Invariants and metrics on periodic crystals

MIF++ discussion seminar with MaThCryst and other crystallographers across the world

Materials Innovation Factory (MIF), Liverpool

Summary of the last discussion

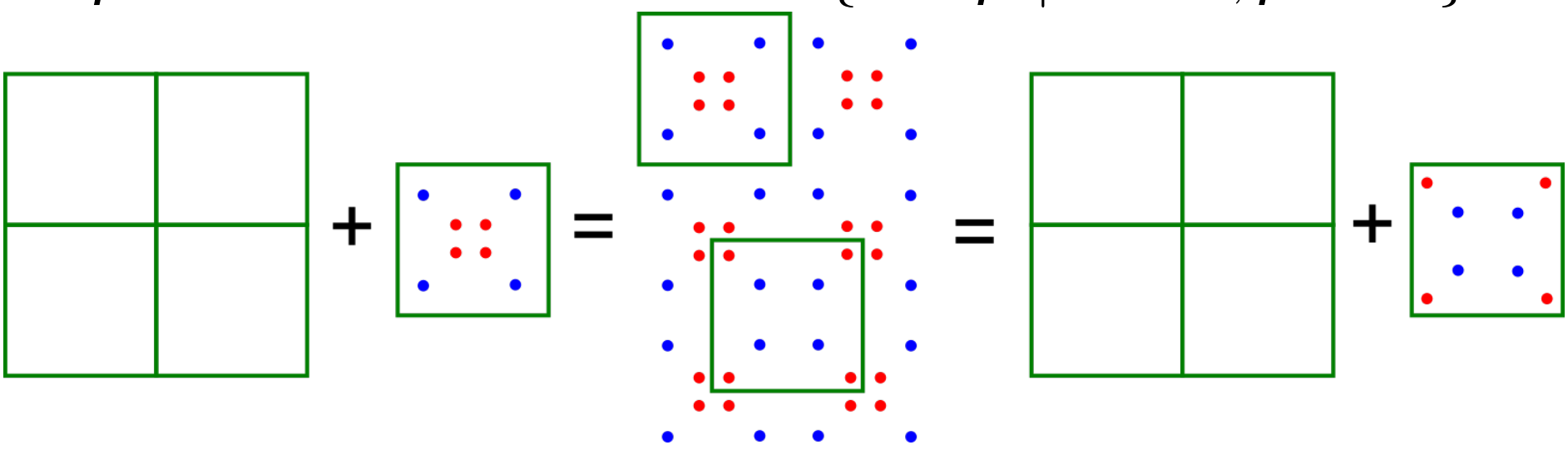
Many thanks to Mois, Davide, Larry, Berthold, Greg for sharing their expertise and references.

On 25th March we discussed that one can define many *equivalence relations* on crystals, 14 Bravais classes, 219 (or 230) space-group types (isomorphism classes) are widely known.

The **strongest equivalence** is *rigid motion* (or isometry including reflections) because crystal structures are determined in a *rigid form* and do not change their properties under rigid motion.

A periodic point set (crystal)

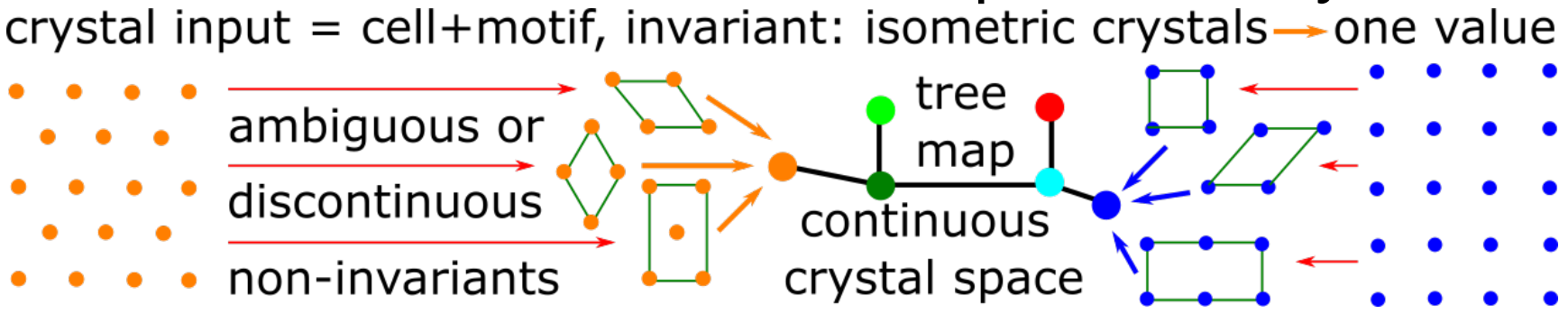
Any basis $v_1, \ldots, v_n$ of $\mathbb{R}^n$ spans the *unit cell* $U = \{\sum_{i=1}^{n} c_i v_i \,:\, 0 \le c_i < 1\}$ and generates the lattice Λ. For any finite *motif* $M \subset U$, the *periodic point set* is $S = \Lambda + M = \{v + p \mid v \in \Lambda, p \in M\}$.



