Geometric Data Science

Vitaliy Kurlin's Data Science group

Olga Anosova, Dan Widdowson, Yury Elkin

Materials Innovation Factory, Liverpool (UK)

and NITMB (National Institute for Theory and Mathematics in Biology), Chicago (US)

THE ROYAL SOCIETY

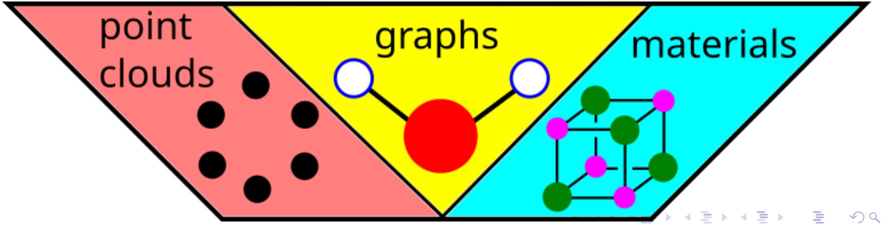
Data in Geometric Data Science

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

Data can be any real (geometric) objects.

Finite case: a cloud of (un)ordered points (atoms in a molecule) or embedded graph (simplicial complex) in $\mathbb{R}^n$ or a metric space.

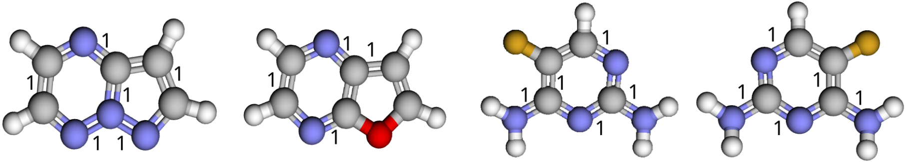
Periodic case: lattices, crystalline materials, ...



Question 1: *same* or different?

Different digital inputs, such as files of atomic coordinates, often represent *same* real object.

What molecules below could be *the same*?



A proper answer needs a definition of *equivalent* objects that are the same in a given problem.

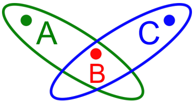
Three axioms of an equivalence

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

By size: molecules $A \sim B$ if they have the same number of atoms. Useful, but $H_2O \sim CO_2$.

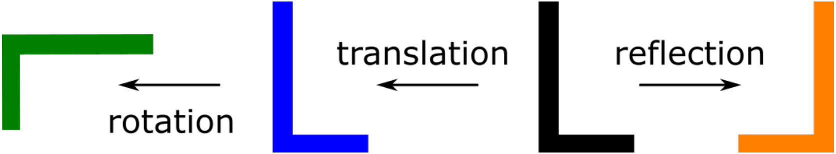
By property: molecules $A \sim B$ if A, B have the same property. Ok, but molecules that share one property can differ by other properties.

By coincidence: molecules $A \sim B$ if A, B have the same xyz files that list all atomic coordinates in an arbitrarily chosen coordinate system.

What is the strongest *practical* equivalence?

Equivalence by rigid motion

Many real-life objects are rigid and should be considered equivalent under **rigid motion** ($\cong$):



Compositions of translations and rotations in $\mathbb{R}^n$ form the group $\mathrm{SE}(n)$. If we allow reflections, we get the Euclidean group $\mathrm{E}(n)$ of **isometries**.

In a metric space, an **isometry** ($\simeq$) is any map that preserves all inter-point distances.

Point sets: ordered vs unordered

If all m points of a given set $C \subset \mathbb{R}^n$ 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 (uniformly) **continuous** under perturbations: perturbing any p_i up to ε changes any distance d_{ij} only up to 2ε. Many point sets are unordered (called *point clouds*).

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

Invariants distinguish objects

Classes of point clouds under rigid motion 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

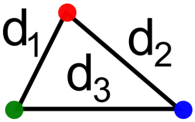
if $I(A) \neq I(B)$ then $A \not\sim B$ meaning that I has

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

I is called *complete* if $I(A) = I(B) \Rightarrow A \sim B$, i.e. I has *no false positives*: $A \not\sim B$ with $I(A) = I(B)$.

