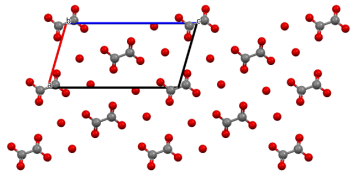


It's a projection with well-defined coordinates.

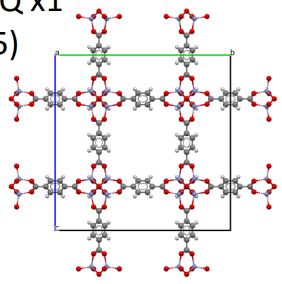
Any crystal has a *unique location* on such maps.

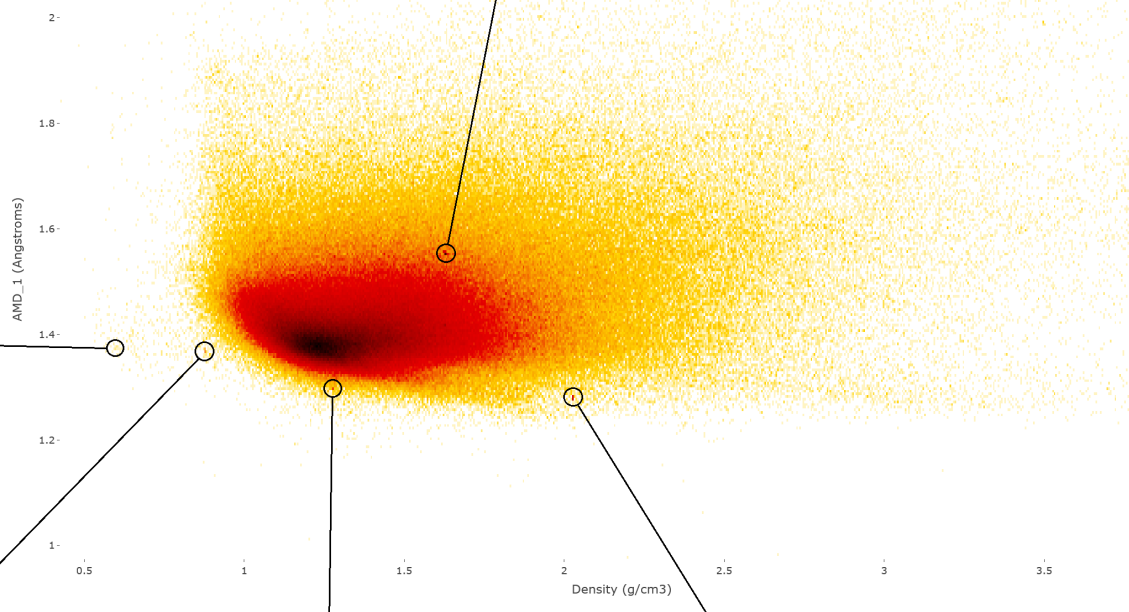
All visible artefacts are explained

OXALOH x100, OXALCH x92,
OXALBH x93, OXACDH x71,
OXACBH x66
(oxalic acid dihydrate)

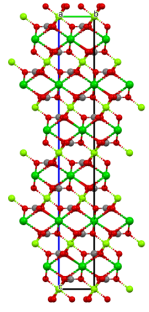


MIBQAR x18,
SAHYOQ x1
(MOF-5)

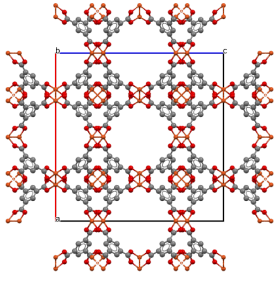


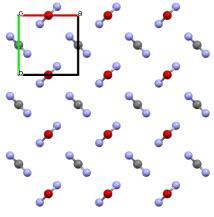


YAYVOM x20



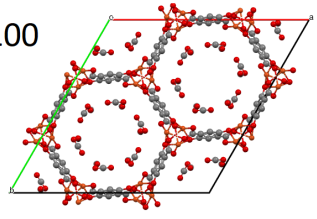
DOTSOV x45





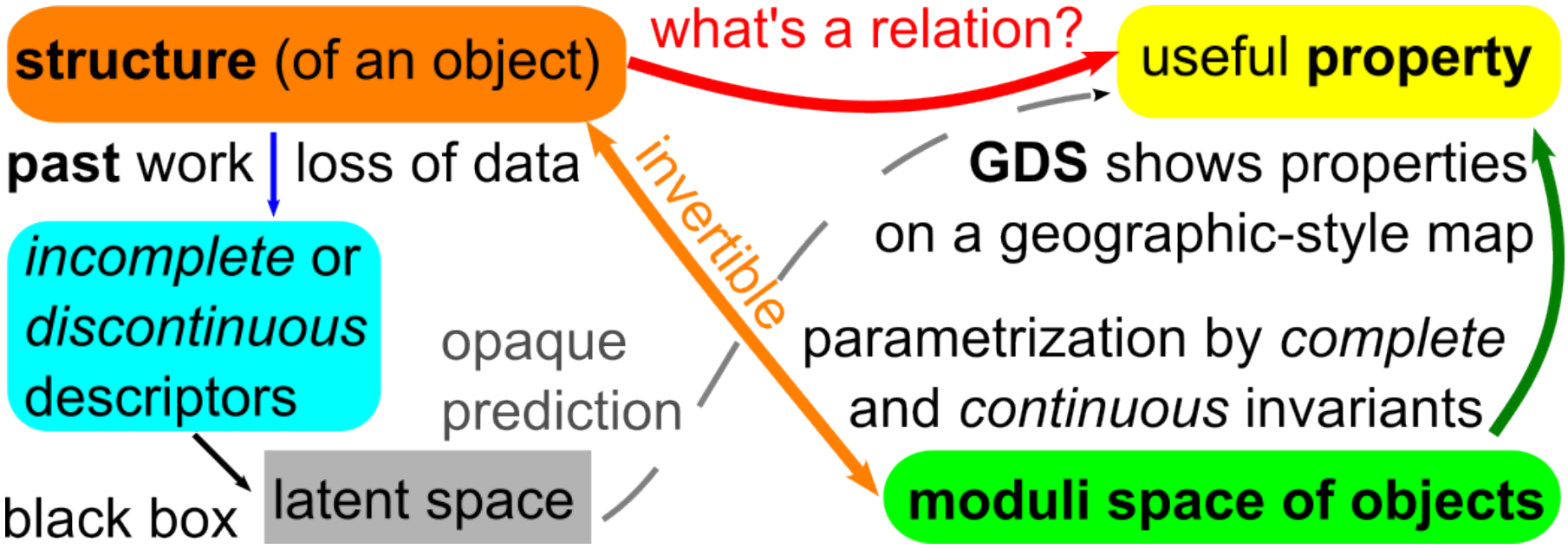
UREAXX x40
(urea)

