Introduction to Geometric Data Science tutorial at SIAM 2024 Math of Data Science

Vitaliy Kurlin's Data Science group :
Olga Anosova, Yury Elkin, Dan Widdowson
Materials Innovation Factory, Liverpool, UK

http://kurlin.org/projects/Geometric-Data-Science.pdf

What to expect in the tutorial

Practical *motivations*, *problem statements*, a few answers leading to *new principles* for molecules and materials, and many more *open questions*.

First 10 min: what *real data* objects do we consider? What data objects are the **same**?

Next 50 min: the **finite** case (*point clouds*) with applications to *proteins* and other *molecules*.

5-min break after the end of the first hour.

Final 50 min: the **periodic** case for *materials*.

What is Geometric *Data* Science?

Data can mean many objects: lattices, periodic crystals, *point clouds*, molecules, protein chains

These data objects are often given via digital representations that can be *highly ambiguous*:

different representations may refer to the *same* object but what do we really mean by **“same”?**

Sacchi et al. **Same or different - that is the question**: identification of crystal forms from crystal structure data. CrystEngComm, 2020.

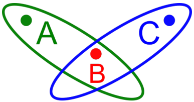
Three axioms of an equivalence

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: $0.5 = 50\% = \frac{1}{2} = 2:4$

Different equivalence relations

Chemical : compounds $A \sim B$ if A, B have the same composition. Ok, but diamond, graphite of pure carbon have vastly different properties.

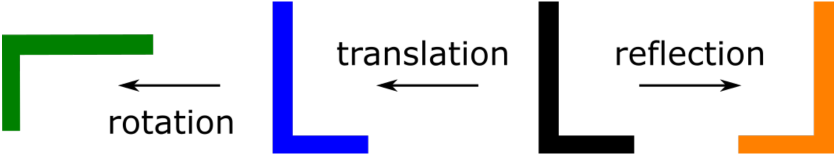
By property : molecules $A \sim B$ if A, B have the same property. Ok, but molecules that share one property can differ by other properties.

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, $Fm\bar{3}m$).

What is the strongest relation?

Many real-life objects are rigid and should be considered equivalent under **rigid motion**

= a composition of translations and rotations;



or *isometry* = rigid motion + reflections in $\mathbb{R}^n$.

In a general metric space, an **isometry** is any map that preserves all inter-point distances.

Point clouds are universal inputs

If all m points of a cloud $C \subset \mathbb{R}^n$ are **ordered** $p_1, \dots, p_m$, then C is *reconstructed* (uniquely up to isometry) from the distances $d_{ij} = |p_i - p_j|$,

which are also Lipschitz **continuous** under perturbations: perturbing any point p_i up to ε changes any distance d_{ij} only up to 2ε.

Point clouds are universal inputs

If all m points of a cloud $C \subset \mathbb{R}^n$ are **ordered** $p_1, \dots, p_m$, then C is *reconstructed* (uniquely up to isometry) from the distances $d_{ij} = |p_i - p_j|$,

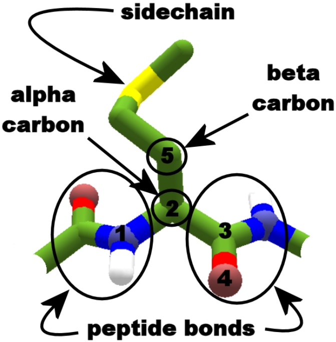
which are also Lipschitz **continuous** under perturbations: perturbing any point p_i up to ε changes any distance d_{ij} only up to 2ε.

In practice, many clouds are **unordered**.

The brute-force way to compare clouds of m unordered points by $m!$ distance matrices is *unrealistic* because of the exponential time.

The backbone of a protein chain

Any *protein chain* has a primary structure: a sequence of 20 amino acid residues whose side chains R_i are joined to alpha-carbon atoms A_i.



A *protein backbone* is a sequence of ordered triplets of the atoms (1) nitrogen N_i, (2) alpha-carbon A_i, (3) carbonyl carbon C_i, where $i = 1, \ldots, m$ (# residues).

Weaker vs stronger equivalences

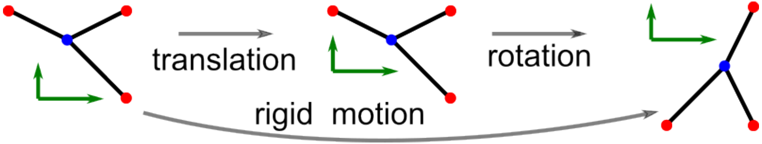
Easy equivalence: backbones are the same if their lengths (number m of residues) are equal.

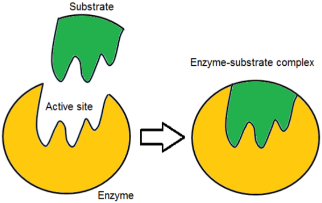
This equivalence by length is *weaker* than **by sequence of amino acids** because many different sequences have the same length.

If backbones $S, Q \subset \mathbb{R}^3$ coincide as ordered sets of atoms, they should (?) have the same sequence. The converse fails for different backbones that have the same sequence.

What protein backbones can be called *different*?

What is the strongest equivalence?



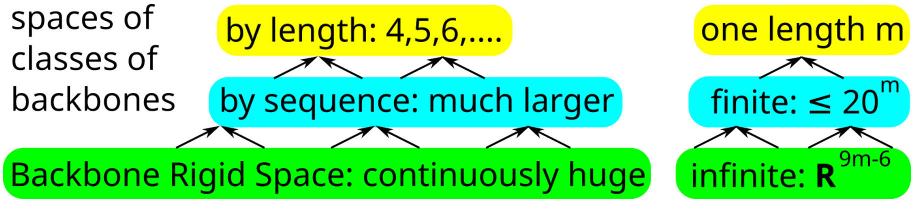


Any *rigid motion* preserves the shape of a backbone, but more flexible deformations can change protein properties.

So *rigid motion* is the **strongest** equivalence $S \cong Q$ on protein backbones S, Q in practice.

Spaces of equivalence classes

All protein backbones form a finite space (of equivalence classes) by *length*, a much larger finite space by *primary structures* (sequences), and a huge infinite *Backbone Rigid Space* of all rigid classes: noise change a backbone class.



How can we distinguish *rigid backbones*?

The importance of invariants

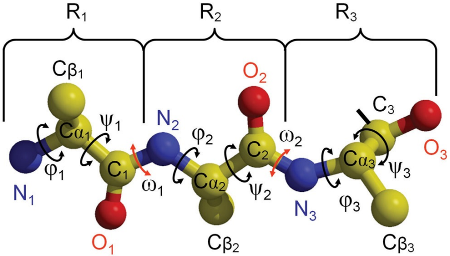
An **invariant** I is a property preserved under given equivalence (rigid motion in the sequel).

If backbones $S \cong Q$ are exactly matched by rigid motion, then $I(S) = I(Q)$. Equivalently, if $I(S) \neq I(Q)$, then $S \ncong Q$ are rigidly different.

The length m and the total sum of $A_i A_{i+1}$ distances between C_α's are rigid invariants.

But atomic coordinates are non-invariants and hence cannot reliably distinguish protein chains.

(In)complete invariants



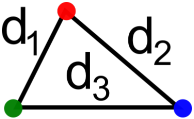
The torsion angles φ_i, ψ_i are invariants of a backbone under rigid motion.