Q2: if different, by how much?

Due to noise, all real objects 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 $d(I(S), I(Q)) = d(I(Q), I(S))$,
$\triangle$ triangle inequality $d_1 + d_2 \geq d_3$.

Rass et al. Trans. Information Forensics 2024:
if we allow the triangle inequality to *fail with any small additive error*, then the results of k-means clustering and DBSCAN can be pre-determined.

General complicated metrics

For any metric spaces (X, d_X) and (Y, d_Y), let $R(X, Y)$ consist of all correspondences R, where $R \subset X \times Y$ covers both X, Y. Then $\mathrm{GH} = \inf\limits_{R} \sup\limits_{(a,b)\in R,(a',b')\in R} |d_X(a,b) - d_Y(a',b')|$ is a metric but cannot be approximated with a factor less than 3 in polynomial time unless P$=$NP by Shmiedl, Discrete Comp. Geometry 2017.

Minimising for all bijections $S \to Q$ in a metric space gives $\mathrm{BD}(S, Q) = \inf\limits_{g:S\to Q} \sup\limits_{p\in S} d(g(p), p)$.

Geo-mapping problem: finite sets

Find a (complete, bi-continuous, and poly-time) *geocode* I for finite 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

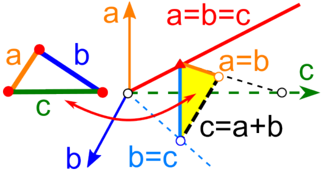
Invertibility: any $S \subset \mathbb{R}^n$ can be reconstructed from its invariant $I(S)$ in a *continuous* way.

Computability: the invariant I, the metric d, and a reconstruction of $S \subset \mathbb{R}^n$ from $I(S)$ can be obtained in polynomial time of the size of S, forbidding *infinite or exponential size* invariants.

Geo-style maps: describe all *realizable values* $I(S)$ that allow us to reconstruct a cloud $S \subset \mathbb{R}^n$, parameterize the moduli space of clouds under isometry to answer **Q3**: *where do they live?*

Euclid's ideal solution for triangles

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



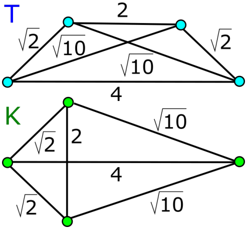
The **Cloud Isometry Space** $\mathrm{CIS}(\mathbb{R}^n; 3)$ is the cone in $\mathbb{R}^3$ $\{0 < a \leq b \leq c \leq a+b\}$ continuously parametrized by three inter-point distances a, b, c.

Genericially complete invariants

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

One can train neural networks to experimentally output invariants, but there is no proof of their *finite completeness* and *Lipschitz continuity*.

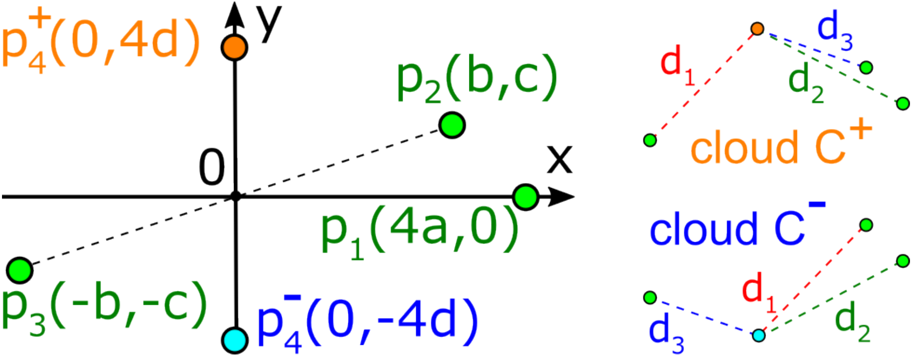
Boutin, Kemper, 2004: the vector of all sorted pairwise distances is *generically complete* in $\mathbb{R}^n$



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

Clouds with the same 6 distances

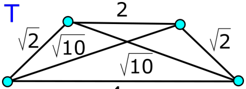
Pairs of non-isometric clouds $\{p_1, p_2, p_3, p_4^{\pm}\}$ in $\mathbb{R}^2$ (depending on 4 parameters $a, b, c, d > 0$) have the same 6 pairwise distances below.