Different pairs (basis, motif) give equivalent sets.

How to classify up to equivalence

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

Invariants vs non-invariants

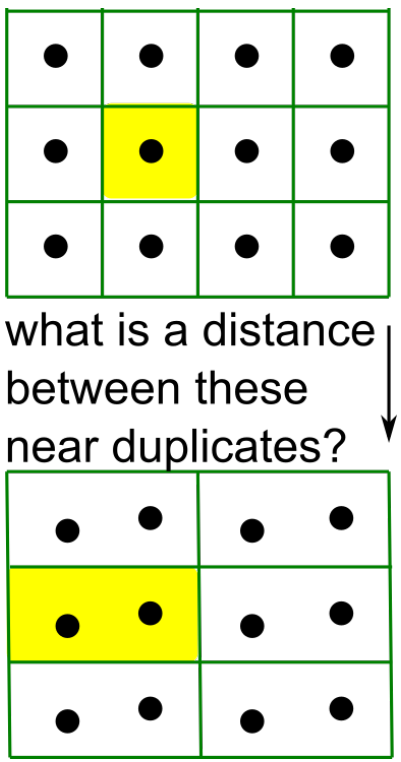
Many non-invariants are surprisingly used: atomic (fractional) coordinates change under any translation in a fixed cell, comparing infinite crystals by their finite subsets is even worse.

Isometry invariants I: chemical composition, space-group type, density. All incomplete: non-isometric $S \ncong Q$ can have $I(S) = I(Q)$.

Structure Tidy and other comparisons start from a conventional or reduced cell (space-group settings), whose shape is an isometry invariant.

Discontinuity of past invariants

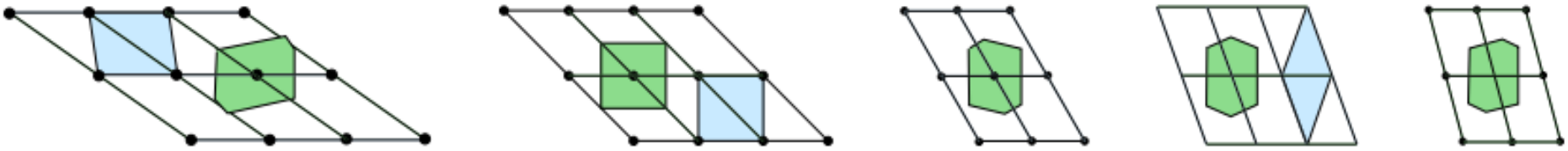
Even if a cell is reduced (Niggli's cell), any such reduction is discontinuous under perturbations.



A reduced cell can double under almost any perturbation. All discrete invariants are discontinuous. Why is *continuity important*? All atoms vibrate, real measurements are noisy, too many crystals. How can we continuously *quantify a crystal similarity*?

JAC 49, 653–664: "difficult to suppose strict boundaries in similarity by universal criteria".

What if we allow perturbations?



Assume $A \sim B$ if A can be perturbed to B by shifting coordinates up to a threshold $\varepsilon > 0$.

Then any objects A, B are joined by a chain of small perturbations $A \sim A_1 \sim \cdots \sim A_n \sim B$, so all A, B are equivalent by the transitivity axiom.

If we fix any threshold $\varepsilon > 0$, we can cluster a finite dataset. The resulting classes can change if we add objects. Is anything better possible?

Metric axioms and metric problem

A metric d is a distance on pairs of objects:

(1) $d(S, Q) = 0$ if and only if $S \sim Q$ equivalent;

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

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

Positivity $d(S, Q) \geq 0$ follows from (1), (2), (3).

