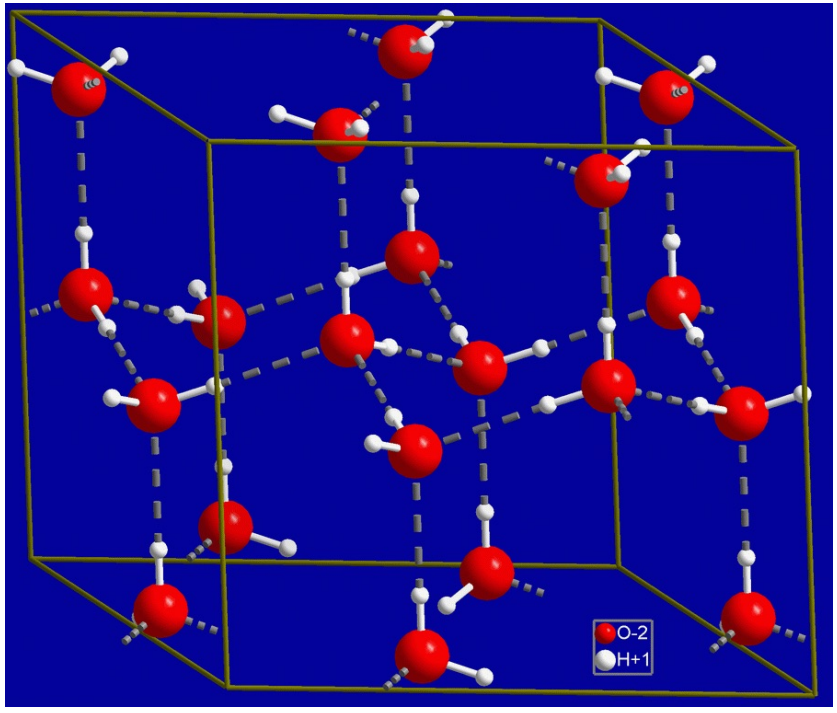
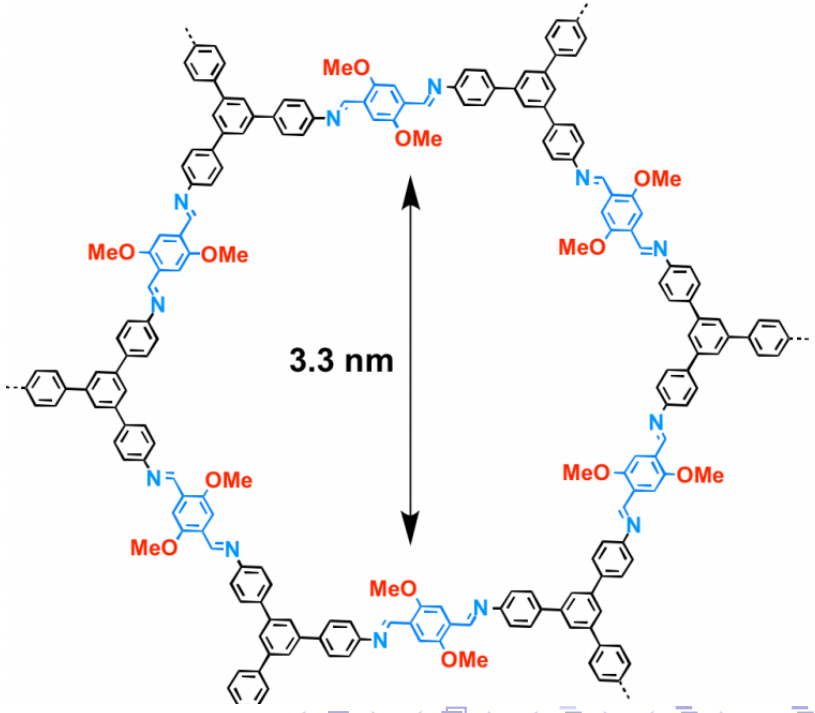
Mathematical Data Science for crystals

Vitaliy Kurlin's group including
Phil Smith, Matt Bright, Dan Widdowson, ...
Materials Innovation Factory (MIF), Liverpool
Royal Acad. Engineering Ind. Fellow, CCDC

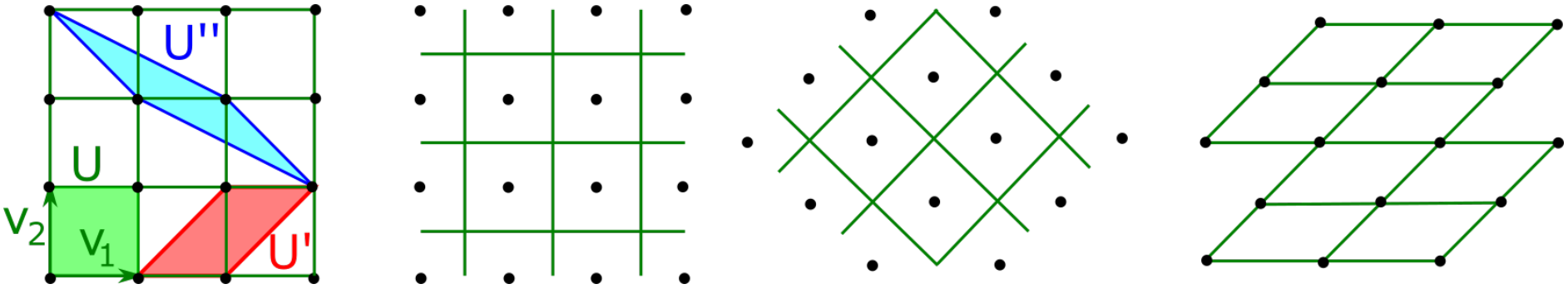
Objects: all periodic crystals

Solid crystalline materials (periodic *crystals*) can have many types, all consist of elementary blocks (*motifs*) of atoms, ions or molecules in a *unit cell* periodically repeated in three directions.