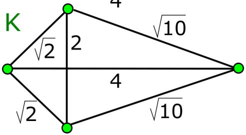
What invariant can distinguish these clouds?

Pointwise Distance Distributions

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



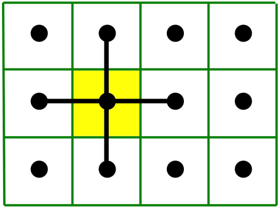
$$\mathrm{PDD}(T;3) = \left(\begin{array}{c|ccc} 1/2 & \sqrt{2} & 2 & \sqrt{10} \\ 1/2 & \sqrt{2} & \sqrt{10} & 4 \end{array}\right) \neq$$



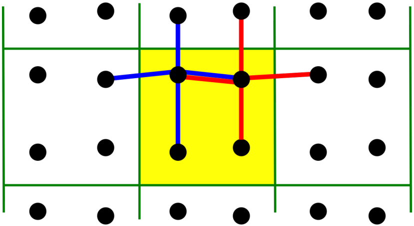
$$\mathrm{PDD}(K;3) = \left(\begin{array}{c|ccc} 1/4 & \sqrt{2} & \sqrt{2} & 4 \\ 1/2 & \sqrt{2} & 2 & \sqrt{10} \\ 1/4 & \sqrt{10} & \sqrt{10} & 4 \end{array}\right).$$

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

Earth Mover's Distance (EMD)



any small perturbation $\longrightarrow$ continuously affects PDD



[Widdowson, VK. NeurIPS 2022] If we perturb all points of a set S up to ε, then 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)=

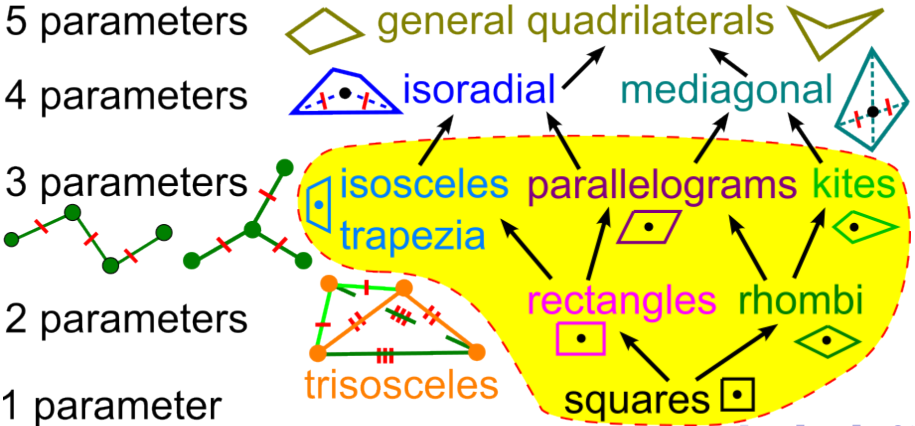
weight 0.5	0.8	1.005	1.005	1.2
weight 0.5	1	1	1.005	1.005

EMD = 0.5 (0.2+0.005) = 0.1025 ⩽ 0.2 bound

EMD minimises a cost of matching weighted rows.

New types of 4-point clouds in $\mathbb{R}^2$

Thm 5.3 in SIAM J Appl. Math. (2026): PDD is *complete* for any $m \le 4$ unordered points in $\mathbb{R}^n$.