The sequence $(\psi_1, \varphi_2, \dots, \varphi_m)$ of $2m-2$ torsion angles is often considered enough to represent a backbone, not in theory and not in the PDB.

The $9m-6$ degrees of freedom need an invariant in $\mathbb{R}^{9m-6}$. An invariant I is called **complete** if $I(S) = I(Q)$ implies that $S \cong Q$.

Invariants need a metric

A complete invariant on its own can answer only the binary question: equivalent or not? Because of noise, all real backbones differ at least slightly, which should be continuously quantified by a **distance metric** d satisfying the 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 $d_1 + d_2 \geq d_3$.

A complete invariant I gives a *discontinuous* metric: $d(I(S), I(Q)) = 0$ for $S \cong Q$, else $d = 1$.

Metrics benefit from invariants

The *Root Mean Square Deviation* (RMSD) between backbones $S, Q \subset \mathbb{R}^3$ of m atoms is

$$\mathrm{RMSD}(S, Q) = \min_f \sqrt{\frac{1}{m}\sum_{i=1}^{m} ||f(s_i) - q_i||^2}$$

minimized over all rigid motions f, where s_i, q_i are corresponding atoms of S, Q, respectively.

A metric, e.g. RMSD, gives distances between backbones. A map of classes needs complete invariants for geographic-style coordinates.

Mapping problem for backbones

Find a map $I : \{\text{backbones}\} \to \mathbb{R}^{9m-6}$ such that

(1) $S \cong Q$ in $\mathbb{R}^3$ if and only if $I(S) = I(Q)$;

(2) any $S \subset \mathbb{R}^3$ can be reconstructed from $I(S)$;

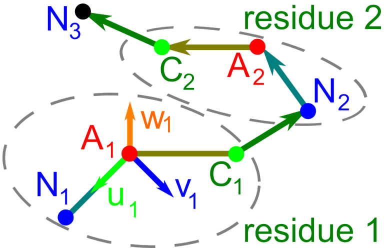
(3) there are a metric d and $\lambda > 0$ such that, for any $\varepsilon > 0$, if Q is obtained from S by perturbing every atom up to ε, then $d(I(S), I(Q)) \leq \lambda\varepsilon$;

(4) there is μ such that, for any backbones S, Q with $\delta = d(I(S), I(Q))$, all their atoms can be matched up to $\mu\delta$ by a rigid motion in $\mathbb{R}^3$;

(5) time $O(m)$ for I, d, reconstruction, alignment.

BRI = Backbone Rigid Invariant

Define the orthonormal basis of each residue: origin at A_i, normalize $\overrightarrow{A_iN_i}$ to $\vec{u}_i$, choose $\vec{v}_i$ in the plane of $\triangle N_iA_iC_i$, finally set $\vec{w}_i = u_i \times v_i$.



$\mathrm{BRI}(S)$ of m residues is the $m \times 9$ matrix whose columns include coordinates of $\overrightarrow{C_{i-1}N_i}$, $\overrightarrow{N_iA_i}$, $\overrightarrow{A_iC_i}$ in the basis $\vec{u}_{i-1}, \vec{v}_{i-1}, \vec{w}_{i-1}$.

The 1st row has 3 coordinates fixing $\triangle N_1A_1C_1$.

Theorem: BRI solves the mapping problem.

Distance metrics on invariants BRI

$\mathrm{BRI}(S)$, the metric L_∞ between invariants, a reconstruction of $S \subset \mathbb{R}^3$ from $\mathrm{BRI}(S)$ uniquely under rigid motion can be found in time $O(m)$.

The $m \times 9$ matrix BRI (with 6 zeros in row 1) flattens to a vector of $9m - 6$ coordinates in Å.

The simplest metric on vectors $L_\infty(\mathrm{BRI}, \mathrm{BRI}') = \max_i |\mathrm{BRI}_i - \mathrm{BRI}'_i|$ computes the maximum deviation of coordinates. We'll also use $\mathrm{RMS} = \sqrt{\frac{1}{9m-6} \sum_{i=1}^{9m-6} |\mathrm{BRI}_i - \mathrm{BRI}'_i|^2}$.

707K+ ‘clean’ chains in the PDB

All 220K+ entries in the PDB have 1M+ chains.

We kept 707K+ ‘clean’ chains after filtering out:

4513 non-proteins (the _entity is not ‘protein’);

178153 disordered chains (occupancy < 1);

201648 chains whose residues have non-consecutive integer indices;

9941 incomplete chains miss some main atoms

4364 chains with non-standard amino acids.

Times of pairwise comparisons

On a typical desktop, the invariants BRI for 704K+ backbones are computed in 3.5 hours.

After comparing all (883+ million) pairs of same-length backbones in 2.5 hours, we found

13403 pairs S, Q with the *exact zero-distance* $L_\infty(\mathrm{BRI}(S), \mathrm{BRI}(Q)) = 0$. The completeness of BRI implies the backbones S, Q in all these pairs can be exactly matched by rigid motion,

but S, Q may not be geometrically identical.

Exact geometric duplicates

9366 pairs turned out to have x, y, z coordinates of the main atoms N, C_α, C in all residues *identical to the last digit* without rigid motion.

763 such pairs are in *different PDB entries*.

In 9 pairs, geometric duplicates surprisingly differ by primary sequences of amino acids,

which seems physically impossible because replacing one amino acid with a different one should affect main atoms at least slightly.

Coincidences and differences

Extra symbols below are model id and chain id.

chain 1	chain 2	# identical C_α	# different acids
1a0t-0-Q	1oh2-0-Q	413	9
2hqe-0-A	2o4x-0-A	217	1, GLN$\neq$GLU

Kay Diederichs accepted the duplication of 1a0t, 1oh2: "*PDB entries can multiply on their own! ... your geometric comparison method identified an error in the PDB*." After our discussions with the PDB validation team, the CIFs of duplicates 1cov, 1gli, 1ruj, 3hhb, 4rhv were changed.

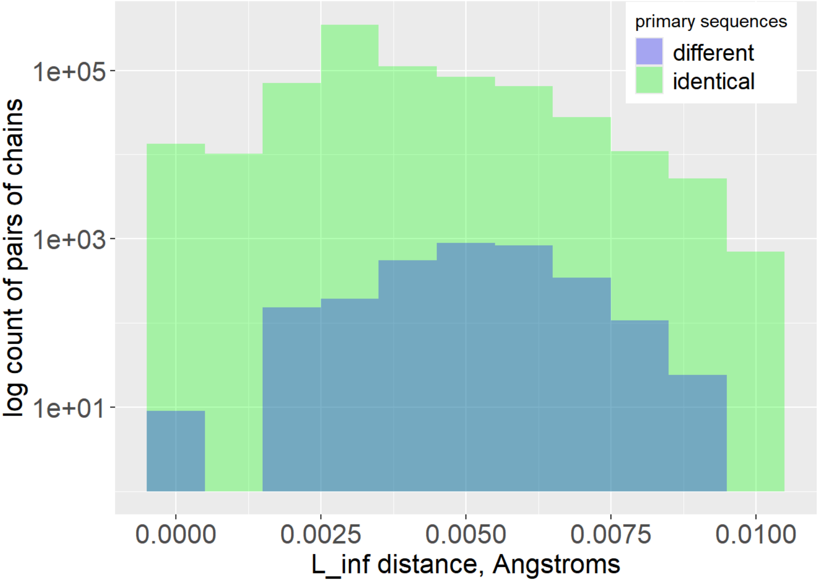
PDB duplicates keep emerging

When a new protein is deposited, it is validated individually, no comparison with all structures.

