Continuous machine learning on Euclidean graphs with unordered vertices

Abstract

Molecular graphs can change their chemical properties under non-rigid deformations in Euclidean space. Hence it is vitally important to distinguish rigid classes of molecular graphs under compositions of translations and rotations. Also, robust outputs of machine learning on molecular graphs embedded in Euclidean space should continuously change under perturbations, motivated by atomic vibrations and experimental noise. We developed a complete invariant that can be inverted back to an embedded graph, uniquely under rigid motion, and has a Lipschitz continuous distance satisfying all metric axioms. For a fixed dimension, the invariant and metric can be computed in polynomial time of the number m of unordered vertices and hence avoiding exponentially many permutations. The new invariants distinguish all chemically different graphs in the world's largest databases of 3D molecules in a few hours on a modest desktop.

1. Motivations for complete and continuous invariant inputs in application-driven ML

This paper formalizes necessary conditions for ML on real data with ambiguous representations and develops complete and Lipschitz continuous invariants satisfying these conditions on any Euclidean graphs and justifying a rigorous concept of a molecular structure. Many real structures from star constellations to molecules are represented by graphs embedded in a Euclidean space (Bonchev, 1991). A *Euclidean graph* $G \subset \mathbb{R}^n$ is a finite set of m unordered (unlabeled) vertices located at distinct points of $\mathbb{R}^n$ and connected by straight-line edges. Forgetting all edges of $G \subset \mathbb{R}^n$ gives us the *vertex set* $V(G) \subset \mathbb{R}^n$ of m unordered points. A Euclidean graph can be disconnected and can have vertices v of any *degree* that is the number of edges whose endpoint is v. Loops and multiple edges (with the same endpoints) do not appear in Euclidean graphs because all edges are straight line segments and can also intersect in theory.

. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Graphs can be considered under any *equivalence* relation that should satisfy the axioms: 1) *reflexivity*: $G \sim G$, 2) *symmetry*: if $G \sim F$ then $F \sim G$, 3) *transitivity*: if $G \sim F$ and $F \sim H$ then $G \sim H$. In chemistry, the simplest equivalence is by chemical composition, which is insufficient in practice, e.g. *stereoisomers* in Fig. 1 (right) have the same chemical compositions and non-equivalent rigid shapes with different chemical properties (Rieder et al., 2023).

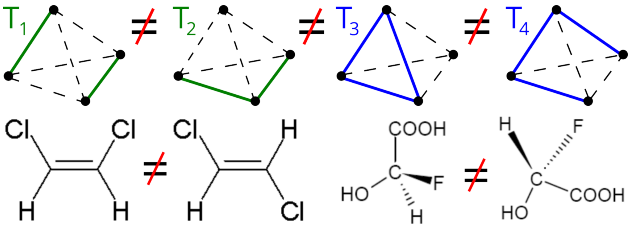

Figure 1. **Top**: graphs $T_1, T_2, T_3, T_4 \subset \mathbb{R}^3$ on the same vertices with solid edges are not isomorphic to each other. **Bottom**: stereoisomers are isomorphic combinatorially, not geometrically.

For molecules, the strongest equivalence (distinguishing as many graphs as practically possible) is a *geometric isomorphism* $G \cong F$, i.e. an orientation-preserving transformation of $\mathbb{R}^n$ that bijectively maps the vertices and edges: $G \to F$. Geometric isomorphisms are also called *rigid motions* (compositions of translations and rotations), which form the special Euclidean group $\mathrm{SE}(n)$. The slightly weaker equivalence (not distinguishing mirror images) is an *isometry*, which is any distance-preserving transformation including reflections. Any geometrically isomorphic molecules have the same chemical properties. If a flexible molecule changes its rigid shape, its functional properties can change, so it is important to distinguish rigid shapes (Wilson et al., 1991).

To reliably distinguish at least some Euclidean graphs $G \subset \mathbb{R}^n$, we need an *invariant* I defined as a numerical descriptor preserved by any rigid motion in $\mathbb{R}^n$. Alternatively, if $I(G) \neq I(F)$, then $G \not\cong F$, so any invariant has *no false negatives* that are pairs of different representatives of *rigidly equivalent graphs* (denoted by $G \cong F$) having equal values of a (non-invariant) descriptor. The number of vertices (or edges) of G is an integer-valued weak invariant that cannot separate any graphs in Fig. 1. The strongest invariant I separating all non-equivalent graphs is called *complete* meaning that if $I(G) = I(F)$ then $G \cong F$. Alternatively, a

complete invariant I has *no false positives* that are pairs of non-equivalent graphs $G \not\cong F$ with $I(G) = I(F)$.

Since all real data (such as inter-point distances) are noisy, a more practically important answer is not binary ('same or different') but should be continuously quantified by a distance metric between isometry classes. The atomic vibrations (Feynman, 1971) imply that rigid classes of molecules graphs on m unordered atoms form a continuous *Graph Isometry Space* $\mathrm{GIS}(\mathbb{R}^3; m)$. Only for triangular graphs with $m = 3$, their space was previously known due to the side-side-side theorem saying that any triangles are isometric if and only if they have the same triple of sides (inter-point distances) a, b, c considered up to 6 permutations. Hence the space of triangular graphs is $\{0 < a \le b \le c \le a+b\} \subset \mathbb{R}^3$, where $c \le a+b$ guarantees that distances a, b, c are realizable by a real triangle.

Problem 1.1 (complete invariant of Euclidean graphs with a polynomial-time continuous metric)**.** *Find a function I : $\{$Euclidean graphs with of unordered vertices in $\mathbb{R}^n\} \to$ a space X with a distance d satisfying the conditions below:*

***(a) completeness** of the invariant: any graphs G, F are related by rigid motion in $\mathbb{R}^n$ if and only if $I(G) = I(F)$;*

***(b) Lipschitz continuity:** there is a constant λ and a metric d satisfying the axioms 1) $d(\alpha, \beta) = 0$ if and only if $\alpha = \beta$, 2) $d(\alpha, \beta) = d(\beta, \alpha)$, 3) $d(\alpha, \beta) + d(\beta, \gamma) \ge d(\alpha, \gamma)$ for all $\alpha, \beta, \gamma \in X$, such that if F is obtained by perturbing every vertex of G up to $\varepsilon > 0$, then $d(I(G), I(F)) \le \lambda\varepsilon$;*

***(c) invertibility:** any Euclidean graph G can be reconstructed (uniquely up to rigid motion in $\mathbb{R}^n$) from $I(G)$;*

***(d) computability:** for a fixed dimension n, the invariant I, d, and a reconstruction of $G \subset \mathbb{R}^n$ from $I(G)$ can be obtained in polynomial time of the number of vertices.*

Condition 1.1(a) means that a complete invariant I has the strongest expressivity (Zhang et al., 2024) by uniquely identifying any Euclidean graph under geometric isomorphism. To be useful for noisy inputs, a complete invariant should continuously change under perturbations in a suitable metric. The axioms in 1.1(b) are the foundations of metric geometry (Melter & Tomescu, 1984) and accepted in chemistry (Weinhold, 1975). If the triangle axiom fails with any additive error, the classical k-means and DBSCAN clustering are open to adversarial attacks in (Rass et al., 2024). If the first axiom is ignored, $d \equiv 0$ satisfies all other axioms. The first axiom implies the completeness of I in 1.1(a) but the continuity is much stronger. Indeed, for any complete invariant I, one can define the discrete metric $d(I(G), I(F)) = 1$ for $G \not\cong F$, which unhelpfully treats all non-equivalent graphs (even near-duplicates) as equally distant. The Lipschitz continuity in 1.1(b) is necessary for smoothness, which is implicitly assumed by any gradient-based optimization.

Condition 1.1(c) requires I to be not only complete and continuous but also efficient to explicitly reconstruct G, even better than a DNA code that is does not explain how to grow a living organism. Computability 1.1(d) prevents brute-force attempts, e.g. defining $I(G)$ as the infinite set of images of G under all rigid motions or taking $m!$ distance matrices over all permutations of m unordered vertices.

The main contribution is the new invariant Nested Centered Distribution, which solves Problem 1.1, including the new Lipschitz continuity, for all Euclidean graphs in $\mathbb{R}^n$.

2. Past work on distances for Euclidean graphs

Ordered clouds. The vertex set $V(G)$ of a Euclidean graph $G \subset \mathbb{R}^n$ is called a *point cloud* C. If all points $p_1, \dots, p_m$ of C are ordered (not under the action of all $m!$ permutations), a complete invariant of C under isometry (compositions of translations, rotations, reflections) is the classical $m \times m$ matrix (Li et al., 2023) of pairwise distances $|p_i - p_j|$ due to Theorem 9 in (Grinberg & Olver, 2019) or, after shifting the center of mass to the origin, the Gram matrix of scalar products $p_i \dot{p}_j$ by Theorem 1 in (Dekster & Wilker, 1987). This multidimensional scaling (Schoenberg, 1935) can also provide an embedding $C \subset \mathbb{R}^k$ preserving all distances of C for a dimension $k \le m$. This embedding $C \subset \mathbb{R}^k$ uses eigenvectors whose ambiguity up to signs gives an exponential time that can be close to $O(2^m)$, not polynomial in the number m of ordered points as in 1.1(d).