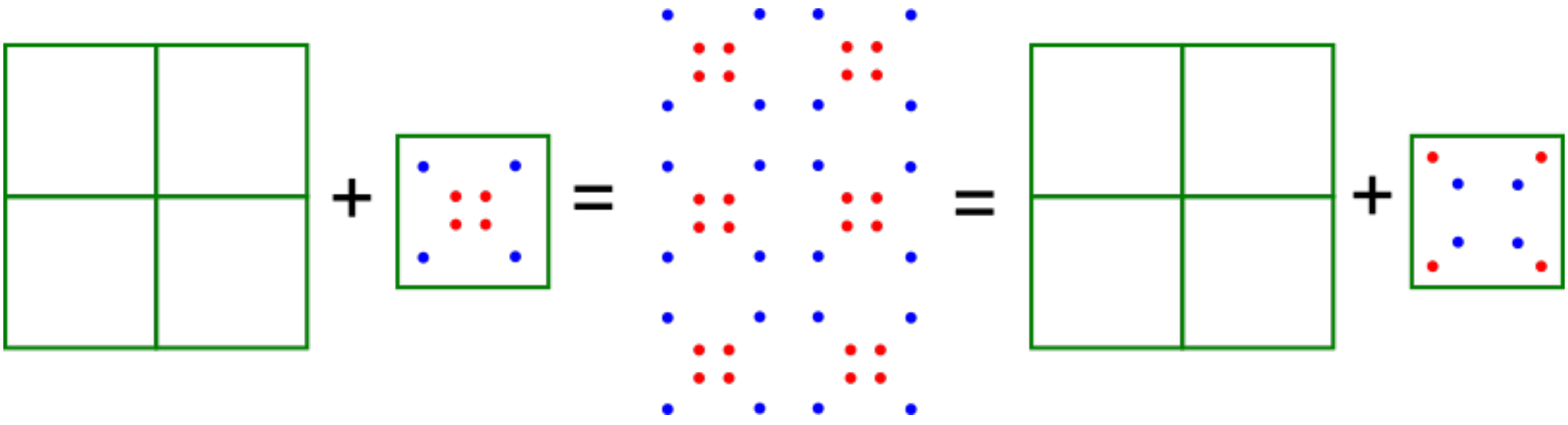




Ambiguity of inputs (cell, motif)



Are the above lattices different or equivalent?



Even if we fix a cell, input ambiguity remains.

Past equivalences of crystals

What crystals are *equivalent*? An equivalence has three axioms: $A \sim A$, if $A \sim B$ then $B \sim A$, transitivity: if $A \sim B \sim C$ then $A \sim C$ (needed for a non-trivial splitting into well-defined classes).

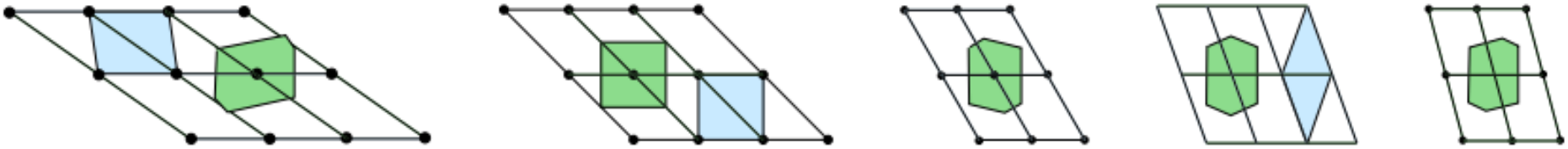
By symmetry group: 230 classes are known, insuffucient to classify 1M+ crystals in the CSD.

Many crystals are often called *similar*. When is such a similarity an *equivalence relation*?

What if crystals have similar density or energy?

Similarity by perturbation

Nomenclature of inorganic structure types (ACA 1990) uses many parameters, but any similarity threshold > 0 makes the classification trivial.



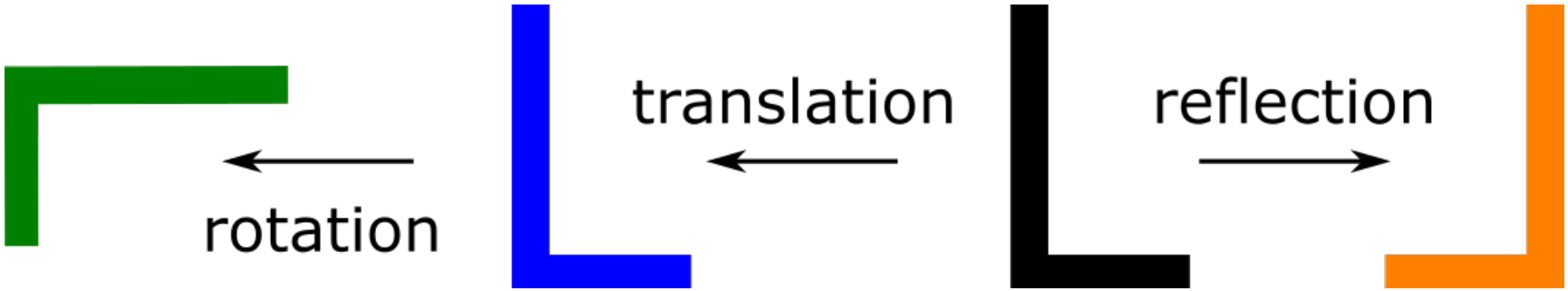
Assume $A \sim B$ if A can be perturbed to B by a small $d > 0$. Then any A, B are joined by a chain of small perturbations $A \sim A_1 \sim \cdots \sim A_n \sim B$, so A, B are equivalent by the transitivity axiom.