Invariants stronger than PDD

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

Some clouds in $\mathbb{R}^3$ have equal PDDs. We have extended the PDD to a *complete* SCD in $\mathbb{R}^n$.

simplest invariants **SRD** and **SPD**
Sorted Radial/Pairwise Distances

quadratic-time invariant **PDD**
Pointwise Distance Distribution

in any metric space

in any Euclidean space

stronger isometry invariant **SDD**
Simplexwise Distance Distribution

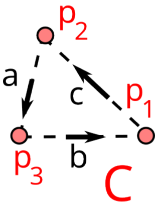
complete isometry invariant **SCD**
Simplexwise Centered Distribution

All these invariants are Lipschitz continuous with constant 2 [Widdowson, VK. CVPR 2023]. Extended in MATCH 2027 and 2303.14161.

Towards complete invariants

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 $q\in C\setminus A$ to all points of A.



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

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

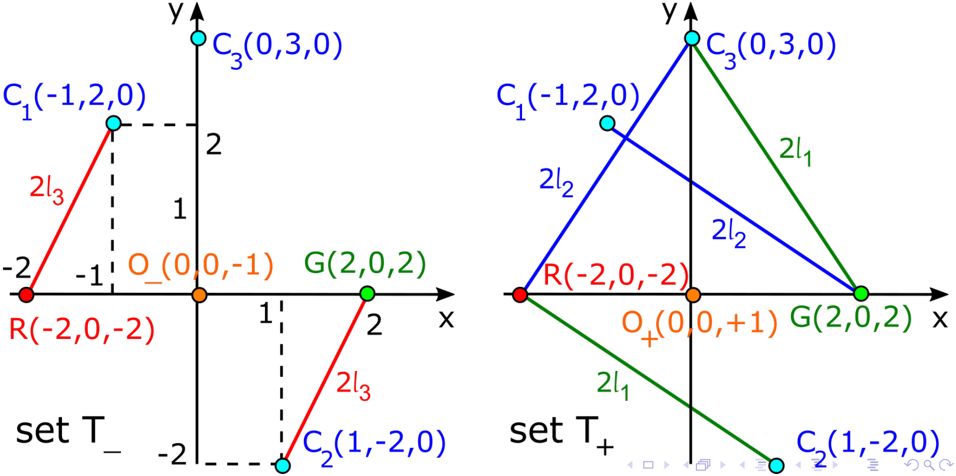
Simplexwise Distance Distribution

Classes of these RDD pairs with the distance matrix of A (under 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 $\mathbb{R}^3$

Thm: for any m-point cloud C in a *metric space*, $\mathrm{SDD}(C; h)$ of size $l = O(m^h/h!)$ is computed in time $O(ml)$ 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* $\mathrm{SDD}(T_\pm;2)$.



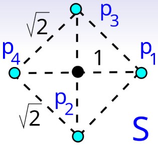
Simplexwise Centered Distribution

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

For any *ordered* subset $A = (p_1, \dots, p_{n-1}) \subset C$, $\mathrm{OCD}(C; A)$ 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 $q \in C \setminus A$.

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

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



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,

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

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.

Thm (CVPR'23): the strength σ is continuous with a Lipschitz constant λ_n, e.g. $\lambda_2 = 2\sqrt{3}$.

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

Theorem (CVPR 2023): 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 $O(m^n/(n-4)!)$, and has Lipschitz constant 2 in the EMD, which is computable in time $O((n-1)!(n^2+m^{1.5}\log^n m)l^2+l^3\log l)$. Here l is the number of different OCDs in SCDs.

The **complete isometry invariant** of $C\subset\mathbb{R}^n$ is the pair $\{\mathrm{SCD}(C),\overline{\mathrm{SCD}}(C)\}$ with reversed signs.

Hierarchy of cloud invariants

Fast: *Sorted Radial Distances* SRD (decreasing distances from 0 to m points) in time $O(m)$.

Stronger: Sorted Pairwise Distances SPD.