Unordered clouds. Computational geometry developed many algorithms for detecting geometric isomorphism (or isometry, also called congruence) between point sets without edges (Huttenlocher et al., 1993; Chew & Kedem, 1992; Chew et al., 1999; Goodrich et al., 1999). For a set $A \subset \mathbb{Q}^n$ of m points, Theorem 3 in (Arvind & Rattan, 2016) computed in time $n^{O(n)} poly(mM)$ a canonizing function $f(A)$, which can be considered a complete isometry invariant of A, where M upper bounds the binary encodings of the rational coordinates in the input. For point clouds under rigid motion (also distinguishing mirror images), Theorem 4.7 in (Widdowson & Kurlin, 2023) described a metric computable in time $O(n(m^{n-1}/n!)^3 \log m)$. (Hordan et al., 2024; Delle Rose et al., 2024; Nigam et al., 2024; Amir et al., 2024; Maennel et al., 2024) also achieved the completeness for point clouds but without a Lipschitz continuous metric as in 1.1(b). Energy potentials written as infinite series of spherical harmonics, are often considered complete representations of atomic environments, which holds in the limit but not for a finite size(Pozdnyakov et al., 2020). For a fixed set of m vertices in general position, one can choose any of $m(m-1)/2$ edges and produce $2^{m(m-1)/2}$ non-isometric graphs. Problem 1.1 for arbitrary graphs is computationally much harder than for point clouds due to exponentially many different graphs on the same vertex set.

The graph isomorphism problem (Grohe & Schweitzer, 2020) for abstract (non-Euclidean) graphs is another version of Problem 1.1 without continuous metrics. The latest advances (Babai, 2019; Helfgott et al., 2017) achieved only quasipolynomial time. While many partial cases were solved, e.g. for planar graphs (embedded in $\mathbb{R}^2$ without intersecting edges), see (Kiefer et al., 2019), the k-dimensional Weisfeiler-Leman test (Leman & Weisfeiler, 1968) fails for 3-regular graphs of size $O(k)$. The key limitation of WL tests is their local nature when invariants are gradually expanded from a vertex or a k-tuple. Then covering a graph on m vertices needs $O(m)$ expansions leading to exponential sizes in m. Section 3.9 in (Dym & Gortler, 2024) discussed that a complete invariant (under all permutations of m vertices) that has a polynomial time in the dimension n would also solve the graph isomorphism problem in polynomial time. Condition 1.1(d) is easier for a fixed dimension n, e.g. $n = 2, 3$ are practical cases. The number m of vertices can be dozens or hundreds, e.g. for molecular graphs in $\mathbb{R}^3$, where vertices are centers of atoms and edges are interatomic bonds that keep atoms together in a stable molecule.

Geometric Deep Learning in (Bronstein et al., 2021) pioneered an axiomatic approach to geometric classifications beyond Euclidean space $\mathbb{R}^n$ in (Bronstein et al., 2017). Some neural networks were proved to be universal (Maron et al., 2019; Zhou, 2020; Abbe & Sandon, 2020) in the sense of approximating any continuous function on given data with sufficiently many layers. This universality property has been strengthened in Problem 1.1 to the full completeness of an explicit invariant that should be computable in polynomial time and invertible to an original graph up to rigid motion. The key challenge was to compute an exact (not approximate) metric that is also Lipschitz continuous.

Equivariants (Kondor & Trivedi, 2018; Cohen et al., 2019; Fuchs et al., 2020; Deng et al., 2021) are defined as descriptors E satisfying $E(f(G)) = T_f(E(G))$ for any rigid motion f and all graphs $G \subset \mathbb{R}^n$, where T_f can be any map, not only the identity as for invariants. Any linear combination of points, e.g. the center of mass, is equivariant but cannot distinguish graphs under translation. Equivariants (Gao et al., 2020; Qi & Luo, 2020; Tu et al., 2022; Batzner et al., 2022) help predict forces acting on atoms to move them to a more optimal configuration. These time-dependent graphs G_t can be studied directly by invariant values $I(G_t)$ without computing intermediate atomic forces.

Many neural networks optimize millions of parameters, e.g. see Table 4 (Goyal et al., 2021), to achieve great accuracies (Dong et al., 2018; Akhtar & Mian, 2018; Laidlaw & Feizi, 2019; Guo et al., 2019; Colbrook et al., 2022) but require re-training on any new data. All known descriptors of molecular graphs (Duvenaud et al., 2015; Choo et al., 2023) have no proofs of all conditions 1.1(a,b,c,d).

Gromov-Wasserstein metrics (Mémoli, 2011) are defined for any metric-measure spaces (Brécheteau, 2019) by minimizing over infinitely many correspondences between points, but cannot be approximated with a factor less than 3 in polynomial time unless P=NP by Corollary 3.8 in (Schmiedl, 2017) and Theorem 3.3 in (Agarwal et al., 2018), see fast algorithms for important cases in (Mémoli et al., 2021; Lim et al., 2023; Majhi et al., 2024). (Nikolentzos et al., 2017; Majhi & Wenk, 2022; Buchin et al., 2023) made significant advances in the related problems of matching and finding distances between fixed Euclidean graphs without considering isometry. Computing a metric between rigid classes is only a small part of Problem 1.1. Indeed, to efficiently navigate on Earth, in addition to distances between cities, we need a map of the planet and hence an invertible continuous invariant I similar to geographic coordinates.

3. Graph invariants: from fastest to complete

Let $|p-q|$ denote the Euclidean distance between any points $p, q \in \mathbb{R}^n$. We always translate any graph $G \subset \mathbb{R}^n$ so that the *center of mass* $O(G) = \dfrac{1}{m} \sum\limits_{p \in V(G)} p$ of the *vertex set* $V(G)$ is at the origin $0 \in \mathbb{R}^n$. Then Problem 1.1 reduces to the $\mathrm{SO}(n)$-invariance under orthogonal transformations.

Definition 3.1 (signed distance $d(p,q)$ and invariants $\mathrm{SRD}, \mathrm{SPD}, \mathrm{PDD}$)**.** *Let $G \subset \mathbb{R}^n$ be any Euclidean graph on m arbitrarily ordered vertices $p_1 \dots, p_m$. If any $p_i, p_j \in V(G)$ are connected by an edge of G, define the* signed distance *as $d(p,q) = |p-q|$, else set $d(p,q) = -|p-q|$.*

(a) *The vector* $\mathrm{SRD}(G)$ *of* sorted radial distances *consists of m distances $|p|$ for all $p \in V(G)$ in decreasing order.*

(b) *The vector* $\mathrm{SPD}(G)$ *of* sorted pairwise distances *consists of all distances $d(p_i, p_j)$ in decreasing order.*

(c) *Let $D(G)$ be the $m \times (m-1)$-matrix whose the i-th row consists of $d(p_i, p_j)$, $j \in \{1, \dots, m\} \setminus \{i\}$, in increasing order. The* Pointwise Distance Distribution $\mathrm{PDD}(G)$ *consists of these unordered rows with equal weights $1/m$.*

If any $k > 1$ rows of $D(G)$ are equal, they can be collapsed in $\mathrm{PDD}(G)$ to a single row with the *weight* k/m. The PDD was defined for clouds as a local distribution of distances in Definition 5.5 of (Mémoli, 2011) and for periodic sets in (Widdowson & Kurlin, 2022) but not for Euclidean graphs.

Table 1. Acronyms of all main invariants and metrics in the paper.

SRD	Sorted Radial Vector	Def 3.1
SPD	Sorted Distance Vector	Def 3.1
PDD	Pointwise Distance Distribution	Def 3.1
CR	Centered Representation	Def 3.3
NCD	Nested Centered Distribution	Def 3.5
NBM	Nested Bottleneck Metric	Def 4.5

The $\mathrm{PDD}(G)$ includes every signed distance twice, once as $d(p,q)$ in the row of a vertex p, and as $d(q,p)$ in the row of a vertex q. Hence $\mathrm{SPD}(G)$ can be obtained from $\mathrm{PDD}(G)$ by (1) combining all distances into one vector, (2) sorting them in decreasing order, and (3) keeping only one copy of every two repeated distances. Example 3.2 shows that the invariant $\mathrm{PDD}(G)$ is strictly stronger than $\mathrm{SPD}(G)$.