We can compare any crystals, not only similar.

Crystals up to isometry

Crystal structures are determined in a **rigid** form and should be studied up to *rigid motion* (a composition of translations and rotations in 3D).

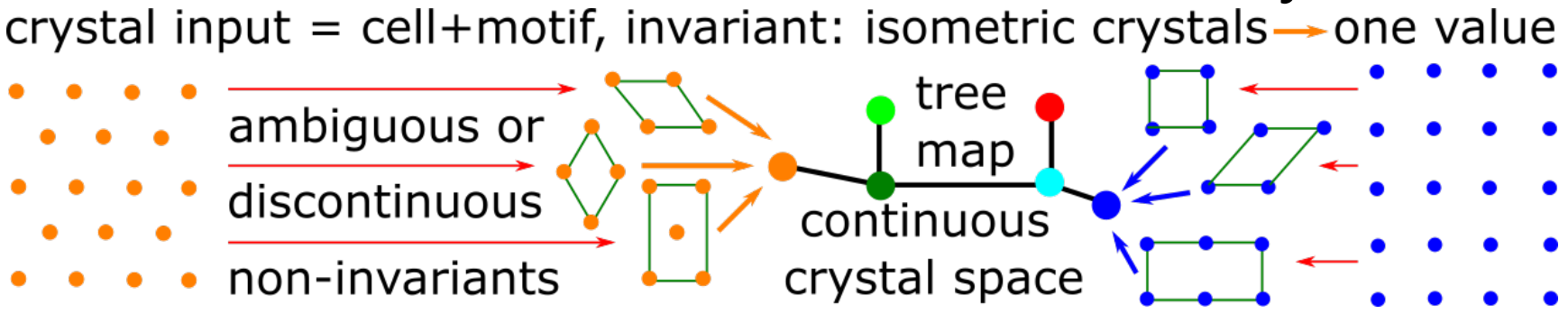
Isometries also include mirror reflections.



Hence a *crystal* is not a single set, but a class of *infinitely many periodic point sets* equivalent to each other up to isometry or rigid motion in 3D.

How can we distinguish crystals?

An **invariant** (number, vector, matrix,...) must take the **same value** on all isometric crystals.



If a **non-invariant** takes two different values on two crystals, then **no conclusion** can be made.

Question: how about non-invariant *big data*?

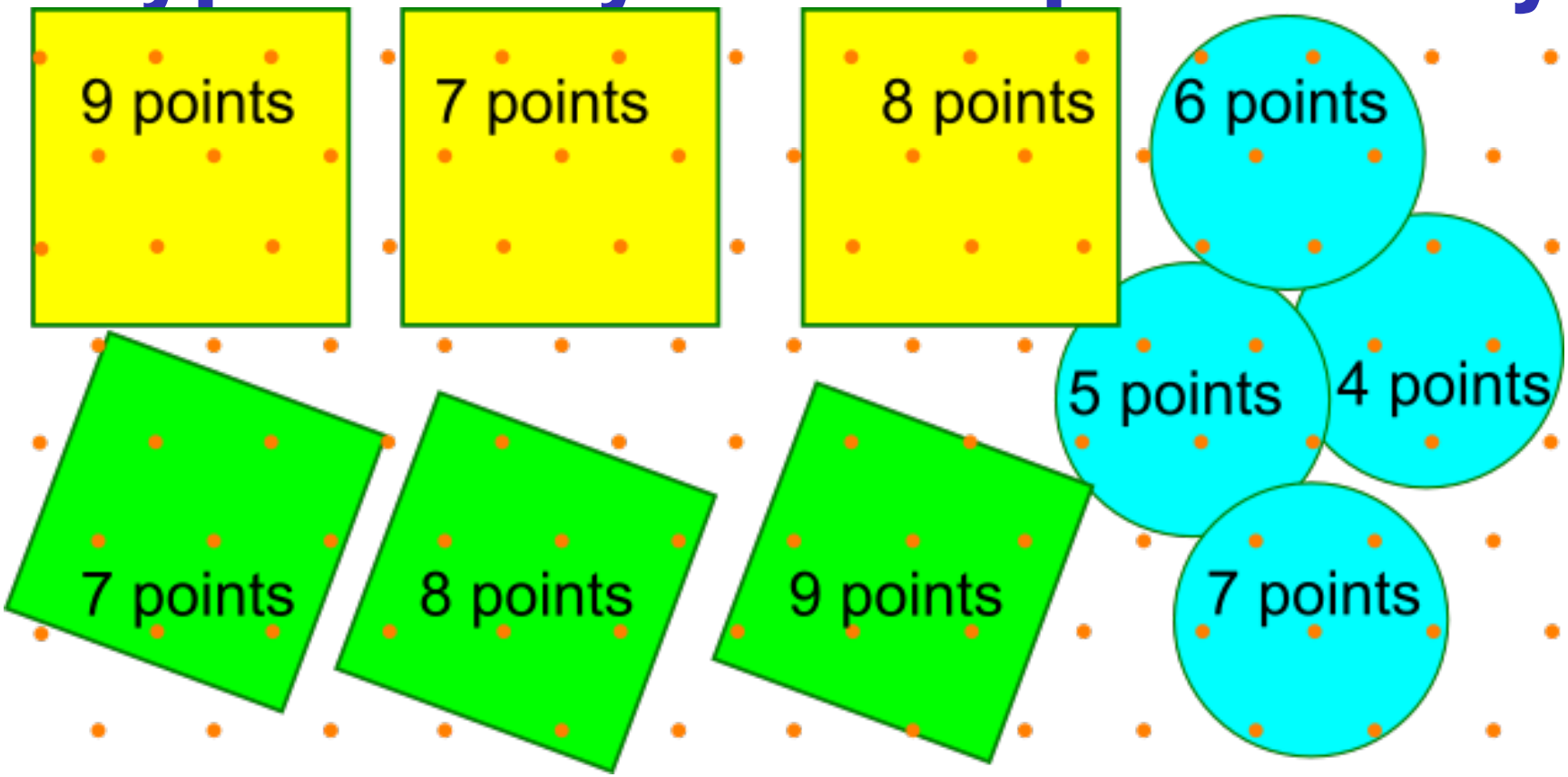
Answer: use invariants.