The triangle axiom justifies a shortest path on a finite set: a distance $d(S, Q)$ cannot be larger than the length of a path from S to Q via any T, implicitly assumed by many clustering tools.

Importance of the first axiom

Past attempts to define a metric on crystals or other objects with non-unique representations:

for a descriptor v (usually a vector, good if v is an invariant), take the Euclidean (or any other) distance on these vectors: $||v(S) - v(Q)||$.

The result is a metric only if $v(S)$ is a *complete invariant*, else there are non-equivalent $S \not\sim Q$ with $v(S) = v(Q)$ and distance 0. Finding a metric involves solving the equivalence problem.

Equivalence vs metric problems

If we can solve the equivalence problem, one can define the discrete metric $d(S, Q) = 1$ for all non-equivalent $S \not\sim Q$, which is discontinuous for slightly different non-isometric crystals.

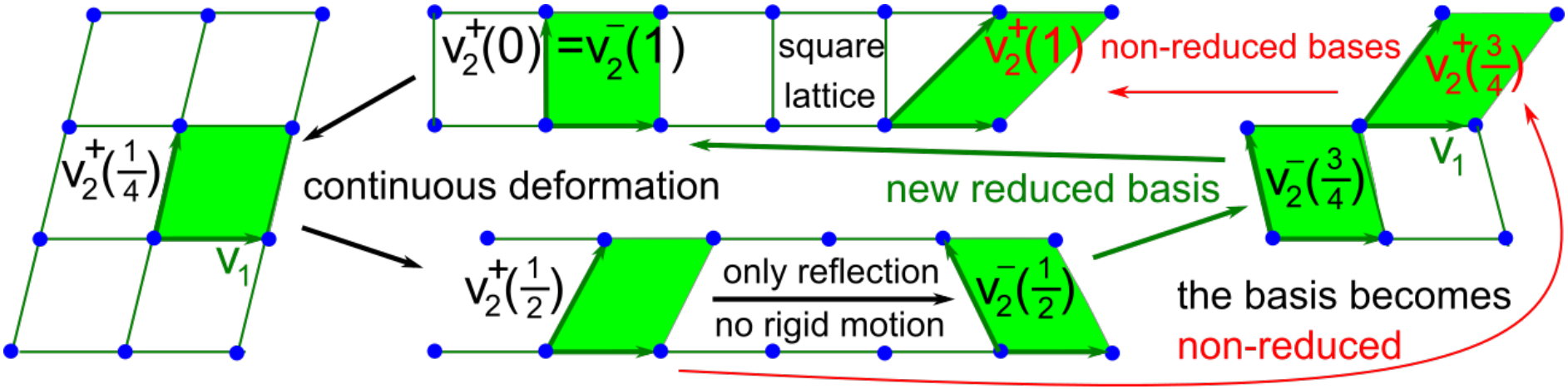
Converse: any metric d detects equivalence $S \sim Q$ if and only if $d(S, Q) = 0$ (first axiom).

A continuous metric $d(S, Q)$ will be better. If Q is obtained by perturbing all atoms of S up to ε, then $d(S, Q) \leq C\varepsilon$ for a constant C and all S, Q.