PDB chain1	PDB chain2	diff. acids	max resolution
1gli-B,D	3hhb-B,D	1/146	2.5Å
2hqe-A	2o4x-A	1/217	2Å
3msg-A	3mua-A	1/330	1.5Å
7u16-A	7u18-A	1/241	2.7Å
7u16-B	7u18-B	1/135	2.7Å
8fdz-A	8fe0-A	1/200	2.5Å

See details in Anosova et al, arxiv:2410.08203.

Many more near-duplicates



Back to clouds of unordered points

Classes of point clouds under isometry (or another equivalence) can be distinguished by

an *invariant* $I : \{\text{objects}\} \to$ simpler space such that if $A \sim B$ then $I(A) = I(B)$ or, equivalently,

if $I(A) \neq I(B)$ then $A \not\sim B$ meaning that I has

no false negatives : different representations $A \simeq B$ of the same object with $I(A) \neq I(B)$.

The size of a cloud is an isometry invariant, the center of mass is not invariant under translation.

Equivariance vs invariance

Let a group, say $G = E(n)$, act on point clouds.

A function $h : \{\text{all clouds}\} \to$ a simpler space is G-**equivariant** if $h(g(C)) = T_g(h(C))$, where T_g is a map depending on g, e.g. T_g is g acting on the mid-point $h(C)$ between closest neighbors.

The *stronger* (restrictive) concept is **invariance** when T_g is the identity. An isometry **invariant** I must satisfy $I(A) = I(B)$ for any isometric clouds $A \simeq B$. Equivalently, $I(A) \neq I(B) \Rightarrow A \not\simeq B$, so I has *no false negatives*, can *distinguish* $A \not\simeq B$.

Do invariants suffice? Yes!

Invariants distinguish objects under equivalence (by definition), while all non-invariants don't!

non-invariant descriptors

non-invariant **equivariants**

$\neq$

non-constant isometry **invariants** of discrete point sets

generically **complete, polynomial-time** computable

continuous under noise **reconstructable**

Equivariants are used to predict forces (vectors at points) that move one cloud to another cloud.

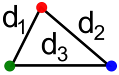
Any such sequence of (rigid classes of) clouds $C_t \subset \mathbb{R}^n$ depending on a time t can be studied in terms of *only invariants* $I(C_t)$ without vectors.

Isometry problem for real data

Find an *invariant* I : {isometry classes of data} $\to$ a simpler space satisfying these conditions:

Completeness: any objects A, B are isometric if and only if $I(A) = I(B)$, so I is a *DNA-style code* with *no false negatives* and *no false positives*.

Lipschitz continuity : there is a *metric* d :



$$d(I(A), I(B)) = 0 \Leftrightarrow A, B \text{ are isometric},$$
$$d(I(A), I(B)) = d(I(B), I(A)),\ d_1 + d_2 \geq d_3$$

and a constant λ : if B is obtained by perturbing each point of A up to ε, then $d(I(A), I(B)) \leq \lambda\varepsilon$.

Time and geography matter!

The conditions above allow a simple solution.

Complete : $I(A) = \{$all isometric images of $A\}$,

$m!$ distance matrices for one m-point cloud, a decomposition in infinitely many basis functions

Time and geography matter!

The conditions above allow a simple solution.

Complete : $I(A) = \{$all isometric images of $A\}$, $m!$ distance matrices for one m-point cloud, a decomposition in infinitely many basis functions

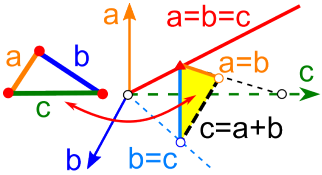
A practical invariant needs *harder* conditions.

Computability: the invariant I and the metric d should be computable in *polynomial time* in the number m of points for a fixed dimension n.

Geo-style maps: describe all *realizable values* $I(A)$ that allow us to reconstruct an object A.

Euclid's ideal solution for triangles

SSS theorem for $m = 3$ points in any $\mathbb{R}^n$. Two triangles are congruent (isometric) *if and only if* they have the same triple of sides a, b, c (up to all 6 permutations). For rigid motion (without reflections), allow only 3 cyclic permutations.



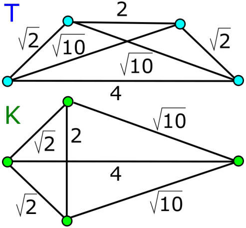
The **Cloud Isometry Space** CIS$(\mathbb{R}^n; 3)$ is the cone in $\mathbb{R}^3$ $\{0 < a \le b \le c \le a + b\}$ bounded by three planes.

Generically complete invariants

Is the problem *open for quadrilaterals* in $\mathbb{R}^2$?

One can train neural networks to experimentally output isometry invariants but it can be hard to prove completeness and continuity under noise.

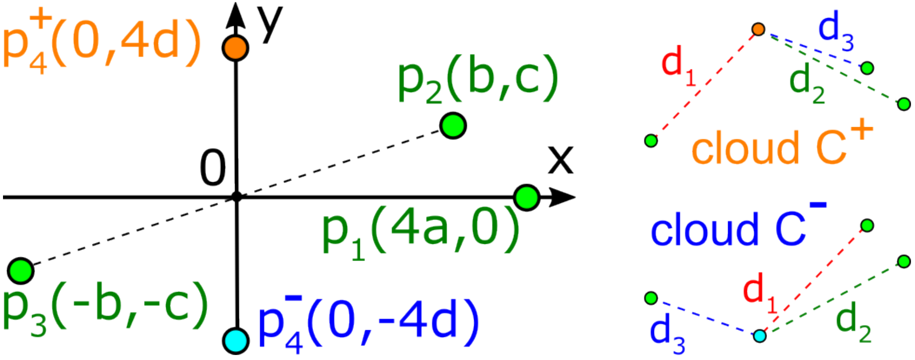
Boutin, Kemper, 2004: the vector of all sorted pairwise distances is *generically complete* in $\mathbb{R}^n$



distinguishing almost all clouds of unordered points except singular examples. These non-isometric clouds have the same 6 pairwise distances.

Pairs of singular quadrilaterals

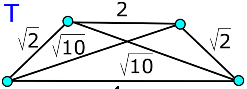
The 5-dimensional space of 4-point clouds has non-isometric $\{p_1, p_2, p_3, p_4^{\pm}\}$ with the same 6 pairwise distances depending on 4 parameters.



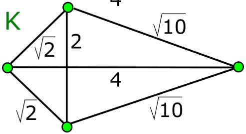
Are there *stronger invariants* of point clouds?

Pointwise Distance Distributions

For a set S of m points $p_1, \dots, p_m$ in a metric space, choose any number $1 \le k < m$ of neighbors and build the $m \times k$ matrix $D(S;k)$.



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



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

Collapse identical rows and assign weights. The matrices PDDs are continuously compared by *Earth Mover's Distance* (EMD), NeurIPS 2022.

Invariants stronger than PDD

Conjecture: PDD is complete for clouds in $\mathbb{R}^2$.