Non-invariants cannot help science

Even if some descriptors or features distinguish objects, it doesn't make them reliable invariants. The average colour (one of $256^3 = 16,777,216$) of clothes can easily distinguish many people but cannot be used for a reliable identification.

A scientifically justified invariant of humans is a DNA code. Data Science for any other objects looks for similar invariants that are complete for an important *equivalence relation* in question.

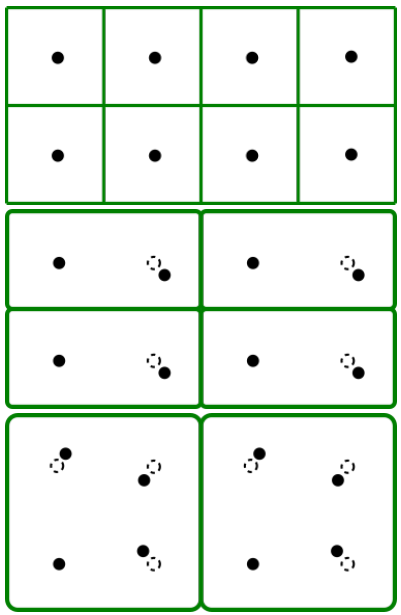
Typical way to lose periodicity



Taking boxes or balls with a *fixed cut-off radius* produces **non-isometric finite sets with no chance** to reconstruct a given periodic set.

Discontinuity of past invariants

Even if a cell is reduced (Niggli's cell), any such reduction is discontinuous under perturbations.



A reduced cell can double under almost any perturbation. All discrete invariants including symmetry groups are discontinuous. How can we continuously *quantify a crystal similarity*?

Why is *continuity important*? All atoms vibrate, real measurements are noisy, too many crystals.

Isometry classification problem

We need a complete and continuous isometry *invariant* I : $\{\text{periodic point sets}\} \to \{\text{numbers}\}$.

1) *Invariance* : if point sets $S \sim Q$ are isometric, then $I(S) = I(Q)$, so I should be well-defined on isometry classes, independent of a unit cell.

2) *Completeness* : if $I(S) = I(Q)$, then S, Q are isometric, so I distinguishes all sets $S \not\sim Q$.

3) *Continuity* : the invariant I slightly changes under perturbations to quantify a similarity.

More classification requirements

4) *Computability* : a polynomial time in a motif size (the number *m* of atoms in a unit cell).

Current brute-force : blind sampling of an infinite space produces 5679 predictions over 12 weeks on a supercomputer, five crystals synthesised.

5) *Inverse design* : a complete invariant should allow us to reconstruct a full 3D crystal so that we can choose a new invariant value and then discover a new crystal with desired properties.

Metric axioms and metric problem

A metric $d \geq 0$ on isometry classes of crystals:

(1) $d(S, Q) = 0$ if and only if S, Q are isometric;

(2) symmetry: $d(S, Q) = d(Q, S)$;

(3) $\triangle$ inequality: $d(S, T) \leq d(S, Q) + d(Q, T)$.

The first metric axiom fails for any non-complete invariant I: if $I(S) = I(Q)$ for non-isometric S, Q, then any distance d between $I(S), I(Q)$ is 0.

The metric problem solves the classification: S, Q are isometric if and only if $d(S, Q) = 0$.

Mercury's RMSD implementation

Given two crystals, Mercury tries to match a number of molecules (15 by default) in both crystals by finding a best rigid motion, outputs the Root Mean Square Deviation RMSD $= \sqrt{\frac{1}{n}\sum_{i=1}^{n} ||p_i - q_i||^2}$ between n matched atoms.

RMSD fails the triangle inequality and is a bounded version of the bottleneck distance $d_B(S, Q) = \inf_{f:S\to Q} \sup_{p\in S} ||f(p) - p||$, which can be $+\infty$, e.g. $S = \mathbb{Z}$, $Q = (1 + \varepsilon)\mathbb{Z}$ for any $\varepsilon > 0$.

New isometry invariants of crystals

Density functions, Proceedings SoCG 2021

+ continuous, + complete for generic crystals,

- slower (cubic time, hours on 5679 crystals)