Example 3.2 (invariants $\mathrm{SRD}, \mathrm{SPD}, \mathrm{PDD}$ for tetrahedral graphs in Fig. 1)**.** ***(a)*** *Since the vertex sets of $T_i \subset \mathbb{R}^3$ are regular tetrahedra with all pairwise distances 1, these graphs have identical $\mathrm{SRD}(T_i)$ of 4 equal circumradii of the same vertex set $V(T_i)$ independent of $i = 1, \dots, 4$.*

The first graph T_1 has two edges contributing $+1$ and four non-edges (dashed lines) contributing -1 to the Sorted Distance Vector $\mathrm{SPD}(T_1) = (+1,+1,-1,-1,-1,-1)$. The graph T_2 also has two edges, so $\mathrm{SPD}(T_2) = \mathrm{SPD}(T_1)$ doesn't distinguish $T_1 \not\cong T_2$ up to rigid motion. Similarly, the graphs $T_3 \not\cong T_4$ are not distinguished by the invariants $\mathrm{SPD}(T_3) = (+1,+1,+1,-1,-1,-1) = \mathrm{SPD}(T_4)$.

(b) *In T_1, every vertex has exactly one edge and two non-edges (dashed lines), hence its signed distances are $+1,-1,-1$. The matrix $\mathrm{PDD}(T_1) = (100\% \mid -1,-1,+1)$ consists of a single row, where the weight 100% indicates that all vertices of T_1 have the same row in* PDD*. The graph T_2 has one vertex (25%) with no edges, two vertices (50%) with one edge, and one vertex (25%) with two edges, so* $\mathrm{PDD}(T_2) = \left(\begin{array}{c|ccc} 25\% & -1 & -1 & -1 \\ 50\% & -1 & -1 & +1 \\ 25\% & -1 & +1 & +1 \end{array}\right) \neq \mathrm{PDD}(T_1)$, *so* PDD *distinguishes the rigidly non-equivalent graphs $T_1 \not\cong T_2$ with $\mathrm{SPD}(T_1) = \mathrm{SPD}(T_2)$. The graph T_3 has one vertex (25%) with no edges and three vertices (75%) with two edges, so* $\mathrm{PDD}(T_3) = \left(\begin{array}{c|ccc} 25\% & -1 & -1 & -1 \\ 75\% & -1 & +1 & +1 \end{array}\right)$. *The graph T_4 has two vertices (50%) with one edge and two vertices (50%) with two edges. Then* $\mathrm{PDD}(T_4) = \left(\begin{array}{c|ccc} 50\% & -1 & -1 & +1 \\ 50\% & -1 & +1 & +1 \end{array}\right)$, *so* PDD *distinguishes the graphs $T_3 \not\cong T_4$ with equal $\mathrm{SPD}(T_3) = \mathrm{SPD}(T_4)$.*

For a graph G with m unordered vertices, $\mathrm{PDD}(G)$ has $m-1$ columns. The reduced version $\mathrm{PDD}(G;k)$ includes only the first k columns for $1 \le k < m-1$. Though PDDs have unordered rows, they can be continuously compared by Earth Mover's Distance (Rubner et al., 2000).

Fig. S4 in (Pozdnyakov et al., 2020) described infinitely many non-isometric pairs of clouds $C, C' \subset \mathbb{R}^3$ with $\mathrm{PDD}(C) = \mathrm{PDD}(C')$. These counter-examples inspired the stronger invariants for graphs below. For simplicity, we will introduce all invariants and metrics in dimension $n = 2$. All higher dimensions $n > 2$ are covered in appendices. While $\mathrm{PDD}(G)$ includes signed distances to a single (arbitrary) vertex $p_i \in V(G)$, a stronger invariant below include triples of signed distances to three base points, one of which is the center of mass of $V(G)$ because any point in $\mathbb{R}^2$ is uniquely determined by its distances to three fixed points.

Definition 3.3 (Centered Representation $\mathrm{CR}(G;A)$ of a graph with $A \subset V(G)$)**.** *Let $G \subset \mathbb{R}^2$ be a graph on m unordered points with the center of mass $p_0 = O(G) = 0$.*

(a) *For any vertex $p_1 \in V(G)$, the matrix $R(G;p_1)$ has $m-1$ unordered columns, one for each vertex $q \in V(G) \setminus \{p_1\}$, consisting of the signed distances $d(q,p_0)$ and $d(q,p_1)$. Here $p_0 = 0$ is not considered as a vertex of G, so $d(q,p_0) = -|q|$. The* Centered Representation $\mathrm{CR}(G;p_1)$ *is the pair $[d(p_0,p_1), R(G;p_1)]$, where $d(p_0,p_1) = -|p_1|$.*

(b) *Fix a* base pair *A of ordered vertices $p_1, p_2 \in V(G)$. Let $\mathrm{sign}(A)$ be the sign of the 2×2 determinant on the vectors p_1, p_2. Let $D(A)$ be the matrix of signed distances between p_0, p_1, p_2. The matrix $R(G;A)$ has $m-2$ unordered columns, one for each vertex $q \in V(G) \setminus A$, consisting of signed distances $d(q,p_0), d(q,p_1), d(q,p_2)$. The* Centered Representation $\mathrm{CR}(G;A)$ *is the triple $[\mathrm{sign}(A), D(A), R(G;A)]$.*

After fixing $p_0 = 0$, the matrix $D(A)$ and $\mathrm{sign}(A)$ help reconstruct base vertices $p_1, p_2 \in \mathbb{R}^2$, uniquely under rotation around 0. Any other $q \in V(G) \setminus A$ is fixed relative to p_0, p_1, p_2 by its column in $R(G;A)$. A positive sign of $d(p_i,p_j)$ indicates an edge between vertices p_i, p_j. This argument will later be formalized in Theorem 4.6(b).

Example 3.4 (CRs for 2-vertex bases in $\mathbb{R}^2$)**.** *Let $G \subset \mathbb{R}^2$ be the triangular cycle on $p_1 = (2,0)$, $p_2 = (-1,1)$, $p_3 = (-1,-1)$, so $O(G) = 0$ and all signed distances are positive, see Fig. 2 (top left). For $A = (p_1,p_2)$,* $\mathrm{sign}(A) = \mathrm{sign}\begin{vmatrix} 2 & -1 \\ 0 & 1 \end{vmatrix} = 1$. *The distance matrix on $0, p_1, p_2$ is* $D(p_1,p_2) = \begin{pmatrix} 0 & -2 & -\sqrt{2} \\ -2 & 0 & \sqrt{10} \\ -\sqrt{2} & \sqrt{10} & 0 \end{pmatrix}$. *Then* $R(G;p_1,p_2) = \begin{pmatrix} -|p_3| \\ |p_3 - p_1| \\ |p_3 - p_2| \end{pmatrix} = \begin{pmatrix} -\sqrt{2} \\ \sqrt{10} \\ 2 \end{pmatrix}$. *Then* $\mathrm{CR}(G;p_1,p_2) = [+1, D(p_1,p_2), R(G;p_1,p_2)]$. *Replacing p_2 with p_3, we find* $\mathrm{sign}(p_1,p_3) = \mathrm{sign}\begin{vmatrix} 2 & -1 \\ 0 & -1 \end{vmatrix} = -1$, $D(p_1,p_3) = \begin{pmatrix} 0 & -2 & -\sqrt{2} \\ -2 & 0 & \sqrt{10} \\ -\sqrt{2} & \sqrt{10} & 0 \end{pmatrix}$, *and* $R(G;p_1,p_3) = \begin{pmatrix} -|p_2| \\ |p_2 - p_1| \\ |p_2 - p_3| \end{pmatrix} = \begin{pmatrix} -\sqrt{2} \\ \sqrt{10} \\ 2 \end{pmatrix}$. *The final triple is* $\mathrm{CR}(G;p_1,p_3) = [-1, D(p_1,p_3), R(G;p_1,p_3)]$.

Though a Centered Representation $\mathrm{CR}(G;p_1,p_2)$ will suffice to reconstruct $G \subset \mathbb{R}^2$ uniquely under rigid motion

in Theorem 4.6(b), $\mathrm{CR}(G;p_1,p_2)$ for all vertices $p_1,p_2 \in V(G)$ should be considered in a joint unordered collection below to guarantee the independence of points p_1,p_2.

Definition 3.5 (Nested Centered Distribution $\mathrm{NCD}(G;h)$)**.** *Let $G \subset \mathbb{R}^2$ be any Euclidean graph with m unordered vertices and the center of mass at the origin $0 \in \mathbb{R}^n$.*

(a) *The* Nested Centered Distribution $\mathrm{NCD}(G;1)$ *of order 1 is the unordered set of Centered Representations* $\mathrm{CR}(G;p_1)$ *from Definition 3.3 for all vertices $p_1 \in V(G)$.*

(b) *For any vertex $p_1 \in V(G)$, the* Centered Distribution $\mathrm{CD}_1(G;p_1)$ *is the unordered set of* $\mathrm{CR}(G;p_1,p_2)$ *for all $p_2 \in V(G) \setminus \{p_1\}$. The* Nested Centered Distribution $\mathrm{NCD}(G;2)$ *of order 2 is the unordered set of* $\mathrm{CD}_1(G;p_1)$ *for all vertices $p_1 \in V(G)$, see Fig. 2 (top). The* mirror image $\overline{\mathrm{NCD}}(G;2)$ *is obtained from* $\mathrm{NCD}(G;2)$ *by reversing* $\mathrm{sign}(p_1,p_2)$ *of 2×2 determinants in all* $\mathrm{CR}(G;p_1,p_2)$.