MOKYAP x18, MOJPOT x100
(MOF CPO-27-Cu)

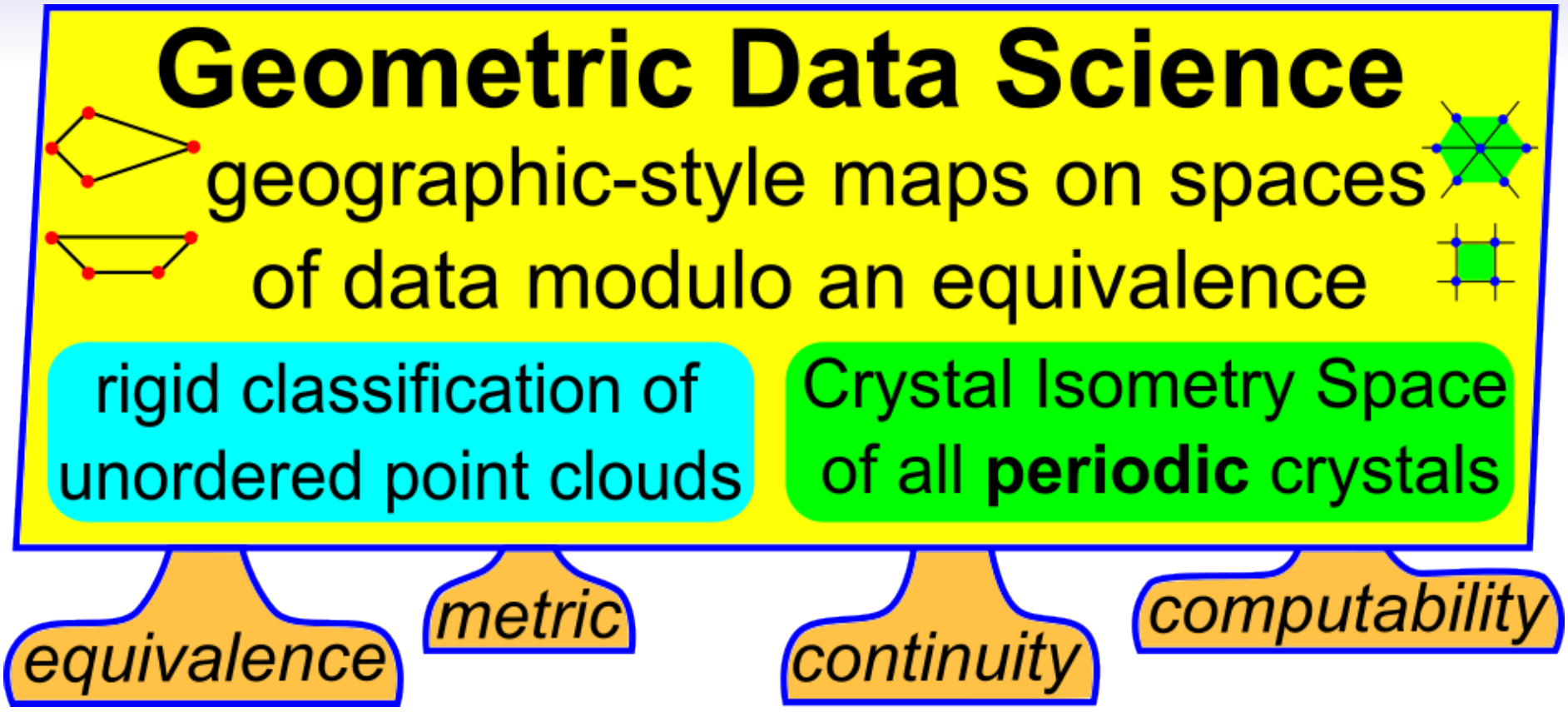


Geo-graphic-style maps in GDS

Non-invariant (or incomplete or discontinuous) descriptors *lose data* in *cramped latent spaces*.



Geometric Data Science puts all (equivalence classes of) real data objects at *uniquely defined locations* on a continuous map (moduli space).



The **key GDS problem** (*complete, continuous, computable, and realizable* invariants) makes sense for **any data** objects (instead of crystals) and **equivalence** relation (instead of isometry).