PDD is not complete for some clouds in $\mathbb{R}^3$, the *stronger invariants* below distinguished them.

simplest isometry invariant
SDV *Sorted Distance Vector*

near-linear time invariant **PDD**
Pointwise Distance Distribution

in any metric space

in any Euclidean space

strongest isometry invariants **SDD**
Simplexwise Distance Distribution

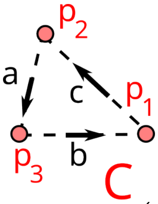
complete isometry invariants **SCD**
Simplexwise Centered Distribution

Published in: Daniel Widdowson, Vitaliy K. Computer Vision & Pattern Recognition 2023.

Higher-order distributions

Let C be a cloud of m unordered points in a *metric space*. $\mathrm{SDD}(C; h)$ for $h = 1$ is $\mathrm{PDD}(C)$.

Any sequence $A \subset C$ of h points has the matrix $\mathrm{RDD}(C; A)$ with $m - h$ permutable columns of distances from $q \in C - A$ to all points of A.



The **Relative Distance Distribution** for $A = \begin{pmatrix} p_2 \\ p_3 \end{pmatrix}$ is $\mathrm{RDD}(C; A) = [a; \begin{pmatrix} c \\ b \end{pmatrix}]$.

$\mathrm{RDD}(C; \begin{pmatrix} p_3 \\ p_1 \end{pmatrix}) = [b; \begin{pmatrix} a \\ c \end{pmatrix}]$, $\mathrm{RDD}(C; \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}) = [c; \begin{pmatrix} b \\ a \end{pmatrix}]$

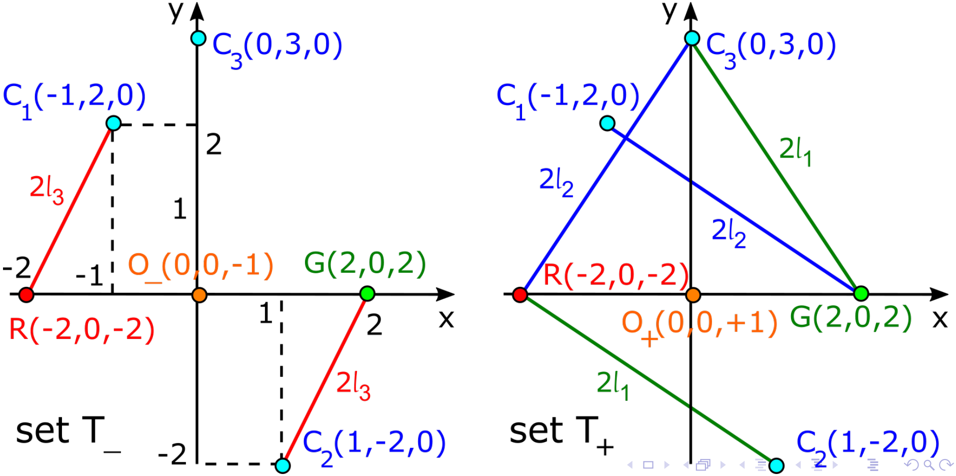
Simplexwise Distance Distribution

Classes of these RDD pairs with the distance matrix of A (up to permutations of points in A) for all h-point unordered subsets $A \subset C$ form **SDD**$(C; h)$. For $h = 2$, the stronger invariant $\mathrm{SDD}(C; 2)$ *distinguished all counter-examples* to the completeness of easier past invariants in $\mathbb{R}^3$

Theorem : for any m-point cloud C in a *metric space*, $\mathrm{SDD}(C; h)$ is computable in time $O(m^{h+1}/(h-1)!)$ and has Lipschitz constant 2 in EMD, time $O(h!(h^2 + m^{1.5}\log^h m)l^2 + l^3 \log l)$.

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

The 6-point sets $T_\pm$ with 3 free parameters have the same $\mathrm{PDD}(T_\pm)$ but *different* $\mathrm{SDD}(T_\pm;2)$.



Simplexwise Centered Distribution

In $\mathbb{R}^n$, fix the center of a cloud C at $p_0 = 0 \in \mathbb{R}^n$.

For any *ordered* subset $A = (p_1, \ldots, p_{n-1}) \subset C$, $\mathrm{OCD}(C; A)$ is the pair of the distance matrix $D(A \cup \{0\})$, matrix M with $m - n + 1$ permutable columns of n distances $|q - p_i|$ for $q \in C - A$.

To reconstruct $C \subset \mathbb{R}^n$ up to rigid motion, we add the *sign of the determinant* on the vectors from each $q \in C - A$ to the points $p_0, \ldots, p_{n-1}$.

SCD(C) is the *unordered set of classes* of $\mathrm{OCD}(C; A)$ for all $(n-1)$-point subsets $A \subset C$.

The key to Lipschitz continuity

The discontinuity of a sign in degenerate cases such as 3 points in a line is resolved by the new **strength** $\sigma(B) = V^2/p^{2n-1}$ of a simplex, where V is the volume, p is the half-perimeter of B.

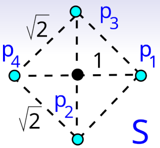
The strength of a triangle $B \subset \mathbb{R}^2$ with sides a, b, c is $\sigma(B) = \dfrac{(p-a)(p-b)(p-c)}{p^2}$, which is 'roughly linear' unlike the 'quadratic' area of B.

Theorem : in $\mathbb{R}^n$, the strength σ is Lipschitz continuous with a constant λ_n, e.g. $\lambda_2 = 2\sqrt{3}$.

Complete invariant SCD in $\mathbb{R}^n$

Theorem : for any n-dimensional cloud C of m unordered points, the *Simplexwise Centered Distribution* $\mathrm{SCD}(C)$ is a **complete invariant** under rigid motion in $\mathbb{R}^n$, computable in time $O(m^n/(n-4)!)$, has Lipschitz constant 2 in the Earth Mover's Distance (EMD), computable in time $O((n-1)!(n^2+m^{1.5}\log^n m)l^2+l^3\log l)$, l is the number of different OCDs in given SCDs.

The **complete isometry invariant** is the pair of $\mathrm{SCD}(C)$ and $\overline{\mathrm{SCD}}(C)$ with reversed signs.



For each 1-point subset $A = \{p\} \subset S$, the distance matrix $D(A \cup \{0\})$ on two points is one number 1. Then $M(S; A \cup \{0\})$ has $m - n + 1 = 3$ columns. For $p_1 = (1, 0)$, we have

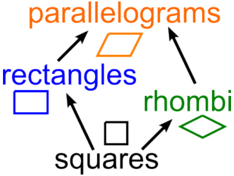
$$M(S; \begin{pmatrix} p_1 \\ 0 \end{pmatrix}) = \begin{pmatrix} \sqrt{2} & \sqrt{2} & 2 \\ 1 & 1 & 1 \\ - & + & 0 \end{pmatrix},$$

whose three columns are ordered as p_2, p_3, p_4. The sign in the bottom right corner is 0 because $p_1, 0, p_4$ are in a straight line. By the rotational symmetry,

$$\mathrm{SCD}(S) \text{ is one } \mathrm{OCD} = [1, \begin{pmatrix} \sqrt{2} & \sqrt{2} & 2 \\ 1 & 1 & 1 \\ - & + & 0 \end{pmatrix}].$$

Cloud Isometry Spaces CIS$(\mathbb{R}^n; m)$

CIS$(\mathbb{R}^n; m)$ is the space of isometry classes of clouds of m unordered points in $\mathbb{R}^n$. For $m = 4$, (sub)classes of quadrilaterals in $\mathbb{R}^2$ are often visualized by a *tree, not on a continuous map*.