The nested structure of $\mathrm{NCD}(G;2)$ helps identify edges between all vertices from G. After any vertex $q \in V(G) \setminus \{p_1,p_2\}$ is uniquely located by using one $\mathrm{CR}(G;p_1,p_2)$, we can use unsigned distances to associate any such q with its unique $\mathrm{CR}(G;p_1,q)$ in the collection $\{\mathrm{CR}(G;p_1,p_i)\}$ for all $p_i \in V(G) \setminus \{p_1\}$. The resulting $\mathrm{CR}(G;p_1,q)$ contains signed distances and hence detects edges from q to all other vertices, see details in the proof of Theorem 4.6(b).

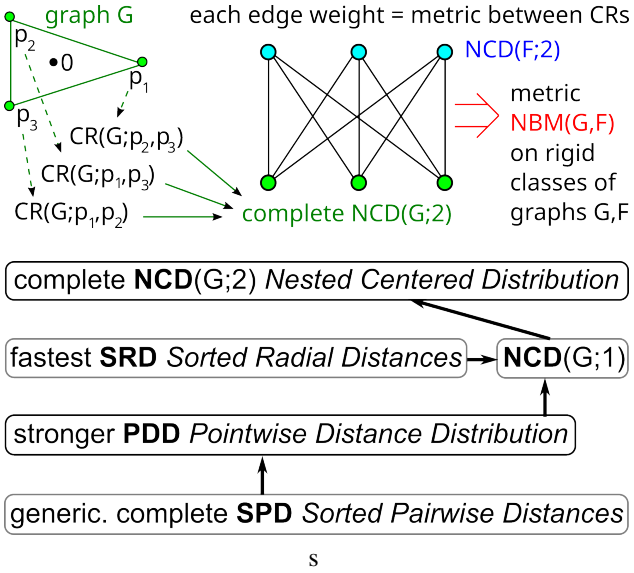

Figure 2. **Top**: building the Nested Centered Distribution NCD in Definition 3.5 from Centered Representations in Definition 3.3 with metrics in section 4. **Bottom**: hierarchy of graph invariants.

$\mathrm{SRD}(G)$ can be considered $\mathrm{NCD}(G;0)$ of order 0, containing signed distances from the center of mass p_0 to all vertices of G, additionally written in increasing order.

4. Continuous metrics on graph invariants

When points $0 \cup A = (p_0,p_1,p_2) \subset \mathbb{R}^2$ pass through a degenerate configuration in a straight line, i.e. p_1,p_2 become collinear, $\mathrm{sign}(A)$ discontinuously changes. To guarantee the Lipschitz continuity, we multiply such a sign by the strength σ below, which smooths the sign change, while the area of the triangle on p_0,p_1,p_2 is not Lipschitz continuous.

Definition 4.1 (*strength* $\sigma(C)$)**.** *Any triple $C = \{p_0,p_1,p_2\} \subset \mathbb{R}^2$ defines a triangle with inter-point distances a,b,c, and half-perimeter $p = \frac{1}{2}(a+b+c)$. The* strength *is* $\sigma(C) = \dfrac{(p-a)(p-b)(p-c)}{p^2}$.

Lemma 4.2 (Theorem 4.4 in (Widdowson & Kurlin, 2023))**.** *Let B be obtained from a set $C \subset \mathbb{R}^2$ of 3 points by perturbing every point within its ε-neighborhood. Then $|\sigma(B)-\sigma(C)| \le 2\varepsilon\lambda_2$ for $\lambda_2 = 2\sqrt{3}$.*

The strength $\sigma(A)$ will be normalized by λ_2 below to guarantee the final Lipschitz constant 2 for a metric in Theorem 4.6(c). For any $k\times k$ matrices M,N of real numbers, the metric L_∞ is $\max\limits_{i,j=1,\dots,k} |M_{ij}-N_{ij}|$. The *bottleneck* distance between any clouds A,B of (the same number of) m unordered points in a metric space with a distance d is $W_\infty(A,B) = \min\limits_{\text{bijections } g:A\to B} \max\limits_{p\in A} d(g(p),p)$.

Definition 4.3 (max metric M_∞ on CRs)**.** *Let Euclidean graphs $G,F \subset \mathbb{R}^n$ have m unordered vertices.*

(a) *For order $h=1$, take any base vertices $p \in V(G)$ and $q \in V(F)$. Define the* max metric $M_\infty(\mathrm{CR}(G;p),\mathrm{CR}(F;q)$ *as the maximum of $||p|-|q||$ and the bottleneck distance W_∞ between the fixed clouds of unordered points $\{(-|p'|,d(p',p)) \mid p' \in V(G)-\{p\}\}$ and $\{(-|q'|,d(q',q)) \mid q' \in V(F)-\{q\}\}$ in $\mathbb{R}^2$.*

(b) *For order $h=2$, take any base sequences $A \subset V(G)$ and $B \subset V(F)$ of two vertices. Consider the $m-2$ columns of $R(G;A)$ from Definition 3.3 as a cloud of $m-2$ unordered points in $\mathbb{R}^2$, also for $R(F;B)$. The* max metric $M_\infty(\mathrm{CR}(G;A),\mathrm{CR}(F;B))$ *is the maximum of* $\dfrac{2}{\lambda_2}|\mathrm{sign}(A)\sigma(0\cup A) - \mathrm{sign}(B)\sigma(0\cup B)|$, $L_\infty(D(A),D(B))$, *and* $W_\infty(R(G;A),R(F;B))$.

The maximum of several distances in Definition 4.3 is needed to guarantee the first metric axiom, i.e. $M_\infty(\mathrm{CR}(G;A),\mathrm{CR}(F;B)) = 0$ should imply that $0\cup A$ should be exactly matched by rotation with $0\cup B$ and then $\mathrm{CR}(G;A) = \mathrm{CR}(F;B)$ up to a permutation of columns will imply that G coincides with F, see Lemma D.7.

To get a metric on Nested Centered Distributions, we will use the distance on bipartite graphs whose edge weights are the max metrics M_∞ on Centered Representations.

Definition 4.4 (Bottleneck Matching Distance $\mathrm{BMD}(\Gamma)$)**.**

Let Γ be a complete bipartite graph with m white vertices and m black vertices so that every white vertex is connected to every black vertex by a single edge e of a weight $w(e) \geq 0$. A vertex matching *of the graph Γ is a collection E of m disjoint edges with $2m$ distinct vertices. The* weight $W(E) = \max\limits_{e \in E} w(e)$ *is the largest weight of an edge in E. The* Bottleneck Matching Distance $\mathrm{BMD}(\Gamma) = \min\limits_{E} W(E)$ *is the minimum weight of a vertex matching E of Γ.*

Since a graph Γ is complete bipartite, any edge from a vertex matching E in Γ joins a white vertex with a black vertex. Then $\mathrm{BMD}(\Gamma)$ is minimized for all bijections E between all white vertices and all black vertices of Γ.

Definition 4.5 (Nested Bottleneck Metric NBM on NCDs)**.** *Let $G, F \subset \mathbb{R}^2$ be any graphs on m unordered vertices.*

(a) *For order $h = 1$, the* Nested Bottleneck Metric $\mathrm{NBM}(\mathrm{NCD}(G;1), \mathrm{NCD}(F;1))$ *is the max metric $M_\infty(\mathrm{CR}(G;p), \mathrm{CR}(F;\beta(p)))$ minimized for all bijections $\beta : V(G) \to V(F)$ between vertices of G and F.*

(b) *For order $h = 2$, any base vertices $p_1 \in V(G)$ and $q_1 \in V(F)$, let the complete bipartite graph $\Gamma(G;p_1;F;q_1)$ have $m-1$ white vertices and $m-1$ black vertices representing $\mathrm{CR}(G;p_1,p_2)$ and $\mathrm{CR}(F;q_1,q_2)$ for all $p_2 \in V(G) - \{p_1\}$ and $q_2 \in V(F) - \{q_1\}$, respectively. Set the* weight $w(e)$ *of an edge e joining the vertices represented by $\mathrm{CR}(G;p_1,p_2)$ and $\mathrm{CR}(F;q_1,q_2)$ as the max metric M_∞ between these distributions, see Definition 4.3. Then Definition 4.4 gives the bottleneck matching distance $\mathrm{BMD}(\Gamma(G;p_1;F;q_1))$.*

Let the complete bipartite graph $\Gamma(G,F)$ have weight $\mathrm{BMD}(\Gamma(G;p_1;F;q_1))$ on each edge connecting vertices representing $p_1 \in V(G)$ and $q_1 \in V(F)$. The Nested Bottleneck Metric $\mathrm{NBM}(\mathrm{NCD}(G;2), \mathrm{NCD}(F;2))$ *is the Bottleneck Matching Distance $\mathrm{BMD}(\Gamma(G,F))$.*

$\mathrm{SRD}(G)$ coincides with $\mathrm{NCD}(G;0)$ after sorting, so NBM can be defined as $L_\infty(\mathrm{SRD}(G), \mathrm{SRD}(F))$ for order $h = 0$. The metrics $W_\infty, M_\infty, \mathrm{NBM}$ compare objects of the same size. To compare graphs with different numbers of vertices, M_∞ in Definition 4.5 can be replaced with Earth Mover's Distance EMD in Definition C.2. All metric axioms and main Theorem 4.6 below are proved in appendices C and D for any dimension $n \geq 2$ and orders $1 \leq h \leq n$.