Isosets, DGMM 2021, arxiv:2103.02749

+ continuous, + complete for all crystals,

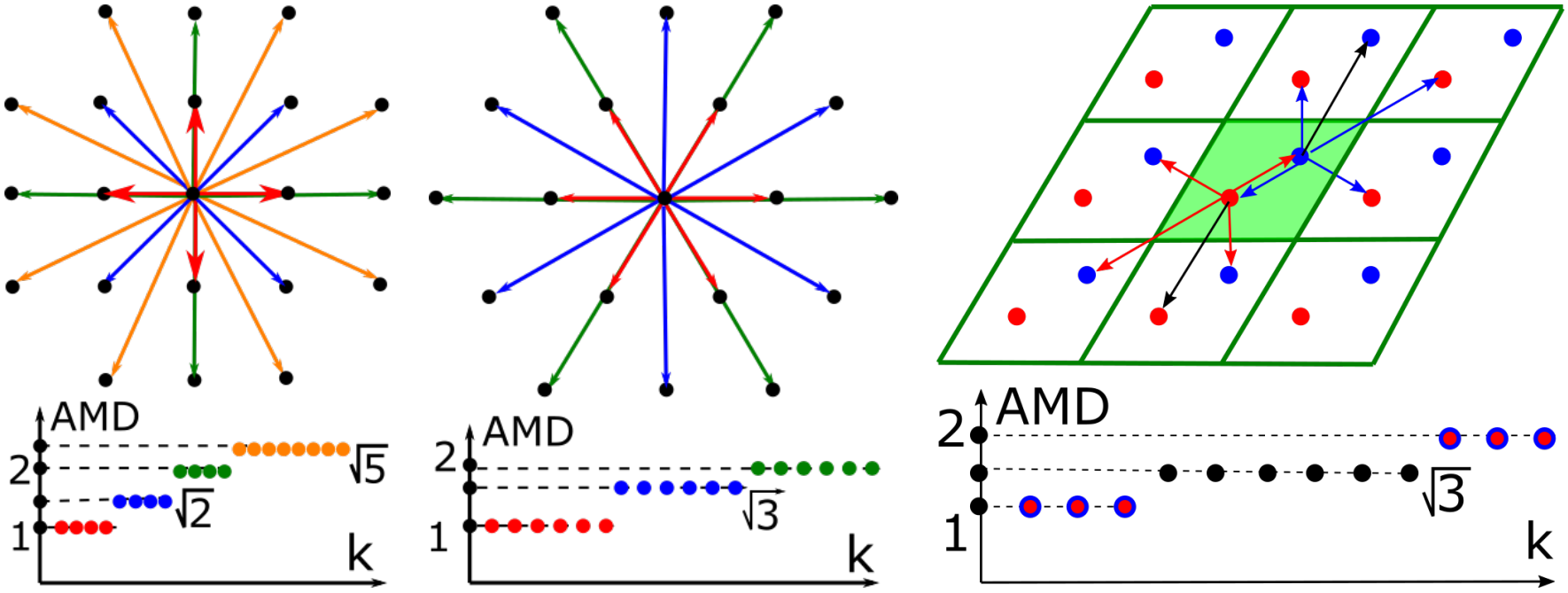
- slower (cubic time), + allow inverse design.

Distance-based invariants, MATCH, to appear

+ simple, + continuous, + fast (near linear time, seconds on 5679 crystals), arxiv:2108.04798,

+ generically complete, + allow inverse design.

Distance-based invariants



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

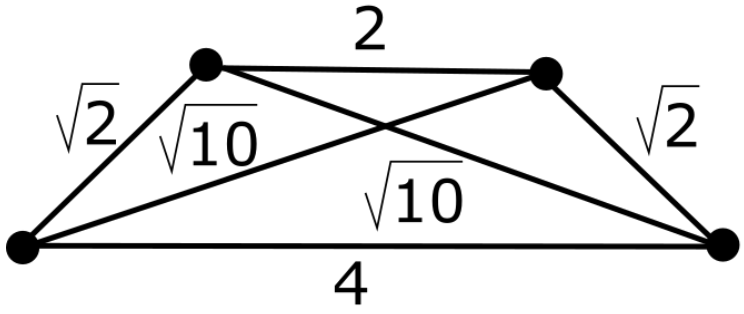
Finite sets up to isometry

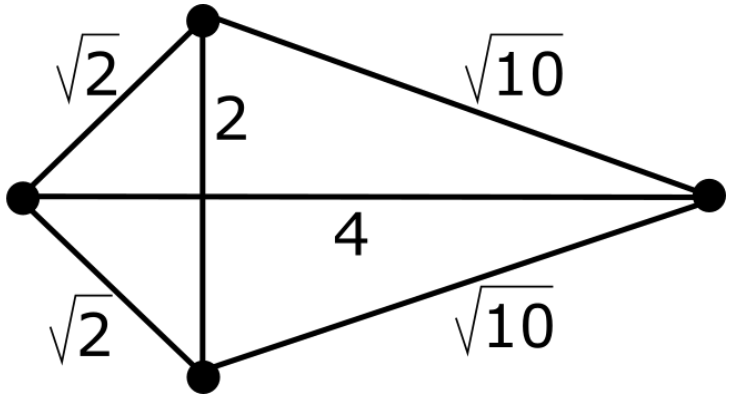
Pozdnyakov et al. *Incompleteness of atomic structure representations*, Phys. Rev. Let. 2020

reviewed many isometry invariants, also for finite sets and suggested several pairs of sets that are not distinguished by inter-point distances.

For a set S of m points and $k \geq 1$, $d_k(p)$ is the distance from p to its k-th nearest neighbor in S. The rows of the $m \times k$ matrix $D(S;k)$ are lists $d_1(p) \leq \cdots \leq d_k(p)$, ordered lexicographically.

Pointwise Distance Distributions





To get $\mathrm{PDD}(S;k)$, collapse identical rows and assign weights. The trapezium and kite differ:

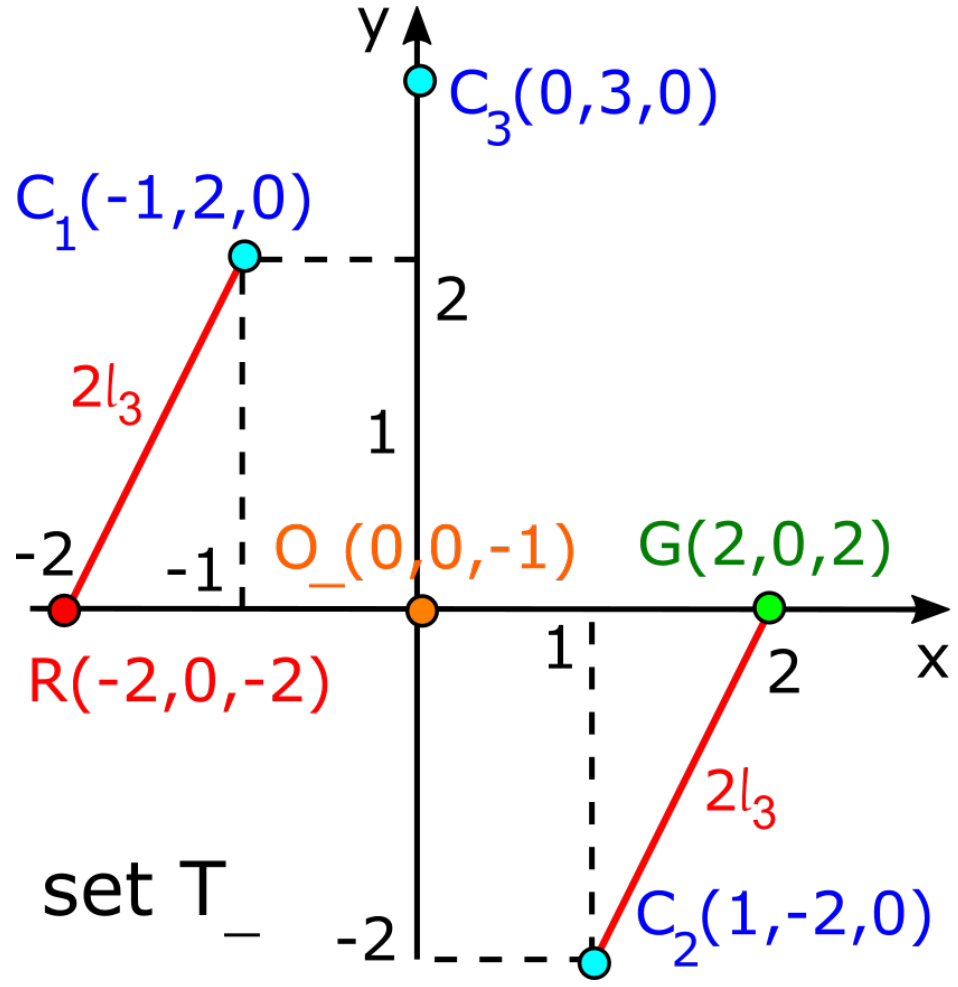
$$\mathrm{PDD}(T;3) = \left(\begin{array}{c|ccc} 1/2 & \sqrt{2} & 2 & \sqrt{10} \\ 1/2 & \sqrt{2} & \sqrt{10} & 4 \end{array}\right) \neq$$

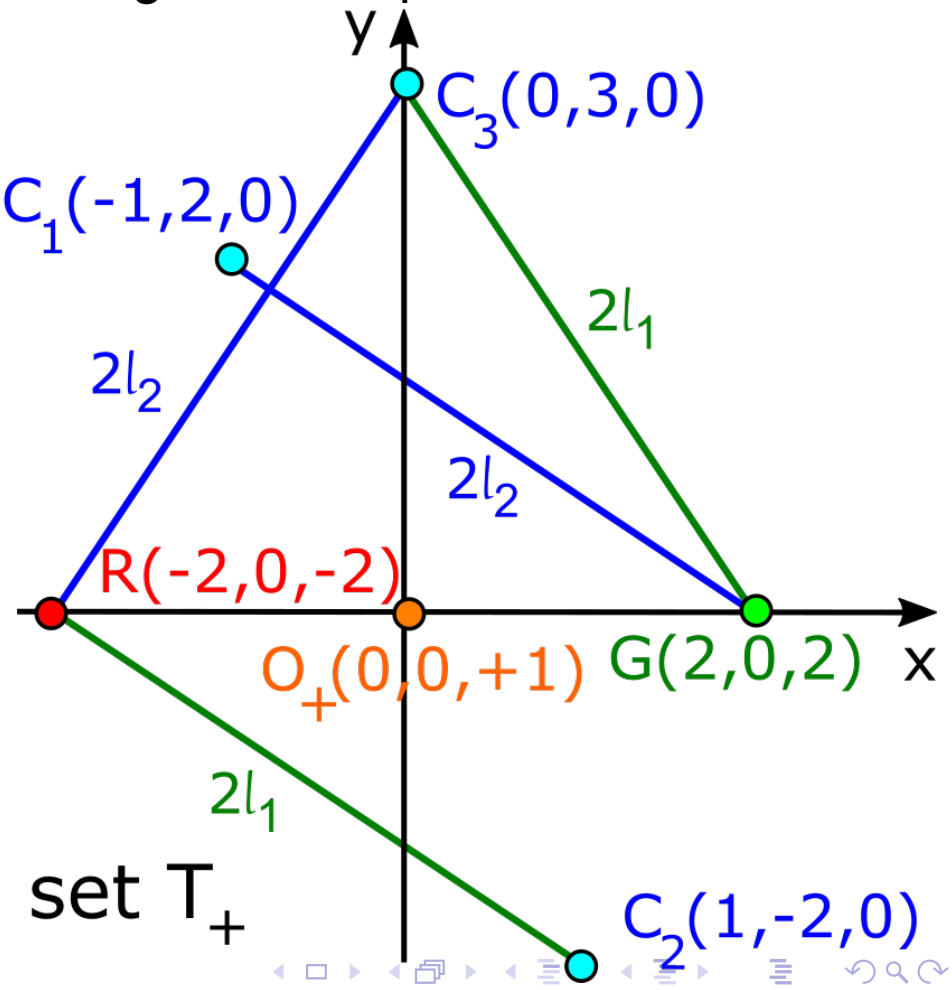
$$\mathrm{PDD}(K;3) = \left(\begin{array}{c|ccc} 1/4 & \sqrt{2} & \sqrt{2} & 4 \\ 1/2 & \sqrt{2} & 2 & \sqrt{10} \\ 1/4 & \sqrt{10} & \sqrt{10} & 4 \end{array}\right).$$

Hard-to-distinguish sets in $\mathbb{R}^3$

The 6-point sets $T_{\pm}$ with free parameters:

$RC_1 = GC_2$, $RC_2 = GC_3$, $RC_3 = GC_1$.