Cell reduction is discontinuous

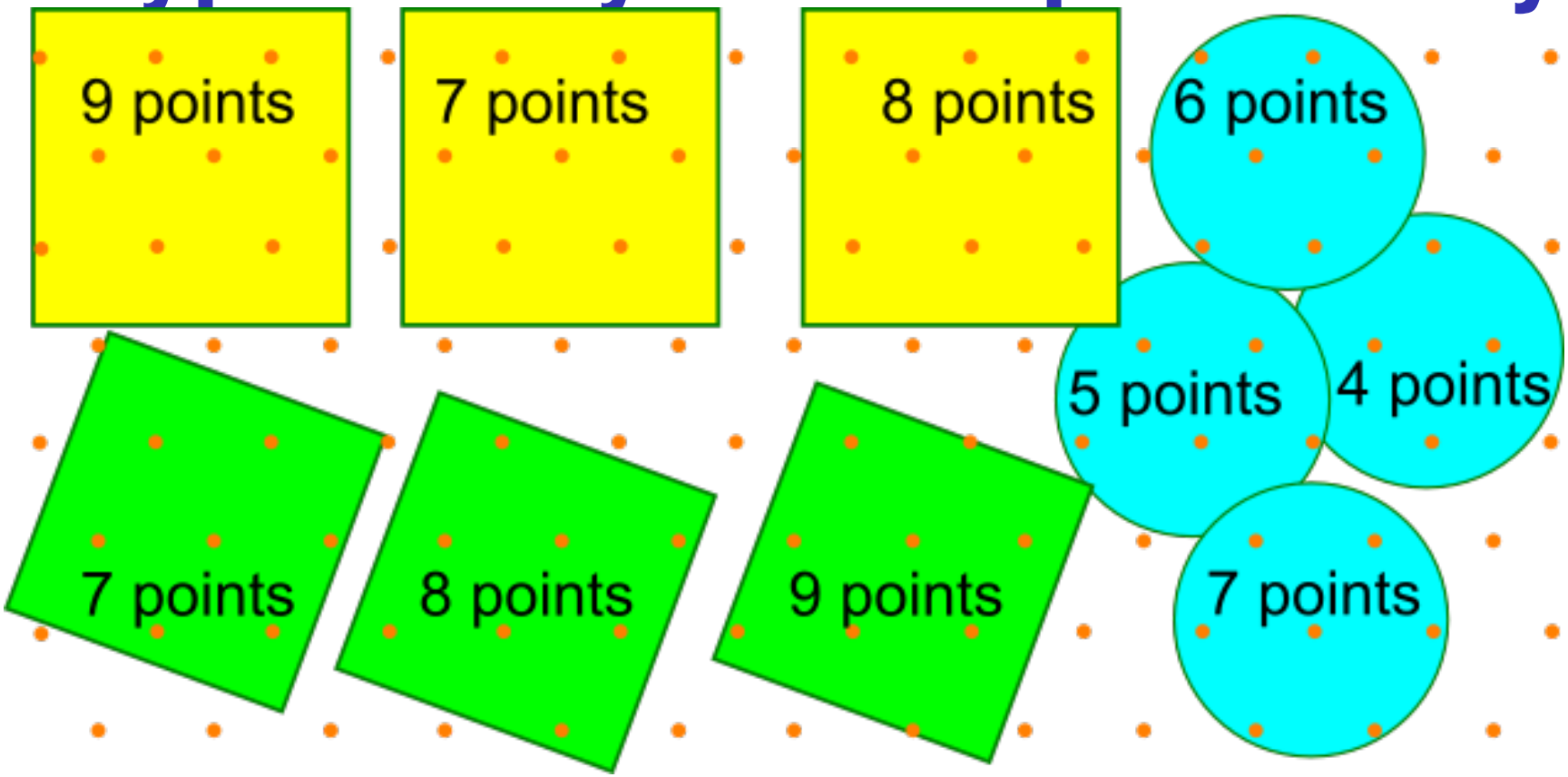
Deform the bases $v_1(t) = (1, 0)$, $v_2^+(t) = (t, 1)$.

We go back to the square lattice over $t \in [0, 1]$.



Even for lattices, any reduced cell (not only Niggli's) is discontinuous: close initial bases can have distant reduced bases, see Theorem 15, MATCH Comm. Math. Comp. Chem., 87(3), 529-559.

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.