Theorem 4.6 (NCD solves Problem 1.1)**.** ***(a)*** *The Nested Centered Distribution $\mathrm{NCD}(G;h)$ in Definition 3.5 is invariant under any rigid motion for all Euclidean graph G on m unordered vertices and can be computed in time $O(n^2 m^{h+1})$ with space $O(n^3 + hm^{h+1})$ for $h \leq n = 2$.*

(b) *$\mathrm{NCD}(G;2)$ is a complete invariant of all graphs $G \subset \mathbb{R}^2$ under rigid motion from the group $\mathrm{SE}(2)$.*

(c) *Perturbing each vertex of a graph $G \subset \mathbb{R}^2$ within its ε-neighborhood changes $\mathrm{NCD}(G;h)$ up to 2ε in both metrics* NBM *and* EMD *for any order $h = 1, 2$.*

(d) *For any graphs $G, F \subset \mathbb{R}^2$ on m unordered vertices, the metrics* NBM *and* EMD *between the invariants $\mathrm{NCD}(G;h)$ and $\mathrm{NCD}(F;h)$ is computed in time $O(m^{2h+1.5} \log^{h+1} m)$ with space $O(m^{2h+1} \log^{h-1} m)$ for $h \leq n = 2$.*

Theorem 4.6(b) implies that any graphs $G, F \subset \mathbb{R}^2$ are related by rigid motion *if and only if* $\mathrm{NCD}(G;2) = \mathrm{NCD}(F;2)$. This equality is interpreted as a bijection $\mathrm{NCD}(G;n) \to \mathrm{NCD}(F;n)$ matching all CRs, which is equivalent to $\mathrm{NBM} = 0$ by the first metric axiom. Since every CR can be stored in a vector form, the complete invariant $\mathrm{NCD}(G;2)$ for $n = 2$ can be considered vectorial.

Table 2 emphasizes that most graphs should be first compared (or represented for machine learning) by simpler and faster invariants, so the complete $\mathrm{NCD}(G;n)$ is used only in rare cases but is still needed to distinguish all graphs.

Table 2. Invariants and metrics on graphs $G \subset \mathbb{R}^2$ with m unordered vertices: from the fastest (linear-time) to complete.

INVARIANT	TIME	METRIC	TIME
$\mathrm{SRD}(G)$	$O(m \log m)$	L_∞	$O(m)$
$\mathrm{SPD}(G)$	$O(m^2)$	L_∞	$O(m^2)$
$\mathrm{PDD}(G)$	$O(m^2 \log m)$	EMD	$O(m^3)$
$\mathrm{NCD}(G;1)$	$O(m^2)$	NBM	$O(m^{3.5} \log^2 m)$
$\mathrm{NCD}(G;2)$	$O(m^3)$	NBM	$O(m^{5.5} \log^3 m)$

Example 4.7 (version of Theorem 4.6(b) for $n = 1$)**.** *For a graph $G \subset \mathbb{R}$ with the center of mass $O(G) = 0$, take any base vertex $p \in G$. Then $\mathrm{sign}(p)$ is the usual sign of $p \in \mathbb{R}$, $D(p)$ is the signed distance $-|p|$, $R(G;p)$ is the $2 \times (m-1)$ matrix whose column for any vertex $q \in V(G) - \{p\}$ consists of the signed distances $d(q,0) = -|q|$ and $d(q,p) = \pm|q-p|$, where the plus sign $+$ indicates an edge between q, p, while the minus sign $-$ means no edge.*

For order $h = 1$, the Centered Representation is the pair $\mathrm{CR}(G;p) = [\mathrm{sign}(p), -|p|, R(G;p)]$. The base vertex p is fixed in the line $\mathbb{R}$ by $\mathrm{sign}(p)$ and $|p|$. Any other vertex $q \in V(G) - \{p\}$ is uniquely determined in $\mathbb{R}$ by its Euclidean distances $|q|$, $|q-p|$ to the origin and the already fixed p. The location of any point $q \in \mathbb{R}$ is characterized by $\mathrm{sign}(q)$ and $|q|$, which helps unambiguously identify its Centered Representation $\mathrm{CR}(G;q)$ in the unordered collection $\mathrm{NCD}(G;1)$ of all these CRs*. The signs of $d(q,q')$ in each $R(G;q)$ determine the presence or absence of an edge of $G \subset \mathbb{R}$ between any vertices $q, q' \in V(G)$.*

5. Experiments on largest molecular databases

The world's largest databases of 3D molecular geometry are QM9 (130K+ entries) (Ramakrishnan et al., 2014) and GD (GEOM_drugs of 31M+ entries) (Axelrod & Gomez-

Bombarelli, 2022), which have hundreds of 3D conformers of *unordered* atoms for each of 621 and 61607 chemical compositions, respectively. The Protein Data Bank has backbones of *ordered* atoms classified by simpler invariants (Anosova et al., 2025). All experiments took a few hours on Ryzen 9 3950X 3.5 GHz, 64 MB of L3 cache, RAM 82GB.

The ICML guide for reviewing application-driven ML says that "novel ideas that are simple to apply may be especially valuable". To demonstrate the chemical importance of the linear-time invariant SRD, we extracted clouds of $k = 10$ neighbors around every atom, see their counts in Table 3.

Table 3. Counts of atoms by chemical elements in QM9 (2,407,753 atoms), GD0 (GEOM_drugs 0th conformers, 12,917,980 atoms).

QM9: H	QM9: C	QM9: N	QM9: O	QM9: F
1,230,122	846,557	139,764	187,996	3,314
GD0: H	GD0: C	GD0: N	GD0: O	GD0: F
5,660,986	5,267,096	842,562	854,400	64,299
GD0: P	GD0: S	GD0: Cl	GD0: Br	GD0: I
1,350	159,648	53,404	14,010	225

Though the data was skewed towards more popular elements H (hydrogen) and C (carbon), a default network in TensorFlow with 80/20 split for train/test achieved over 98% accuracy in predictions of the chemical element of a central atom by distances to only $k = 3$ nearest neighbours, see Table 4. Appendix A has all implementation details.

Table 4. Accuracies in percentages for predicting the chemical element of a central atom by a 4-layer network using *only the* k *shortest distances* to atomic neighbors within a molecular graph.

data	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
QM9	94.63	98.64	98.24	98.54	98.77

In chemistry, both ML and non-ML predictions of elements achieved only 86% on similar size data, see Table 7 summarized in (Vasylenko et al., 2025), because the underlying descriptors were not invariant, e.g. under permutations of atoms, which creates exponentially many representations of the same molecule, incomplete, or their similarities failed the triangle axiom, e.g. see (Steck et al., 2024).

High accuracies in Table 4 are rigorously explained by the cascade comparisons on all atomic clouds (environments) from QM9. Split all clouds from by the 1st distance (to the nearest neighbor of a central atom p) rounded to 3 decimal places in Å. This is a typical experimental precision, where $1\text{Å} = 10^{-10}m$ is approximately the smallest interatomic distance. Second, split each subset with equal 1st distances by 2nd distances, and so on up to $k = 5$ distances. All 2.4M+ atomic clouds of different elements in QM9 were separated by the shorest distances to only 4 atomic neighbors.

The hierarchy of invariants in Fig. 2 and Table 2 transparently explained the reconstruction of chemical elements from distances to k nearest neghbors and inspired the harder task to reconstruct a chemical composition from a molecule-level (not atomwise) invariant of only atomic centers.

For molecular graphs from QM9, we computed the pseudometric L_∞ (max absolute difference of corresponding coordinates) on all 873,527,974 pairs of SRDs, then 8,735,279 distances L_∞ on the stronger SPDs for the 1% closest pairs, then 87,352 EMDs on PDDs for the 1% closest pairs, distances NBM on $\text{NCD}(G;1)$ and $\text{NCD}(G;2)$ for the top 10K closest pairs, and 64 NBMs on complete $\text{NCD}(G,3)$.

The invariants in Table 5 distinguish all chemically different molecules with NBM on complete invariants giving the largest separation. All chemical compositions in QM9 and GD were distinguished by the vector SRD of Euclidean distances (rounded to 3 decimal places in Å) from the molecular center of mass to 5 and 7 farthest atoms, respectively.

This transparent reconstruction of the full chemistry from precise enough atomic geometry gives hope to rigorously infer other molecular properties from geometric invariants.