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	molecule A	molecule B
SRD	2.06 pm	$H_3C_4N_3O_2$	$H_4C_5N_2O_1$
SPD	5.51 pm	$H_3C_4N_5$	$H_3C_5N_3O_1$
PDD	5.15 pm	$H_3C_4N_5$	$H_3C_5N_3O_1$
SCD	7.05 pm	$H_4C_5N_4$	$H_4C_6N_2O_1$

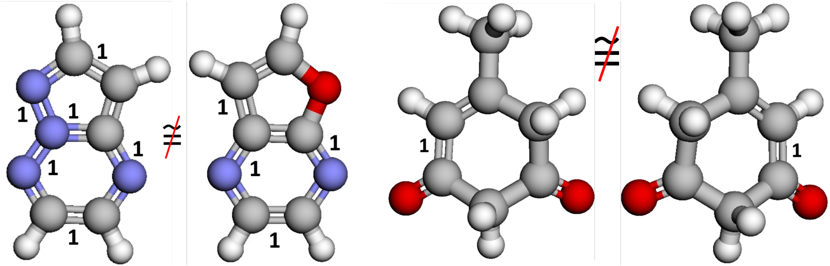
The smallest interatomic distance $\approx$ 95 pm.

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

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

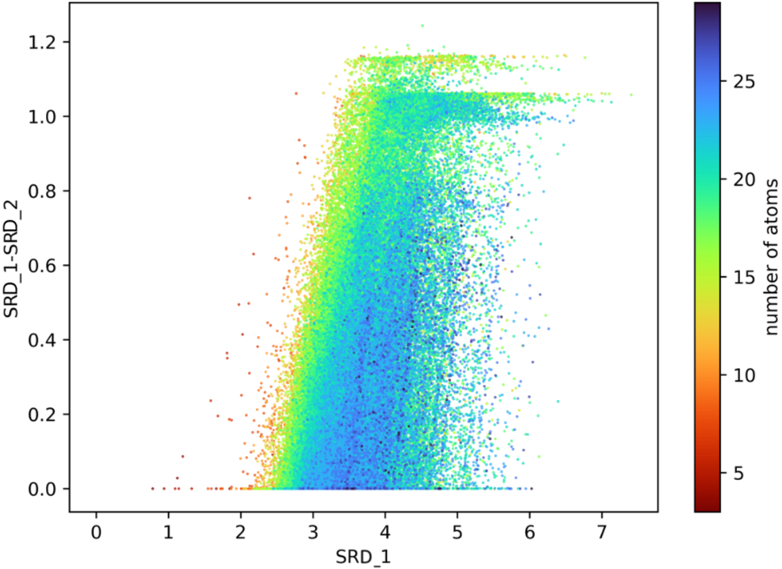
Closest molecules in QM9

Left: closest chemically different molecules 123532 and 24513 have a distance $\approx$ 7 pm.



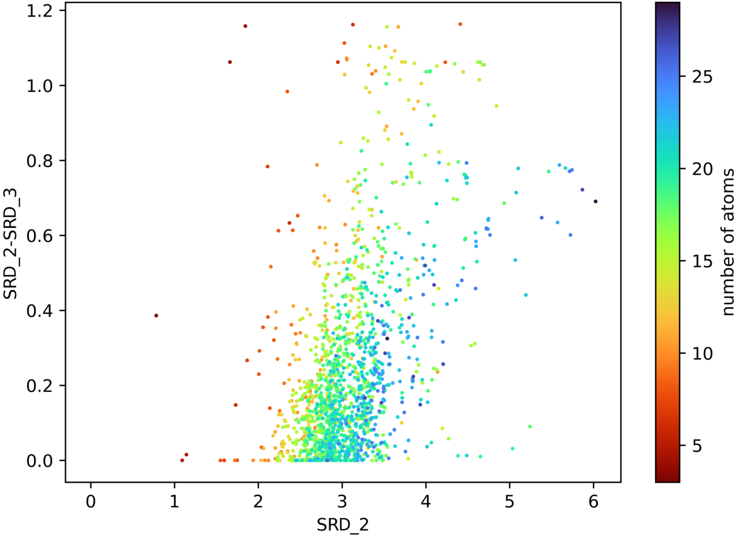
Right: almost mirror images 70954 and 74130 have $\mathrm{EMD} \approx$ 0.04 pm on PDDs, distinguished by a distance $\approx$ 162 pm under rigid motion.