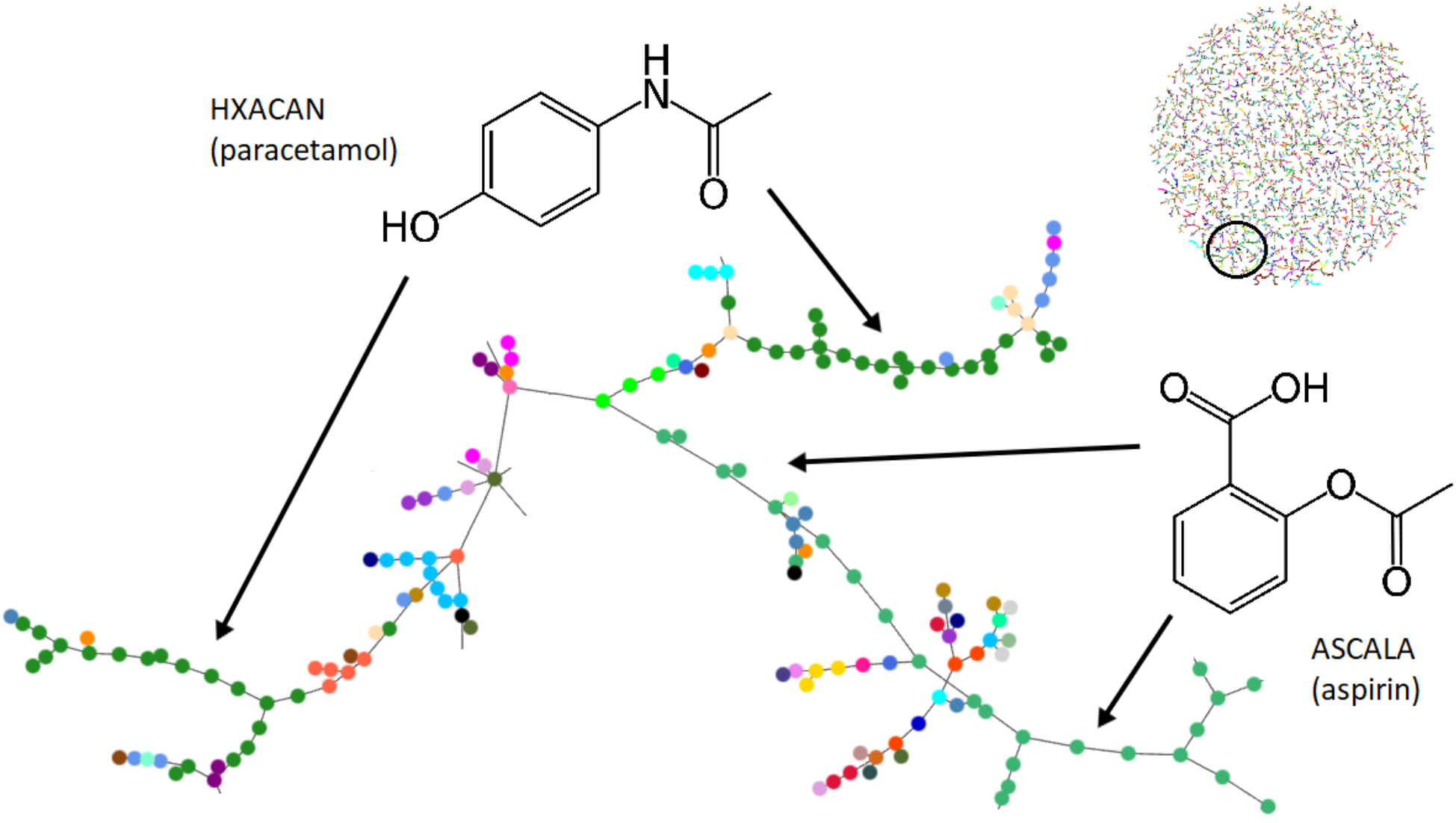
Higher-order $\mathrm{PDD}^{(h)}$ invariants

Matrices $\mathrm{PDD}^{(h)}(S;k)$ for $h \geq 2$ include ordered distances from h-point subsets of S to k nearest neighbours, distinguish all known non-isometric sets, continuous in the Earth Mover's Distance.

Generic periodic sets can be **reconstructed** from $\mathrm{PDD}(S;k)$ and lattice invariants for big k.

Based on k-nearest neighbours, $\mathrm{PDD}^{(h)}(S;k)$ is found in time $O(h^2km^h\log(hm)\log^2 k)$ with some constants depending on S, arxiv:2108.04798.