Table 5. Chemically different molecules (given by QM9 ids) are geometrically distinguished by invariant metrics, see Fig. 3 (right).

smallest distances in Å, molecule A $\neq$ molecule B
SRD, $L_\infty = 0.021$, $H_3C_4N_3O_2$(131923)$\neq$$H_4C_5N_2O$(5365)
SPD, $L_\infty = 0.055$, $H_3C_4N_5$(123533)$\neq$$H_3C_5N_3O$(24547)
PDD, EMD $= 0.051$, $H_3C_4N_5$(123533)$\neq$$H_3C_5N_3O$(24521)
NCD, NBM $= 0.071$, $H_4C_5N_4$(123532)$\neq$$H_4C_6N_2O$(24513)

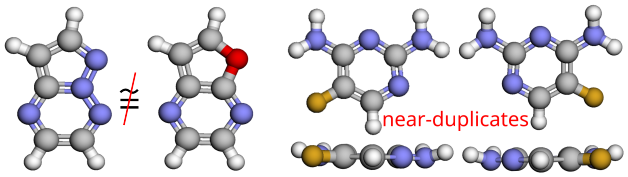

Figure 3. **Left**: the smallest NBM ≈ 0.07Å on $\text{NCD}(G;3)$ for chemically different molecules 123533 and 24521. **Right**: near-duplicate (almost flat) molecules 123532 and 24513 have the same composition and tiny EMD $\approx 2.37 \times 10^{-7}$Å (not distinguishing mirror images) but a $100\times$ higher NBM $\approx 2.95 \times 10^{-5}$Å.

For QM9 molecul graphs, Fig. 4 and 9 NBM distances for different NCD invariants of orders $h = 1, 2, 3$.

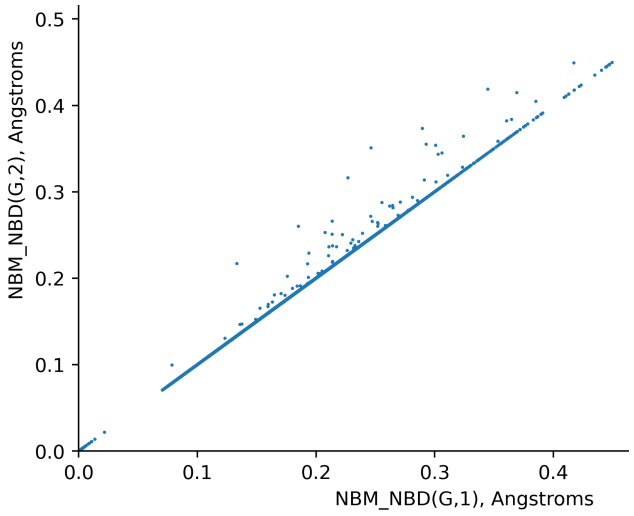

Figure 4. Each dot is a comparison of molecular graphs from QM9: $x = \mathrm{NBM}$ on $\mathrm{NCD}(G;1)$ vs $y = \mathrm{NBM}$ on $\mathrm{NCD}(G;2)$.

6. Discussion: conclusions and limitations

The comparisons of molecular graphs from QM9 and GD imply that all chemically different molecules are rigidly different, see the smallest distance $\mathrm{NBM} \approx 0.07$Å on complete invariants in Table 5. So the map $\{\text{molecules}\} \to \{\text{graphs on atomic centers (without chemical elements)}\}$ is injective on rigid classes and can be inverted on its image.

Hence the most important property (chemical composition) is reconstructable from precise enough geometry. Using only a few radial distances (5 at the atomic level and 7 at the molecular level, rounded to 3 decimal places) for uniquely identifying all chemical elements in QM9 and GD demonstrates the transparency of application-driven ML.

The solution to Problem 1.1 settled the long-standing challenge of properly defining a *molecular structure*. A traditional approach is to describe such a structure as "a set of unlabeled configurations that are relatively similar to each other", quoted from the paragraph to the left of the caption of Fig. 1 in (Lang et al., 2024). If this 'similarity' is treated as an equivalence allowing perturbations of atoms up to a positive threshold, sufficiently many perturbations can make all molecules (of the same number of atoms) equivalent by the transitivity axiom. A justified way to resolve this paradox is to embrace uncertainty and continuously quantify this similarity not by ignoring any perturbations up to a threshold but by computing an exact distance satisfying all metric axioms and Lipschitz continuity in Problem 1.1.

The question of whether to put close neighbors like near-duplicates in Fig. 3 (left) into one cluster of the "same" molecules is rather administrative similar to assigning close houses to one village (cluster) instead of different ones.

Studying molecules by fixing a composition is similar to drawing artificial boundaries between countries on Earth. Because some molecules of different compositions have close shapes as in Fig. 5, they should have similar properties. Now any properties of molecules should be possible to predict only from the complete invariant $\mathrm{NCD}(G;3)$ even without chemistry in the same way as any precise geographic location uniquely determines all physical properties of this place such as the average annual temperature. Chemical compositions can be still helpful similar to the location's altitude, which easier predicts (say) the average temperature than theoretically sufficient geographic coordinates.

Any vertex p and edge of G can have an *attribute* and a *weight* respected by any isometry that maps one graph to another. These vertex attributes and edge weights can be incorporated as extra columns and rows in CRs from Definition 3.3, and then incorporated into NCD and NBM. We can compare graphs of different numbers of vertices because EMD works for both PDD and NCD as weighted distributions of any finite size. This comparison splits the vertices from $V(G)$ into parts (subvertices) that are optimally 'transported' to a splitting of another vertex set $V(F)$.

The main contribution is Theorem 4.6 and its extension in Theorem D.1 to all dimensions $n \geq 2$ fully solving Problem 1.1. The limitation is the time $O(n^2 m^{n+1})$ of the complete invariant $\mathrm{NCD}(G;n)$ of any graphs $G \subset \mathbb{R}^n$. For a fixed dimension n, this polynomial complexity resolves two exponential-size challenges: $m!$ permutations of m unordered vertices and up to $2^{m(m-1)/2}$ non-isometric graphs with up to $m(m-1)/2$ edges on m fixed vertices in $\mathbb{R}^n$.

In practice, all comparisons and property predictions can start from much faster (linear-time) invariants SRD and only in cases of close distances (potential confusions) progress to stronger invariants $\mathrm{SPD}, \mathrm{PDD}, \mathrm{NCD}$. This hierarchical (cascade) computation can better address the curse of dimensionality instead of the one-size-fits-all approach.

A map f : objects $\to$ descriptors $\to$ properties is invertible only if objects are faithfully represented by complete invariants. Any non-invariant maps a single object to (usually infinitely) many values or representations. Any incomplete invariant can fail to differentiate between objects with different properties. Hence a *generative* approach (inverting f above) can succeed only after the *discriminative* problem is solved. The space $\mathrm{GRS}(\mathbb{R}^3;m)$ of rigid classes of all graphs on m vertices in $\mathbb{R}^3$ contains all possible shapes of molecules (all already known and also all *not yet discovered* ones). The complete invariant $\mathrm{NCD}(G)$ of $G \subset \mathbb{R}^3$ defines geographic-style coordinates on a continuous map of $\mathrm{GRS}(\mathbb{R}^3;m)$ containing QM9 and GD. Since the space $\mathrm{GRS}(\mathbb{R}^3;m)$ is high-dimensional, we really need complete invariants to separate all known molecules and look for unexplored gaps containing new future molecules.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abbe, E. and Sandon, C. On the universality of deep learning. *Advances in Neural Information Processing Systems*, 33:20061–20072, 2020.

Agarwal, P. K., Fox, K., Nath, A., Sidiropoulos, A., and Wang, Y. Computing the gromov-hausdorff distance for metric trees. *ACM Transactions on Algorithms*, 14(2): 1–20, 2018.

Akhtar, N. and Mian, A. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018.

Amir, T., Gortler, S., Avni, I., Ravina, R., and Dym, N. Neural injective functions for multisets, measures and graphs via a finite witness theorem. *Advances in Neural Information Processing Systems*, 36, 2024.

Anosova, O., Gorelov, A., Jeffcott, W., Jiang, Z., and Kurlin, V. A complete and bi-continuous invariant of protein backbones under rigid motion. *MATCH Communications in Mathematical and in Computer Chemistry (to appear), arxiv:2410.08203*, 2025.

Antunes, L. M., Grau-Crespo, R., and Butler, K. T. Distributed representations of atoms and materials for machine learning. *npj Computational Materials*, 8(1):44, 2022.

Arvind, V. and Rattan, G. The parameterized complexity of geometric graph isomorphism. *Algorithmica*, 75:258–276, 2016.

Axelrod, S. and Gomez-Bombarelli, R. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022.

Babai, L. Canonical form for graphs in quasipolynomial time: preliminary report. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1237–1246, 2019.

Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J. P., Kornbluth, M., Molinari, N., Smidt, T. E., and Kozinsky, B. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453, 2022.

Bonchev, D. *Chemical graph theory: introduction and fundamentals*, volume 1. CRC Press, 1991.

Brécheteau, C. A statistical test of isomorphism between metric-measure spaces using the distance-to-a-measure signature. *Electronic J Statistics*, 13:795–849, 2019.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. Geometric deep learning: going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34 (4):18–42, 2017.

Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P. Geometric deep learning: grids, groups, graphs, geodesics, and gauges. *arXiv:2104.13478*, 2021.

Buchin, M., Chambers, E., Fang, P., Fasy, B. T., Gasparovic, E., Munch, E., and Wenk, C. Distances between immersed graphs: Metric properties. *La Matematica*, pp. 1–26, 2023.

Bunch, J. R. and Hopcroft, J. E. Triangular factorization and inversion by fast matrix multiplication. *Mathematics of Computation*, 28(125):231–236, 1974.

Chen, C., Ye, W., Zuo, Y., Zheng, C., and Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31(9): 3564–3572, 2019.

Chew, P. and Kedem, K. Improvements on geometric pattern matching problems. In *Scandinavian Workshop on Algorithm Theory*, pp. 318–325, 1992.

Chew, P., Dor, D., Efrat, A., and Kedem, K. Geometric pattern matching in d-dimensional space. *Discrete & Computational Geometry*, 21(2):257–274, 1999.

Choo, H. Y., Wee, J., Shen, C., and Xia, K. Fingerprint-enhanced graph attention network (fingat) model for antibiotic discovery. *Journal of Chemical Information and Modeling*, 2023.

Cohen, T. S., Geiger, M., and Weiler, M. A general theory of equivariant cnns on homogeneous spaces. *Advances in neural information processing systems*, 32, 2019.

Colbrook, M. J., Antun, V., and Hansen, A. C. The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smale's 18th problem. *Proc. National Academy of Sciences*, 119(12):e2107151119, 2022.

Dekster, B. V. and Wilker, J. B. Edge lengths guaranteed to form a simplex. *Archiv der Mathematik*, 49(4):351–366, 1987.

Delle Rose, V., Kozachinskiy, A., Rojas, C., Petrache, M., and Barceló, P. Three iterations of (d- 1)-wl test distinguish non isometric clouds of d-dimensional points. *Advances in Neural Information Processing Systems*, 36, 2024.

Deng, C., Litany, O., Duan, Y., Poulenard, A., Tagliasacchi, A., and Guibas, L. J. Vector neurons: A general framework for so(3)-equivariant networks. In *Proceedings of the International Conference on Computer Vision*, pp. 12200–12209, 2021.

Deza, E. and Deza, M. M. *Encyclopedia of distances*. Springer, 2009.

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., and Li, J. Boosting adversarial attacks with momentum. In *Computer vision and pattern recognition*, pp. 9185–9193, 2018.

Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R. P. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28, 2015.

Dym, N. and Gortler, S. J. Low-dimensional invariant embeddings for universal geometric learning. *Foundations of Computational Mathematics*, pp. 1–41, 2024.

Efrat, A., Itai, A., and Katz, M. J. Geometry helps in bottleneck matching and related problems. *Algorithmica*, 31(1):1–28, 2001.

Feynman, R. *The Feynman lectures on physics. Chapter 1: atoms in motion*, volume 1. 1971.

Fisikopoulos, V. and Penaranda, L. Faster geometric algorithms via dynamic determinant computation. *Computational Geometry*, 54:1–16, 2016.

Fredman, M. L. and Tarjan, R. E. Fibonacci heaps and their uses in improved network optimization algorithms. *Journal of the ACM*, 34(3):596–615, 1987.

Fuchs, F., Worrall, D., Fischer, V., and Welling, M. Se(3)-transformers: 3d roto-translation equivariant attention networks. *Advances in neural information processing systems*, 33:1970–1981, 2020.

Gao, X., Hu, W., and Qi, G.-J. Graphter: Unsupervised learning of graph transformation equivariant representations via auto-encoding node-wise transformations. In *Proceedings of Computer Vision and Pattern Recognition*, pp. 7163–7172, 2020.

Goldberg, A. and Tarjan, R. Solving minimum-cost flow problems by successive approximation. In *Proceedings of STOC*, pp. 7–18, 1987.

Goodrich, M. T., Mitchell, J. S., and Orletsky, M. W. Approximate geometric pattern matching under rigid motions. *Transactions on Pattern Analysis and Machine Intelligence*, 21(4):371–379, 1999.

Goyal, A., Law, H., Liu, B., Newell, A., and Deng, J. Revisiting point cloud shape classification with a simple and effective baseline. In *International Conference on Machine Learning*, pp. 3809–3820, 2021.

Grinberg, D. and Olver, P. J. The n body matrix and its determinant. *SIAM Journal on Applied Algebra and Geometry*, 3(1):67–86, 2019.

Grohe, M. and Schweitzer, P. The graph isomorphism problem. *Communications of the ACM*, 63(11):128–134, 2020.

Guo, C., Gardner, J., You, Y., Wilson, A. G., and Weinberger, K. Simple black-box adversarial attacks. In *International Conference on Machine Learning*, pp. 2484–2493, 2019.

Helfgott, H. A., Bajpai, J., and Dona, D. Graph isomorphisms in quasi-polynomial time. *arXiv:1710.04574*, 2017.

Hopcroft, J. E. and Karp, R. M. An nˆ5/2 algorithm for maximum matchings in bipartite graphs. *SIAM Journal on Computing*, 2(4):225–231, 1973.

Hordan, S., Amir, T., Gortler, S. J., and Dym, N. Complete neural networks for euclidean graphs. In *AAAI Conference on Artificial Intelligence*, volume 38 (11), pp. 12482–12490, 2024.

Horn, R. A. and Johnson, C. R. *Matrix analysis*. Cambridge University Press, 2012.

Huttenlocher, D. P., Klanderman, G. A., and Rucklidge, W. J. Comparing images using the Hausdorff distance. *Transactions on pattern analysis and machine intelligence*, 15 (9):850–863, 1993.

Kiefer, S., Ponomarenko, I., and Schweitzer, P. The weisfeiler–leman dimension of planar graphs is at most 3. *Journal of the ACM*, 66(6):1–31, 2019.

Kondor, R. and Trivedi, S. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International Conference on Machine Learning*, pp. 2747–2755, 2018.

Laidlaw, C. and Feizi, S. Functional adversarial attacks. *Adv. Neural Information Proc. Systems*, 32, 2019.

Lang, L., Cezar, H. M., Adamowicz, L., and Pedersen, T. B. Quantum definition of molecular structure. *Journal of the American Chemical Society*, 146(3):1760–1764, 2024.

Leman, A. and Weisfeiler, B. A reduction of a graph to a canonical form and an algebra arising during this reduction. *Nauchno-Technicheskaya Informatsiya*, 2(9):12–16, 1968.

Li, Z., Wang, X., Huang, Y., and Zhang, M. Is distance matrix enough for geometric deep learning? *arXiv:2302.05743*, 2023.

Liberti, L. and Lavor, C. *Euclidean distance geometry*. Springer, 2017.

Lim, S., Mémoli, F., and Smith, Z. The gromov–hausdorff distance between spheres. *Geometry & Topology*, 27(9): 3733–3800, 2023.

Maennel, H., Unke, O. T., and Müller, K.-R. Complete and efficient covariants for 3d point configurations with application to learning molecular quantum properties. *arXiv:2409.02730*, 2024.

Majhi, S. and Wenk, C. Distance measures for geometric graphs. *arXiv:2209.12869*, 2022.

Majhi, S., Vitter, J., and Wenk, C. Approximating gromov-hausdorff distance in euclidean space. *Computational Geometry*, 116:102034, 2024.

Maron, H., Ben-Hamu, H., Serviansky, H., and Lipman, Y. Provably powerful graph networks. *Advances in neural information processing systems*, 32, 2019.

Melter, R. A. and Tomescu, I. Metric bases in digital geometry. *Computer vision, graphics, and image Processing*, 25(1):113–121, 1984.

Mémoli, F. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11:417–487, 2011.

Mémoli, F., Smith, Z., and Wan, Z. The Gromov-Hausdorff distance between ultrametric spaces: its structure and computation. *arXiv:2110.03136*, 2021.

Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., and Cubuk, E. D. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.

Nemec, L. Principal component analysis (pca): A physically intuitive mathematical introduction. https://towardsdatascience.com/principal-component-analysis-pca-8133b02f1ibd, 2022.

Nigam, J., Pozdnyakov, S. N., Huguenin-Dumittan, K. K., and Ceriotti, M. Completeness of atomic structure representations. *APL Machine Learning*, 2(1), 2024.

Nikolentzos, G., Meladianos, P., and Vazirgiannis, M. Matching node embeddings for graph similarity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Oliynyk, A. O., Antono, E., Sparks, T. D., Ghadbeigi, L., Gaultois, M. W., Meredig, B., and Mar, A. High-throughput machine-learning-driven synthesis of full-heusler compounds. *Chemistry of Materials*, 28(20): 7324–7331, 2016.

Pozdnyakov, S. N., Willatt, M. J., Bartók, A. P., Ortner, C., Csányi, G., and Ceriotti, M. Incompleteness of atomic structure representations. *Phys. Rev. Lett.*, 125:166001, 2020. URL arXiv:2001.11696.