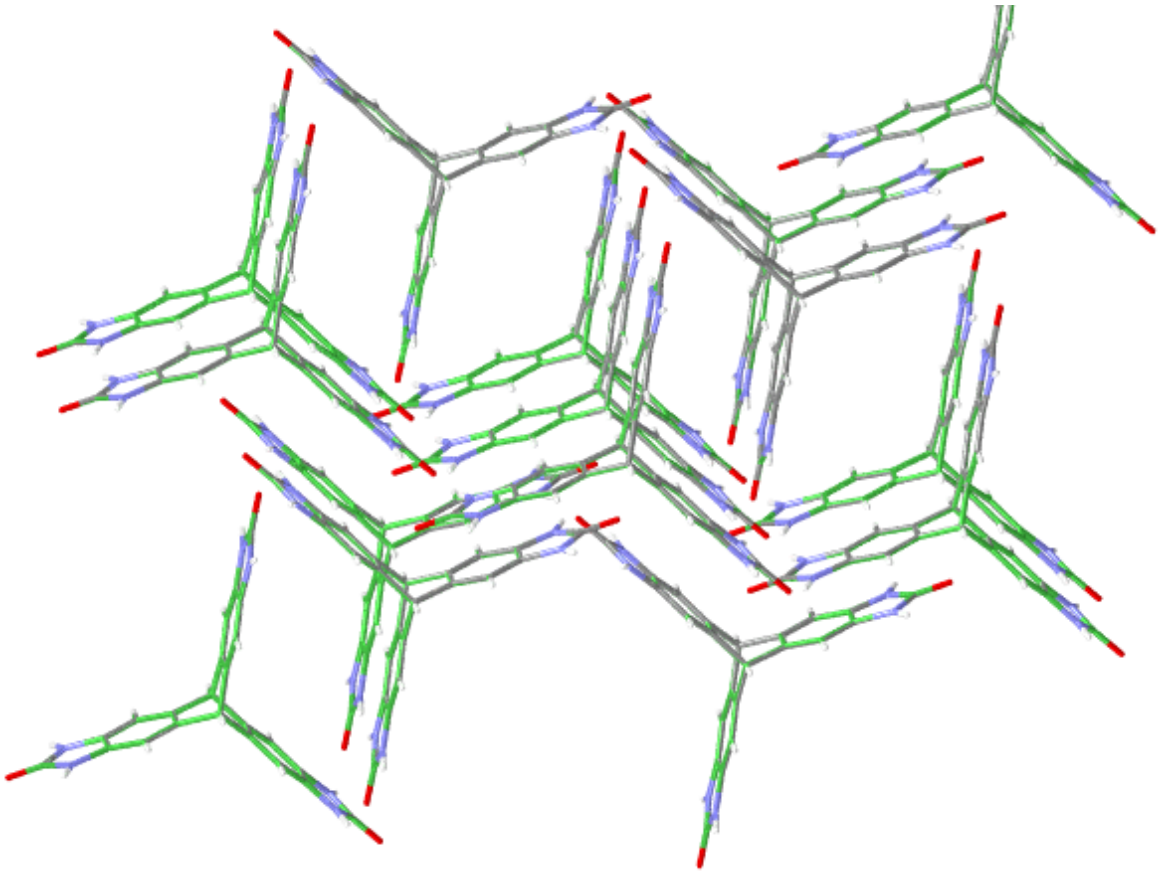
Mercury's RMSD implementation

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 $\varepsilon > 0$.

A partial match of molecules



Crystals 14 and 15 are overlaid: $RMSD = 0.004$Å, but this number irregularly grows if we match more molecules in crystal 14 and T2-δ.

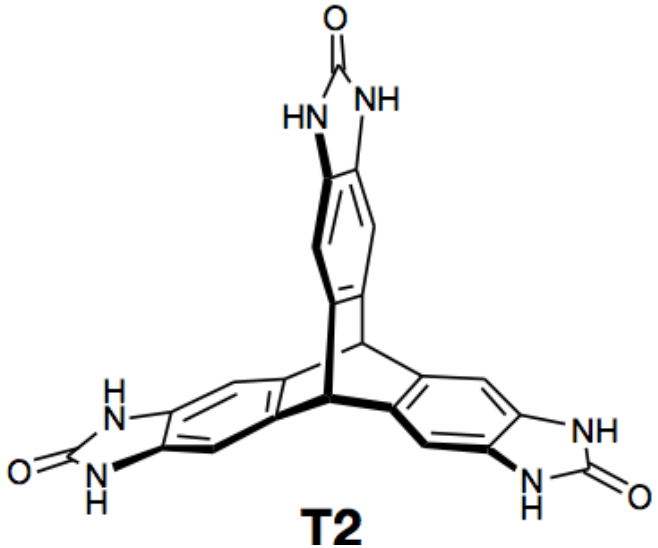
# matched molecules	5 of 5	8 of 10	10 of 15	11 of 20	16 of 25	18 of 30	21 of 35
RMSD, 1Å $= 10^{-10}$m	0.603	0.681	0.812	0.825	0.99	1.027	1.079
running time, seconds	0.168	0.422	2.026	14.61	63.51	151.4	759.3