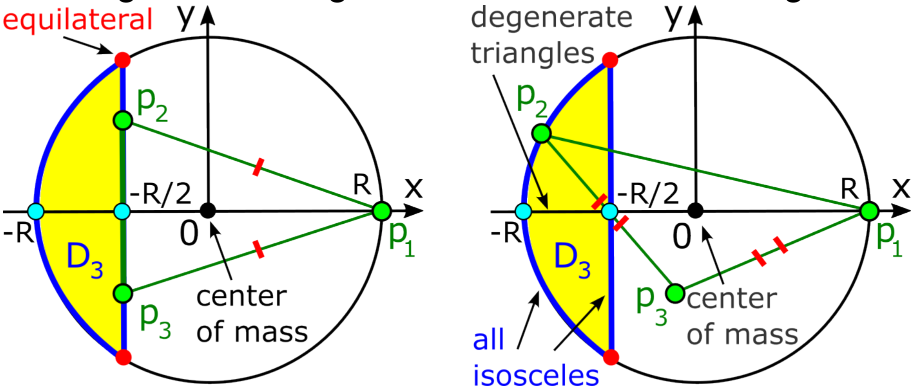


CIS$(\mathbb{R}^2; 4)$ has dimension 5. Applying uniform scaling gives the smaller quotient **CSS**$(\mathbb{R}^2; 4)$ of dimension 4.

The tetrahedron on any 4 points in $\mathbb{R}^2$ has volume 0 expressed via 6 pairwise distances, which are *unsuitable for a geographic-style map*.

Cloud Similarity Space CSS$(\mathbb{R}^2; 3)$

Triangles under rigid motion + uniform scaling.

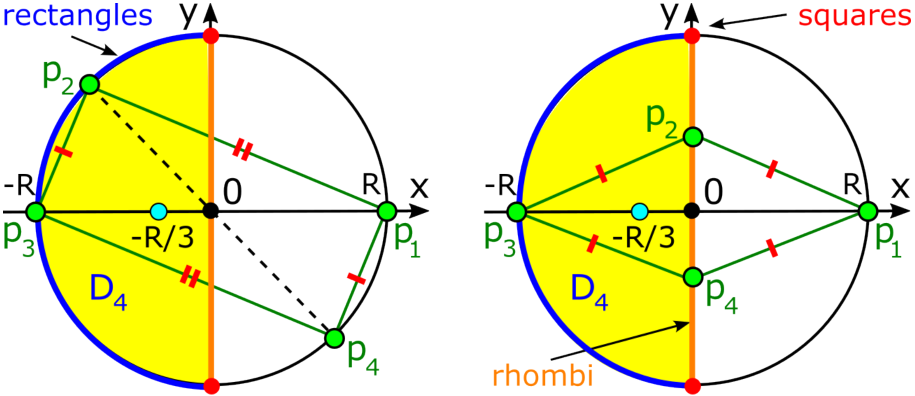


Fix the center of mass at 0, point p_1 at $(R, 0)$. A position of p_2 in the yellow region determines p_3.

Continuous map of paralellograms

Fix the center at 0, points p_1, p_3 at $(\pm R, 0)$.

Then p_2 in the yellow region determines p_4.



Rectangles and rhombi live on the boundary.

Hierarchy of rigid invariants

Fast: *Sorted Radial Vector* SRV = decreasing distances from 0 (centre of mass) to m points.

Stronger: Sorted Distance Vector SDV, $O(m^2)$.

Even stronger: $\mathrm{PDD}(C; m-1)$ in time $O(m^2)$.

Complete: $\mathrm{SCD}(C)$ in time $O(m^3)$ for $C \subset \mathbb{R}^3$.

The QM9 database has 130K+ molecules with atomic coordinates and 873,527,974 pairs of molecules of the same size. The hierarchy of the invariants above *distinguished all pairs in QM9* within a few hours on a desktop computer.

Principle of Molecular Rigidity

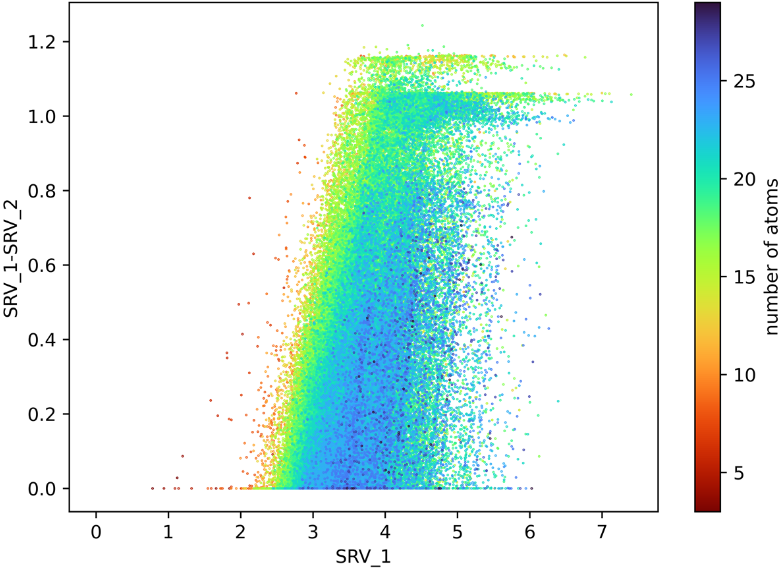
Chemically different molecules *differ rigidly*.

invariant	distance, Å	molecule A	molecule B
SRV	0.02057	$H_3C_4N_3O_2$	$H_4C_5N_2O_1$
SDV	0.05505	$H_3C_4N_5$	$H_3C_5N_3O_1$
PDD	0.05145	$H_3C_4N_5$	$H_3C_5N_3O_1$
SCD	0.07054	$H_4C_5N_4$	$H_4C_6N_2O_1$

The map: $\{$molecules$\}$ $\rightarrow$ $\{$ clouds of atomic centers$\}$ is **injective** modulo rigid motion in $\mathbb{R}^3$.

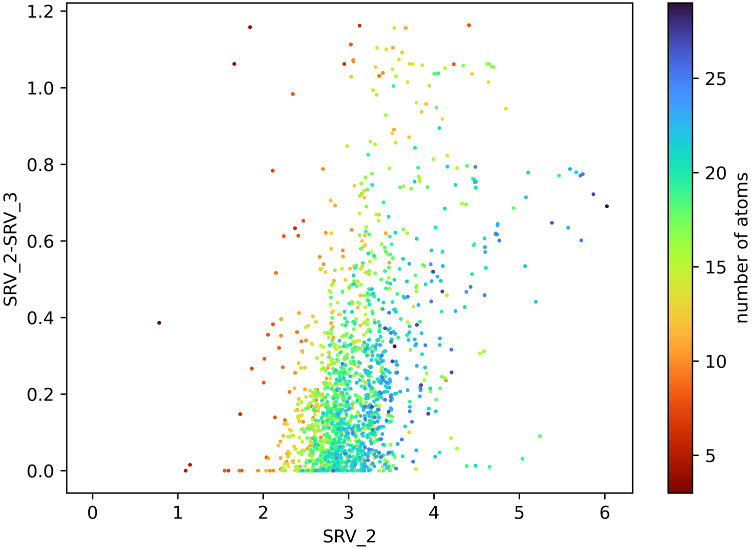
New definition : a *molecular structure* is a class of atomic clouds under rigid motion in $\mathbb{R}^3$.