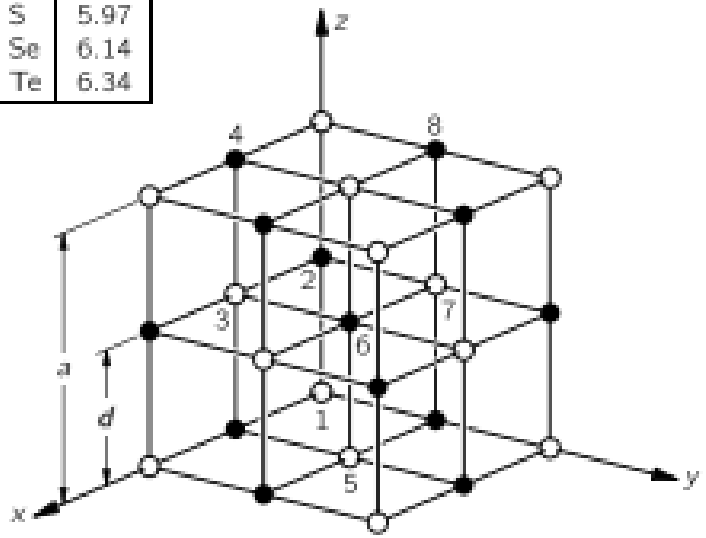
A tree of 12576 crystalline drugs



Crystal Isometry Principle

Map: {**all crystals**} $\rightarrow$ {**periodic point sets**}
taking only atomic centres should be injective.

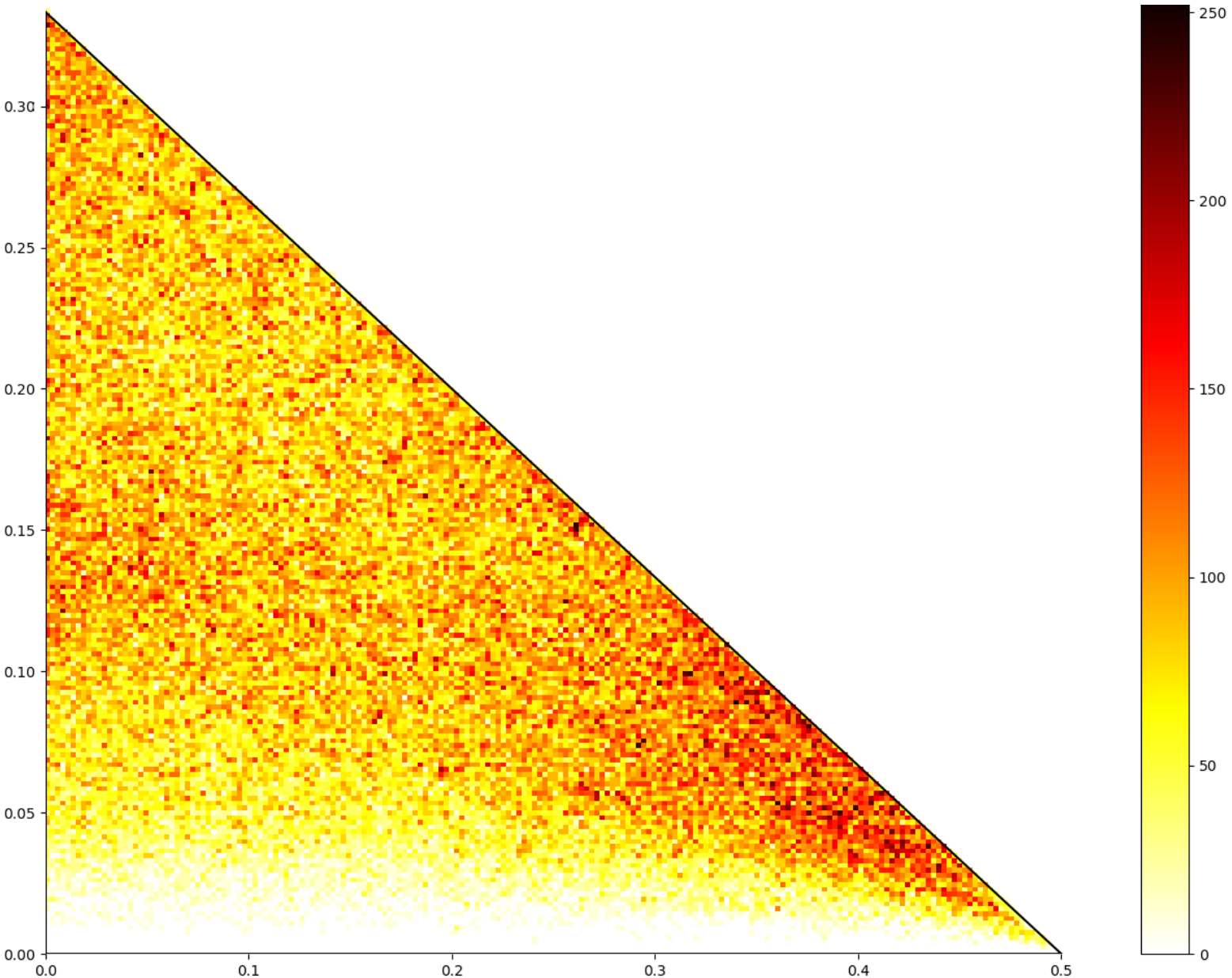
Crystal	●	○	a (Å)
Rocksalt	Na	Cl	5.64
Sylvine	K	Cl	6.28
	Ag	Cl	5.54
	Mg	O	4.20
Galena	Pb	S	5.97
	Pb	Se	6.14
	Pb	Te	6.34



Nearest neighbor distance $d = a/2$

400M+ pairwise comparisons of all 660K+ periodic crystals in the CSD detected 5 pairs with identical geometry, different chemistry, physically impossible: 5 journals are investigating.

145K+ orthorhombic lattices



Summary: maths for crystals

Equivalence of crystals: isometry, rigid motion.

Past descriptors: not invariants, discontinuous.

Isometry invariants PDD (Pointwise Distance Distributions) are complete in general position, simpler and faster than persistence that is a weaker isometry invariant of finite sets in TDA.

Crystal Isometry Principle (CRISP) justifies that all crystals live in one *continuous space* parameterised by complete invariants.