CSP: Crystal Structure Prediction

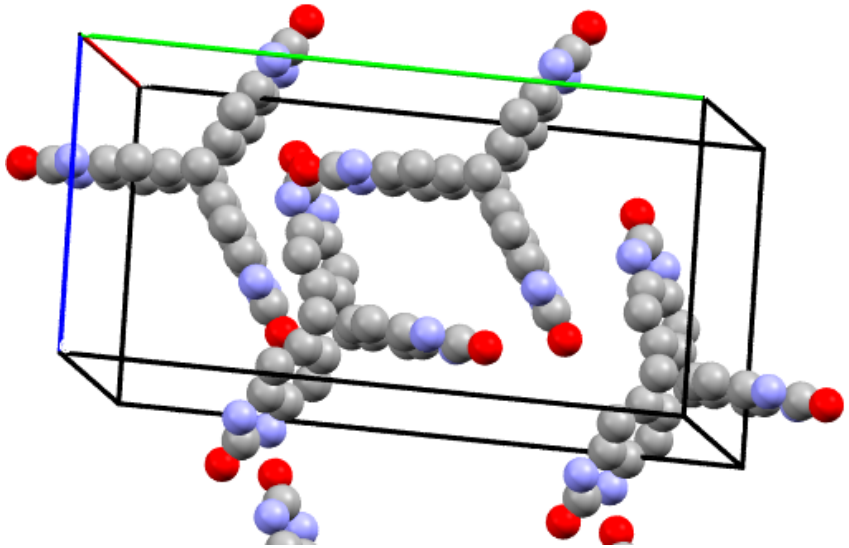
Aim : discover crystals with target properties,

predict *stable* crystals by ‘randomly’ placing several molecules or atoms into a ‘random’ cell.

The lattice energy is iteratively minimized to get most stable crystals that can be synthesized.

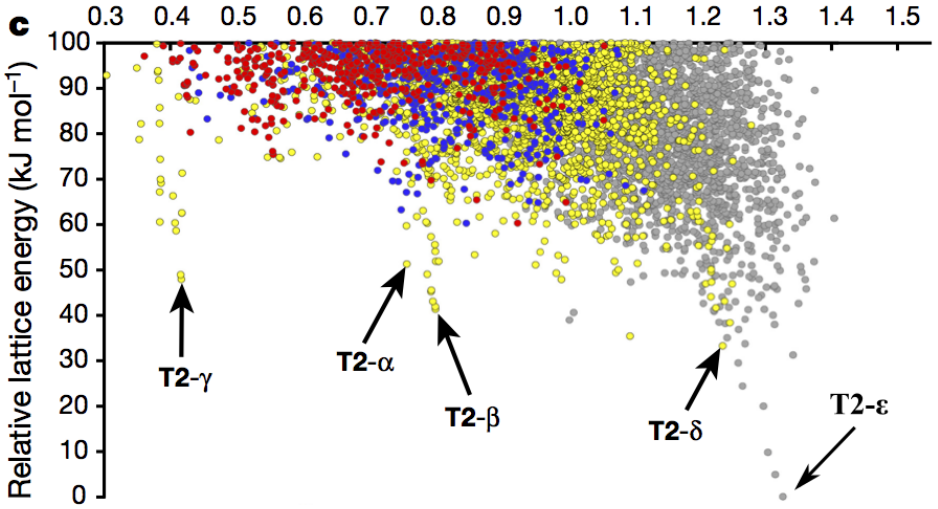


T2



Embarrassment of over-prediction

coined by Prof Sally Price FRS (UCL) says that the state-of-the-art CSP software outputs too many *approximate local minima* of the energy: random start, perturb atoms, compute again, ...



The plot in Nature 2017 shows 5679 predicted crystals of T2 molecules, only 5 were synthesised.

Each crystal is represented by (density, energy), insufficient to completely map a crystal space.