Geo-graphic-style maps in GDS



All QM9 molecules. The units are Å $= 10^{-10}$ m.

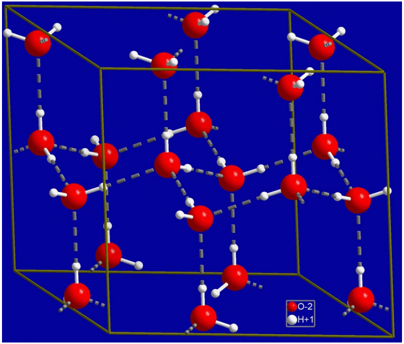
Use further invariants to zoom in

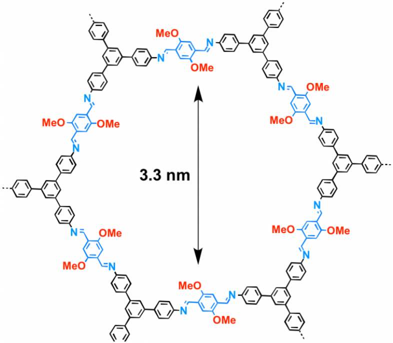


QM9 subset: $\mathrm{SRV}_1 = \mathrm{SRV}_2$.

Objects: all periodic crystals

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

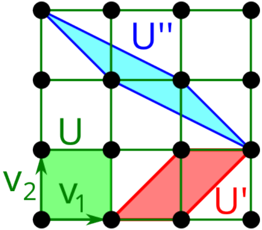


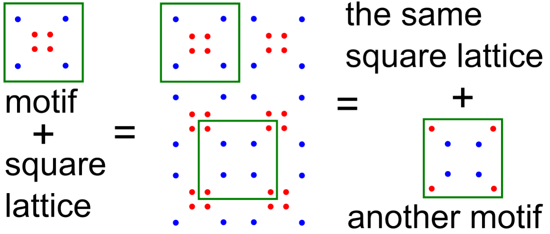


Questions: What is a crystal? What crystals are the same? If different, how much different?

A periodic point set (crystal)

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





Different pairs (basis, motif) give equivalent sets.

Was a crystal structure defined?

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

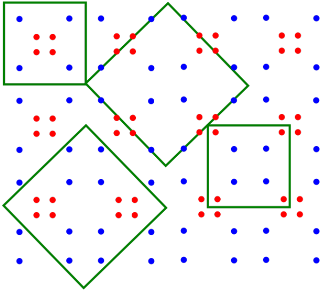
IUCr ACTIVITIES

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

Definitions are not final without equivalence.

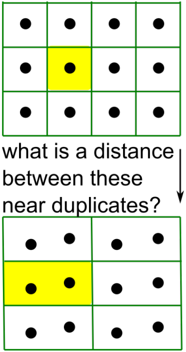
Isostructural ‘definition’

IUCr online dictionary: “crystals are said to be *isostructural* if they have the same structure ... $CaCO_3$, $NaNO_3$, $FeBO_3$ are isostructural”.



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

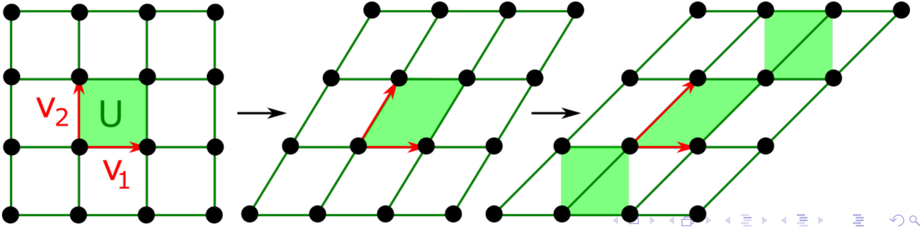
Discontinuity of conventional cells



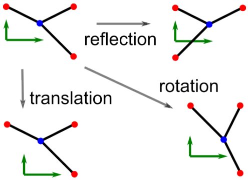
Any *reduced or conventional* cell is discontinuous under noise and atomic displacements.

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.



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 $\mathbb{R}^3$.

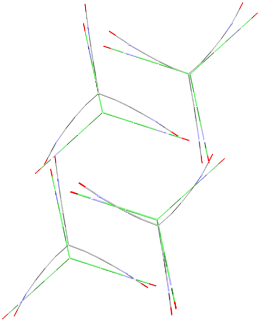
Slightly weaker : **isometry** = rigid motion + reflections = any map preserving distances.

Anosova et al. IUCrJ 2024: one periodic set $\neq$

a **crystal structure** $=$ a *rigid class* of crystals

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

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!

Descriptors vs isometry invariants

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

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

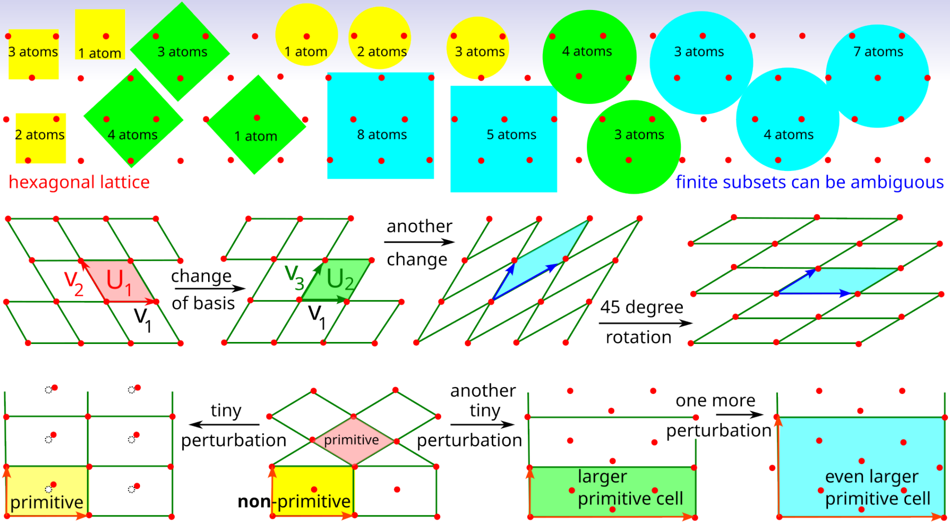
If $S \simeq Q$ are *isometric*, then $I(S) = I(Q)$; or if $I(S) \neq I(Q)$, then $S \not\simeq Q$ are not isometric.

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

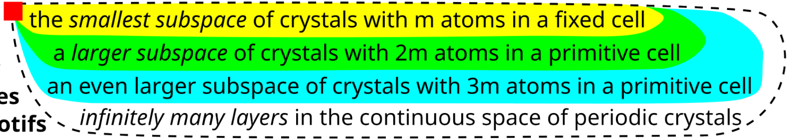
invariants can distinguish some, possibly not all crystals: density, symmetry

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

continuous **fast & reconstructable**



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



Isometry classification problem

Find an easy continuous and complete isometry invariant I for periodic 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 Q is obtained by perturbing every point of S up to ε, then $d(I(S), I(Q)) \leq \lambda\varepsilon$.