Geo-graphic-style maps of QM9



All QM9 mole-cules. The units: 1Å = 100 pm.

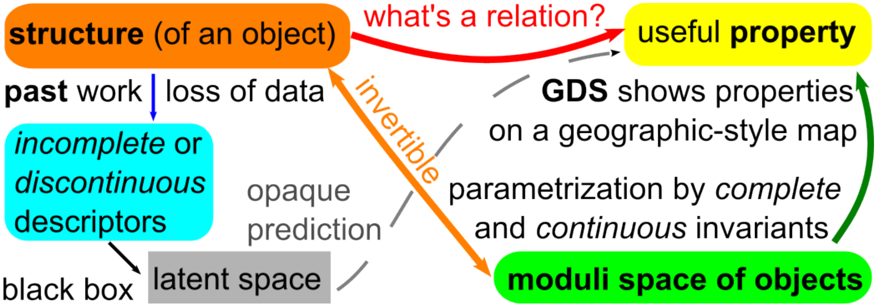
Use further invariants to zoom in



QM9 subset: $\mathrm{SRD}_1 = \mathrm{SRD}_2$ equal radii from a centre.

Geo-graphic-style maps in GDS

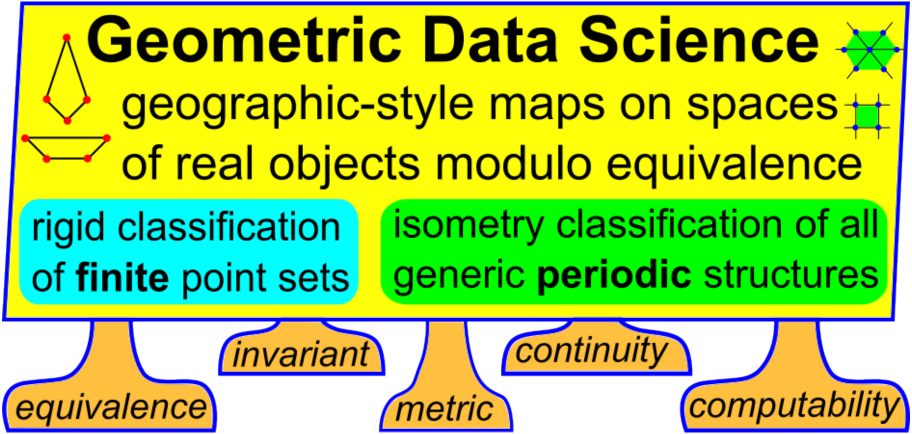
Non-invariant (or incomplete or discontinuous) descriptors *lose data* in *cramped latent spaces*.



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

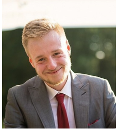
Collaborations are welcome!

Questions for any objects: Same or different? If different, by how much? Where do they all live?

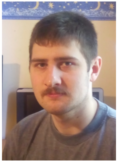


Many thanks to the key co-authors!

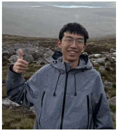
Dr Olga Anosova. Geometry of proteins. Mapping a continuous space of proteins.



Dr Dan Widdowson. Geometry of crystals. Developing the *Crystal Geomap* software for visualising large materials databases.



Dr Yury Elkin. Geometry of molecules. Mergegrams, nearest neighbour search, and invariants of molecular structures.



Mr Ziqiu Jiang (Henry). Atomic clashes in the Protein Data Bank and AlphaFold.