CSP needs a continuous metric

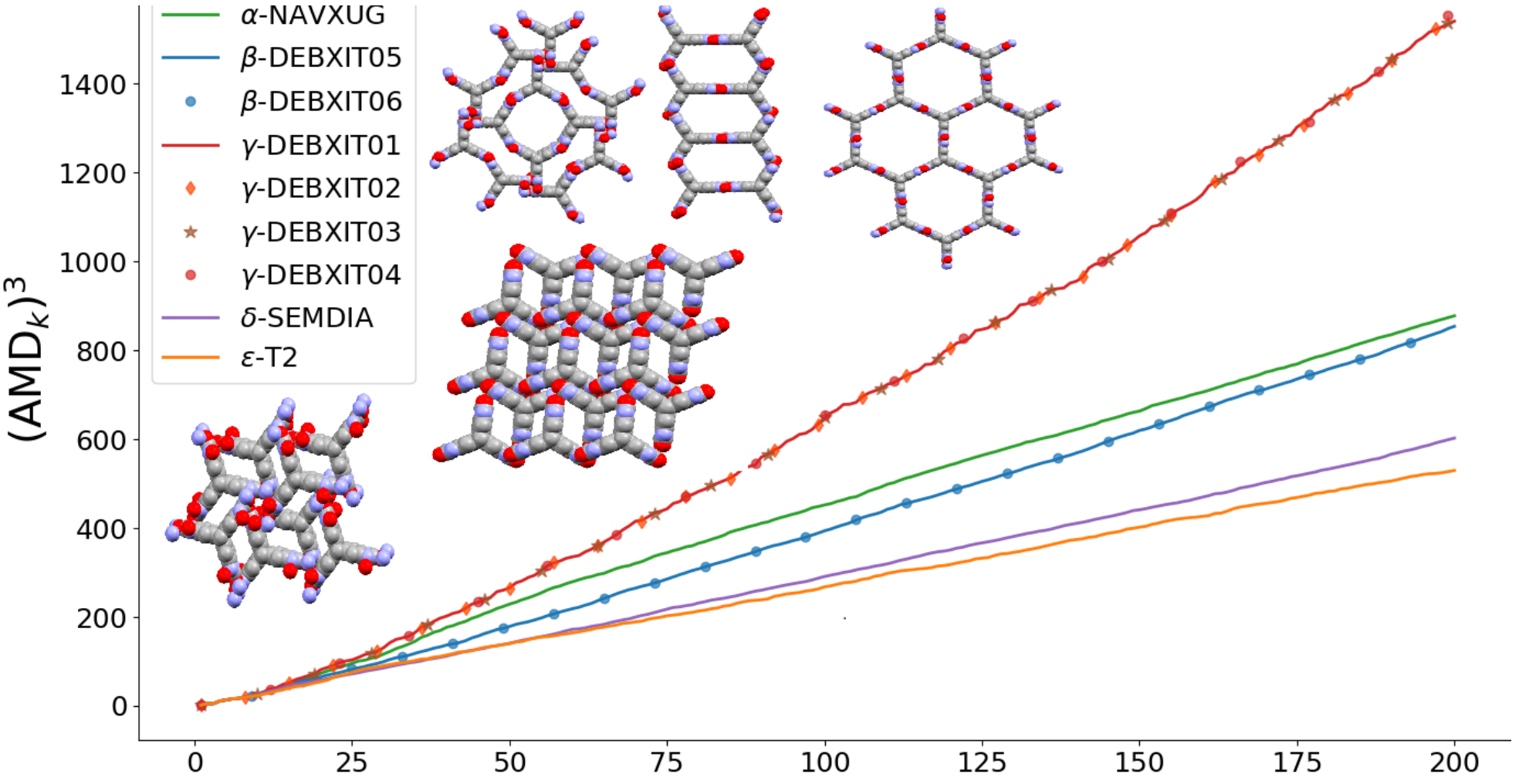
Any optimisation of simulated crystals outputs slightly distinct approximations of the same local energy minimum. Repeated runs produce many nearly identical crystals whose similarity is hard to spot because of different cells, space groups.

cell_length_a	7.3760
cell_length_b	12.3050
cell_length_c	13.8278
cell_angle_alpha	95.0406
cell_angle_beta	74.5727
cell_angle_gamma	72.5670