Harder practical requirements

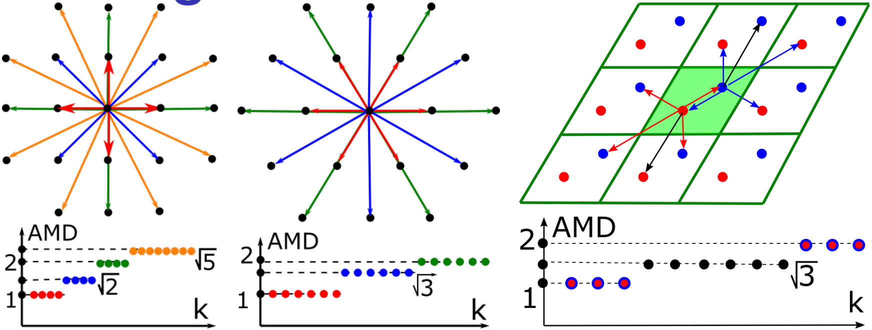
Reconstruction (inverse design): any $S \subset \mathbb{R}^n$ can be reconstructed from its invariant $I(S)$.

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* $\mathrm{AMD}_k = \frac{1}{m} \sum_{i=1}^{m} d_{ik}$.

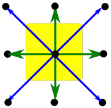
Pointwise Distance Distribution

For a periodic set S with m points $p_1, \ldots, p_m$ in a motif, choose any number $k \geq 1$ of neighbors, build the matrix $\mathrm{PDD}(S; k)$ with at most m rows.

periodic crystal structure → periodic set S of points at all atomic centres → for any point p_i in a unit cell, the i-th row of D(S;k) has distances $d_{i1} \leq \ldots \leq d_{ik}$ to k nearest neighbours of p in the infinite S

↓

order rows lexicographically collapse identical rows, and assign weights: PDD(S;k)



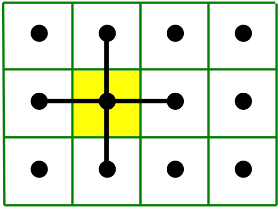
distances from p to 8 neighbours:

PDD(S;8)= | 1 | 1 | 1 | 1 | 1 | $\sqrt{2}$ | $\sqrt{2}$ | $\sqrt{2}$ | $\sqrt{2}$ |

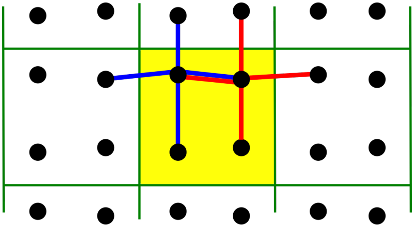
the final single row has weight 1

For different S, the matrices $\mathrm{PDD}(S; k)$ can have different numbers of rows with weights but can be compared by Earth Mover's Distance.

Earth Mover's Distance (EMD)



any small perturbation → continuously affects PDD



Continuity. If we perturb all points of a set S within their ε-neighborhoods, 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 minimizes a cost of matching weighted rows.

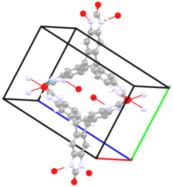
Key results from NeurIPS 2022

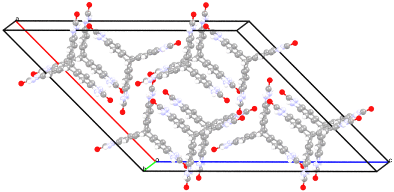
Increasing a number k of neighbors only adds more columns, k is a *degree of approximation*.

Theorem. Any *generic* periodic point set S (with distinct inter-point distances ignoring periodicity) can be uniquely reconstructed from the lattice invariants and $\mathrm{PDD}(S; k)$ with all distances up to a double covering radius of S in dimension 2, 3.

Theorem. 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’

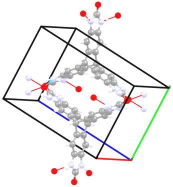


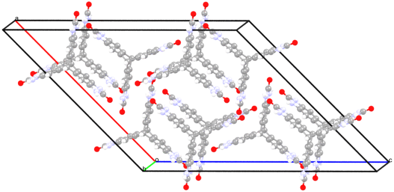


T2-14 vs T2-15 crashed Platon comparisons.

360B+ pairwise comparisons of PDD invariants over *minutes on a modest desktop* for 850K+ periodic crystals in the Cambridge Structural Database (CSD)

5 pairs of ‘needles in a haystack’





T2-14 vs T2-15 crashed Platon comparisons.

360B+ pairwise comparisons of PDD invariants over *minutes on a modest desktop* for 850K+ 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 *37+ thousand years* for all pairwise comparisons on the same machine.

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

The big **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), replace some atoms, and claim as *new* (CCDC has implemented PDD).

Nature papers in November 2023

Google's GNoME paper (led by Dogus Cubuk): DFT computations predicted 2+ million crystals.

Berkeley's A-lab paper (led by Gerbrand Ceder) claimed to have synthesized 43 of 58 crystals.

The review by Palgrave and Schoop in PRX Energy (March 2024): "*none of the materials by A-lab were new* : the large majority were misclassified, and a smaller number were correctly identified but already known", see our analysis in Widdowson et al, arxiv:2410.13796.

Google's GNoME database

Google has made public 384K+ '*stable*' crystals (close to the boundary of the convex hull of known crystals): any such '*stable*' crystal can be perturbed to get many more slightly different (new?) 'stable' crystals, also on the boundary.

The review "Artificial Intelligence Driving Materials Discovery?" (Chemistry of Materials, April 2024) by R.Seshadri and A.Cheetham found "*scant evidence for compounds that fulfill the trifecta of novelty, credibility, and utility*".

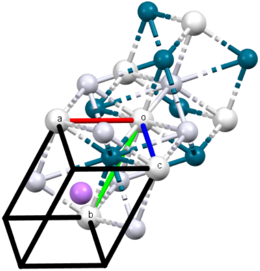
Thousands of (near)-duplicates

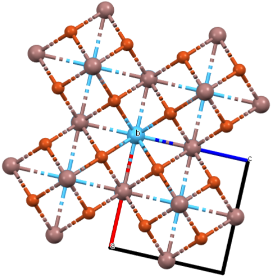
Many GNoME's crystals have geometric near-duplicates in the ICSD and Materials Project, measured by EMD on PDDs with $k = 100$.

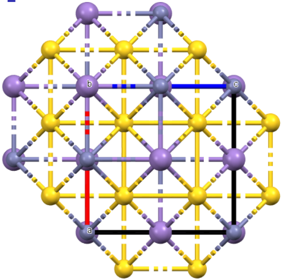
$\mathrm{EMD} \leq$	10^{-5}	10^{-4}	10^{-3}Å	0.01	0.02	0.03
ICSD	38	303	757	2454	6002	13165
Mat. Proj.	83	452	848	3457	10725	24416

Since the smallest inter-atomic distance is about $1\AA = 10^{-10}$m, any perturbations of atoms up to a small fraction of $1\AA$ can be considered noise.

Example near-duplicates







These crystals are perturbations up to 10^{-4} $\AA$.

crystal	database	ID	composition
1st	GNoME	01cd76eb18	LiScPdPt
2nd	ICSD	54594	Cu_2HfIn
3rd	Mat. Project	1186003	$MnZnAu_2$

Identical CIFs in the GNoME

Filtering by unit cells detected numerous duplicates.

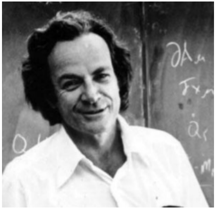
group size = #CIFs	CIFs are identical texts	all numbers coincide	rounding to 4 digits	rounding to 2 digits
10	0	0	0	1
9	0	1	1	0
7	0	1	1	2
6	0	2	2	3
5	0	2	3	16
4	1	6	8	81
3	43	21	43	557
2	1,089	411	872	6,950
total	2,311	3,258	4,259	18,328

The largest group of 9+1 duplicates

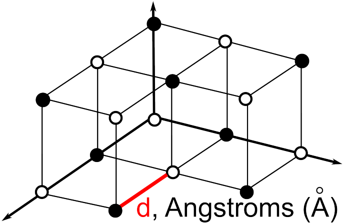
GNoME id	chemical formula	all digits are equal
082738d51d	$Dy_1Y_6Ho_{13}Cd_6Ru_2$	in a group of 9
1fba8c028f	$Dy_2Y_4Ho_{14}Cd_6Ru_2$	9
39fe92e2ee	$Tb_2Y_4Ho_{14}Cd_6Ru_2$	9
6d47ae3d9f	$Tb_3Y_3Ho_{14}Cd_6Ru_2$	9
703ed1d823	$Tb_6Ho_{14}Cd_6Ru_2$	9
78fcd9d814	$Tb_1Y_5Ho_{14}Cd_6Ru_2$	9
976f8cb279	$Y_6Ho_{14}Cd_6Ru_2$	9
a30e9d8c9b	$Tb_5Y_1Ho_{14}Cd_6Ru_2$	9
b8c0e953e2	$Tb_4Y_2Ho_{14}Cd_6Ru_2$	9
a18d30a9fc	$Tb_6Ho_{14}Cd_6Re_2$	in a group of 1

CRISP: Crystal Isometry Principle

Map: {**real 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**.

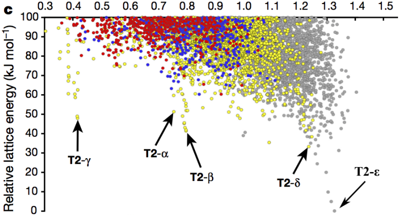


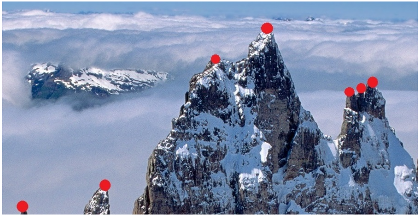
●	○	d (Å)
Na	Cl	2.82
K	Cl	3.14
Ag	Cl	2.77
Mg	O	2.10
Pb	S	2.98
Pb	Se	3.07
Pb	Te	3.17



Vision of the crystal universe

Sally Price FRS: *embarrassment of over-prediction*, too many optimized structures: only local peaks, no locations.

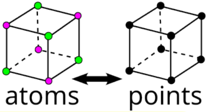




all **real periodic crystals** of atoms with *chemical elements*

different crystals ⟷ different structures

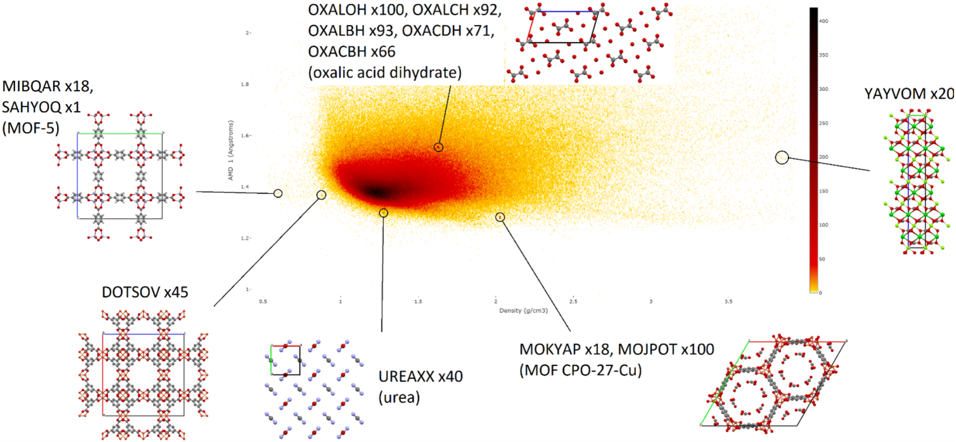
geometry *suffices* to reconstruct chemistry



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

isometry classes of point sets outside the crystal space

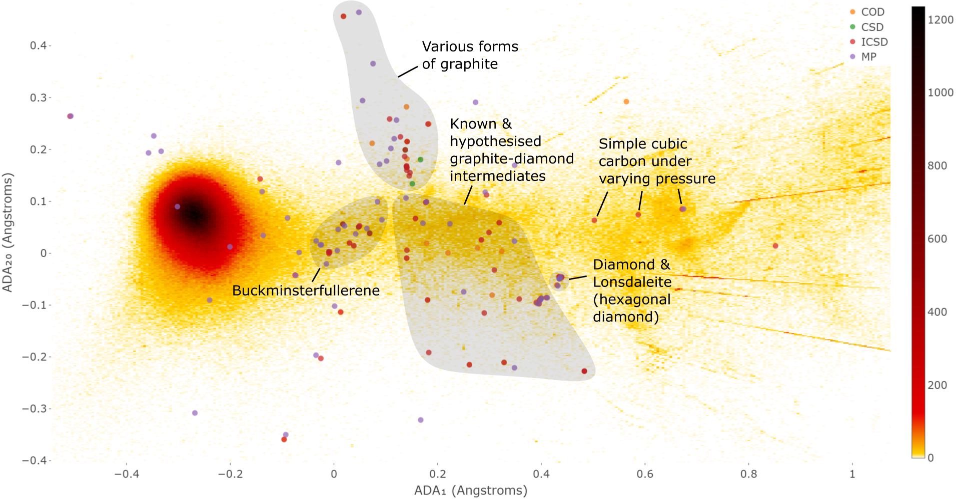
CSD in meaningful coordinates



It's a projection with well-defined coordinates.

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

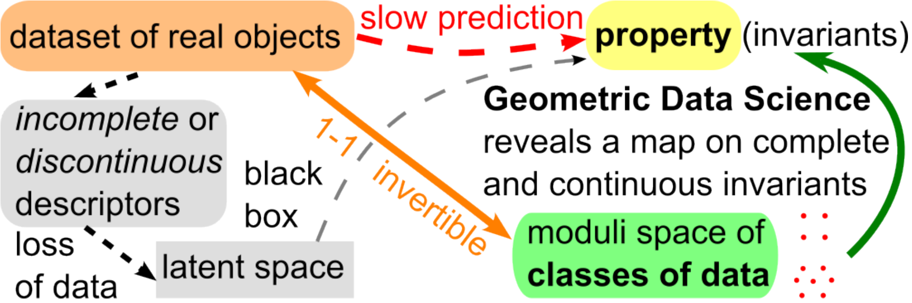
Carbon allotropes on a crystal map



$\mathrm{ADA}_1 =$ average distance to the 1st atomic neighbor (adjusted by subtracting the proved asymptotic curve).

Next: structure → properties

In the past, properties were slowly predicted by incomplete or discontinuous descriptors.



Without losing data, one can express properties of structures as *functions of invertible invariants* on geographic-style maps of moduli spaces.

Collaborations are welcome!

We can similarly explore continuous spaces of other data objects (graphs, images, meshes) under rigid motion or other equivalences.