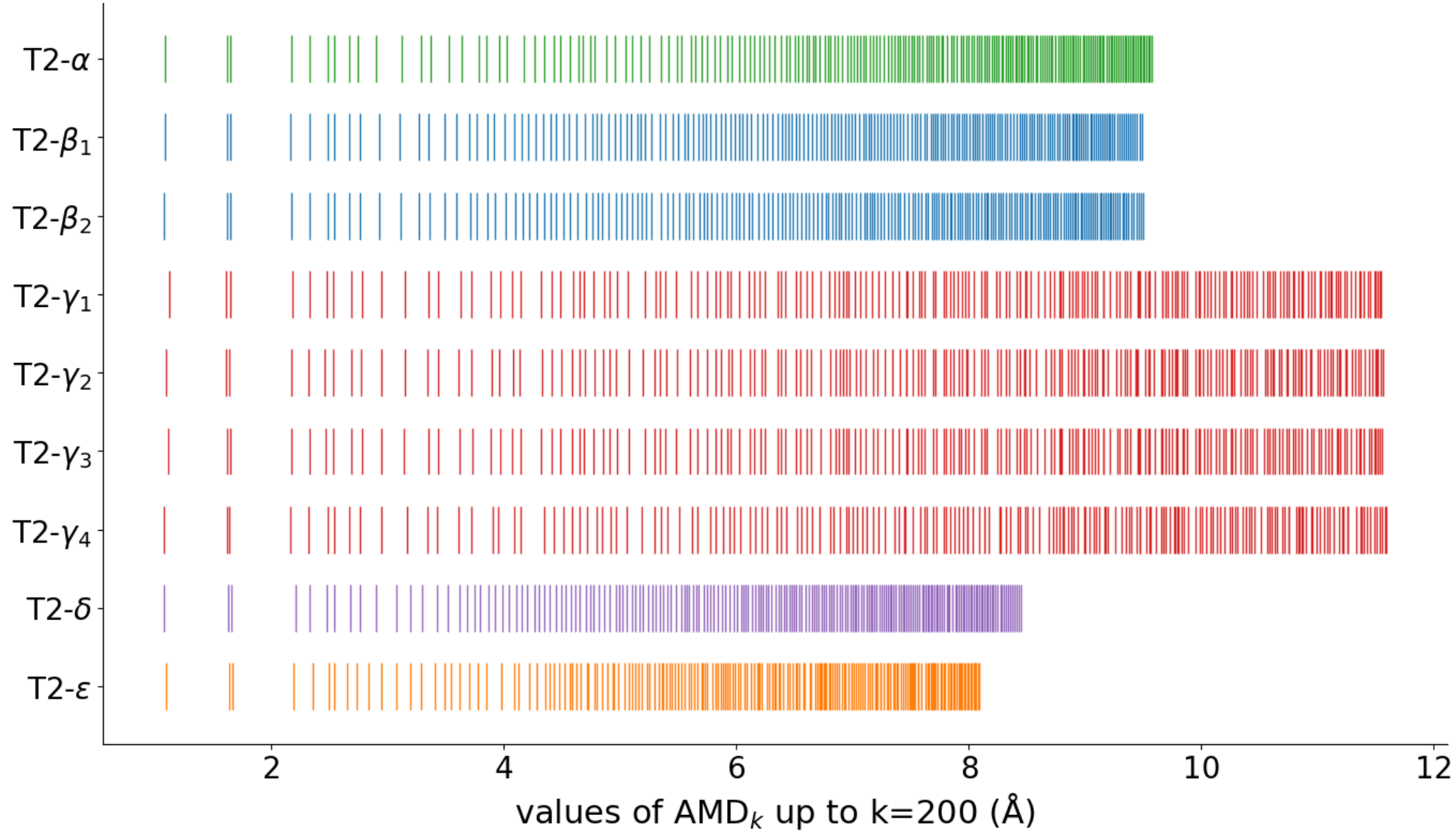
cell_length_a	26.6601
cell_length_b	7.3750
cell_length_c	32.1616
cell_angle_alpha	90.0000
cell_angle_beta	134.1164
cell_angle_gamma	90.0000

Crystals 14 and 15 have 2 and 8 T2 molecules in the unit cells above, but are nearly identical.

Near-duplicates in exp. databases



AMDigrams of exp. T2 crystals



The mapping problem for crystals

Find a map I on the space of isometry classes (CRISP) of periodic point sets such that

(a) *invariant* : S, Q are isometric $\Rightarrow I(S) = I(Q)$;

(b) *complete*: $I(S) = I(Q) \Rightarrow S, Q$ are isometric;

(c) *metric* : I allows us to define *continuous* d

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

(2) symmetry $d(S, Q) = d(Q, S)$ and

(3) $\triangle$ inequality $d(S, Q) + d(Q, T) \geq d(S, T)$;

More conditions for a good map

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

660K+ periodic crystals in the CSD require 200B+ pairwise comparisons (now done :-).

5) *Inverse design* : a complete invariant should allow us to reconstruct a 3D crystal so that we can choose a new invariant value (unexplored place on a geographic map) for a realizable periodic point set and discover a new crystal.

From discrete to continuous

Crystallography has continuously evolved from the 19th century to the 21st. Biological analogy:

space-group types $\leftrightarrow$ Linnaean taxonomy,

group-sub(super)group relations $\leftrightarrow$
Darwin's evolution theory of species,

continuous complete invariants of crystals $\leftrightarrow$
DNA-style crystallography (materials genome).

We are open to collaboration. Come to the ECM33 satellite MACSMIN on 5-9 September