Qi, G.-J. and Luo, J. Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods. *Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2168–2187, 2020.

Ramakrishnan, R., Dral, P. O., Rupp, M., and Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

Rass, S., König, S., Ahmad, S., and Goman, M. Metricizing the euclidean space towards desired distance relations in point clouds. *IEEE Transactions on Information Forensics and Security*, 2024.

Rieder, S. R., Oliveira, M. P., Riniker, S., and Hünenberger, P. H. Development of an open-source software for isomer enumeration. *Journal of Cheminformatics*, 15(1):10, 2023.

Rubner, Y., Tomasi, C., and Guibas, L. The Earth Mover's Distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

Sato, R., Cuturi, M., Yamada, M., and Kashima, H. Fast and robust comparison of probability measures in heterogeneous spaces. *arXiv:2002.01615*, 2020.

Schmiedl, F. Computational aspects of the Gromov–Hausdorff distance and its application in non-rigid shape matching. *Discrete Comp. Geometry*, 57:854–880, 2017.

Schoenberg, I. Remarks to Maurice Frechet's article "Sur la definition axiomatique d'une classe d'espace distances vectoriellement applicable sur l'espace de Hilbert. *Annals of Mathematics*, pp. 724–732, 1935.

Shirdhonkar, S. and Jacobs, D. Approximate earth mover's distance in linear time. In *Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

Sippl, M. J. and Scheraga, H. A. Cayley-menger coordinates. *Proceedings of the National Academy of Sciences*, 83(8): 2283–2287, 1986.

Steck, H., Ekanadham, C., and Kallus, N. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM on Web Conference 2024*, pp. 887–890, 2024.

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., and Jain, A. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98, 2019.

Tu, E., Wang, Z., Yang, J., and Kasabov, N. Deep semi-supervised learning via dynamic anchor graph embedding in latent space. *Neural Networks*, 146:350–360, 2022.

Vasylenko, A., Antypov, D., Schewe, S., Daniels, L. M., Claridge, J. B., Dyer, M. S., and Rosseinsky, M. J. Digital features of chemical elements extracted from local geometries in crystal structures. *Digital Discovery*, 2025.

Ward, L., Agrawal, A., Choudhary, A., and Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2(1):1–7, 2016.

Weinhold, F. Metric geometry of equilibrium thermodynamics. *The Journal of Chemical Physics*, 63(6):2479–2483, 1975.

Weston, L., Tshitoyan, V., Dagdelen, J., Kononova, O., Trewartha, A., Persson, K. A., Ceder, G., and Jain, A. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *Journal of chemical information and modeling*, 59(9):3692–3702, 2019.

Widdowson, D. and Kurlin, V. Resolving the data ambiguity for periodic crystals. *Advances in Neural Information Processing Systems*, 35:24625–24638, 2022.

Widdowson, D. and Kurlin, V. Recognizing rigid patterns of unlabeled point clouds by complete and continuous isometry invariants with no false negatives and no false positives. In *Proceedings of CVPR*, pp. 1275–1284, 2023.

Wilson, S. R., Cui, W., Moskowitz, J. W., and Schmidt, K. E. Applications of simulated annealing to the conformational analysis of flexible molecules. *Journal of computational chemistry*, 12(3):342–349, 1991.

Zhang, B., Fan, C., Liu, S., Huang, K., Zhao, X., Huang, J., and Liu, Z. The expressive power of graph neural networks: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 37:1455–1474, 2024.

Zhou, D.-X. Universality of deep convolutional neural networks. *Applied and computational harmonic analysis*, 48 (2):787–794, 2020.

Zhou, Q., Tang, P., Liu, S., Pan, J., Yan, Q., and Zhang, S.-C. Learning atoms for materials discovery. *Proceedings of the National Academy of Sciences*, 115(28):E6411–E6417, 2018.

A. Extra details of experiments on the world's largest 3D molecular databases QM9 and GD

The default 4-layer network from TensorFlow used a "sequential" mode, 3 epochs, and the settings in Table 6.

Table 6. Parameters of the TensorFlow network for predictions in Table 4.

LAYER (TYPE)	OUTPUT SHAPE	NUMBER OF PARAMETERS
DENSE (DENSE)	(NONE, 32)	352
BATCH_NORMALIZATION	(NONE, 32)	128
RE_LU (RELU)	(NONE, 32)	0
DENSE_1 (DENSE)	(NONE, 5)	165

Past maps of QM9 in Fig. 5 based on eigenvalues are too dense without clear separation. Even if we zoom in, these two or three incomplete invariants will not provide any extra separation. The complete invariants NDP contain much more geometric information.

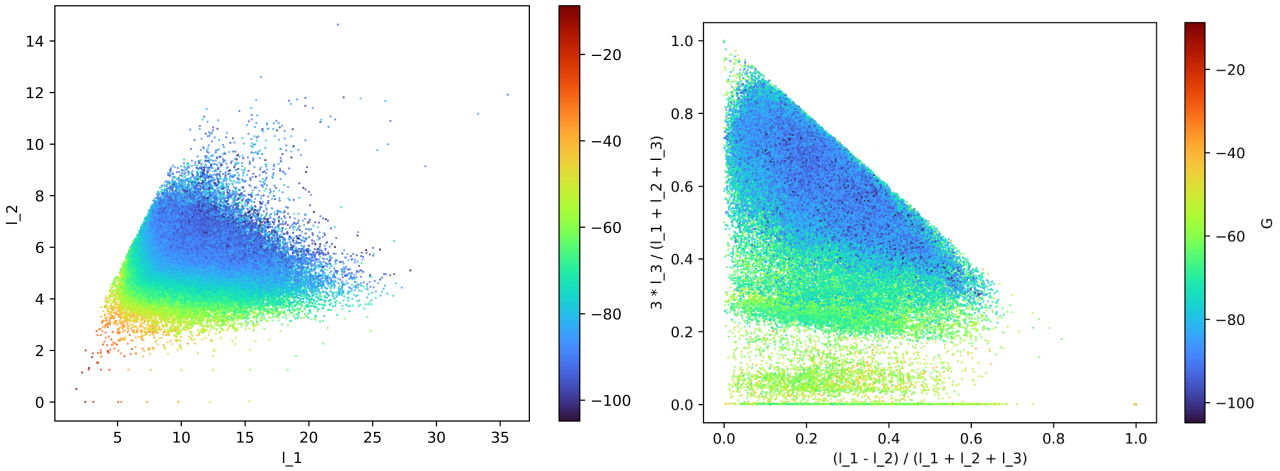

Figure 5. **Left**: each dot represents one QM9 molecule whose atomic cloud has two largest roots $l_1 \geq l_2$ of eigenvalues (moments of inertia (Nemec, 2022) or elongations in two principal directions) in Angstroms ($1\mathring{A} = 10^{-10}m \approx$ smallest interatomic distance). The color represents the free energy G characterizing molecular stability. **Right**: each dot represents one QM9 molecule whose atomic cloud has coordinates x, y expressed via the roots $l_1 \geq l_2 \geq l_3 \geq 0$ of three eigenvalues.

Fig. 7 shows the simplest geographic-style map of QM9 as a finite sample within $\bigcup_{m=3}^{29} \mathrm{GRS}(\mathbb{R}^3; m)$ projected to the invariants $\mathrm{SRD}_1 \geq \mathrm{SRD}_2$. All molecules on the horizontal axis $y = \mathrm{SRD}_1 - \mathrm{SRD}_2 = 0$ have $\mathrm{SRD}_1 = \mathrm{SRD}_2$ (due to two equidistant atoms from the center of mass) and can be projected (like any subset of QM9) to other coordinates as in Fig. 8. Molecular properties can be visualized on these geographic maps as 'mountainous' landscapes.

Table 7. Past ML and non-ML predictions of chemical elements have lower accuracies than by distance invariants in Table 4.

METHOD	DESCRIPTION	ACCURACY	REFERENCE
LEAF	LOCAL COORDINATION GEOMETRY	86%	(VASYLENKO ET AL., 2025)
MATSCHOLAR	ML-DERIVED FROM LITERATURE	81%	(WESTON ET AL., 2019)
MAT2VEC	ML-DERIVED FROM LITERATURE	80%	(TSHITOYAN ET AL., 2019)
ATOM2VEC	ML-DERIVED FROM COMPOSITIONAL CONTENT	79%	(ZHOU ET AL., 2018)
GNOME	FREQUENCY OF ELEMENTS AT THE SAME ATOMIC SITES	79%	(MERCHANT ET AL., 2023)
MAGPIE	ELEMENTAL PHYSICAL CHARACTERISTICS	78%	(WARD ET AL., 2016)
OLIYNYK	ELEMENTAL PHYSICAL CHARACTERISTICS	75%	(OLIYNYK ET AL., 2016)
MEGNET	ML-DERIVED FROM ATOM, BOND AND GRAPH ATTRIBUTES	73%	(CHEN ET AL., 2019)
SKIPATOM	ML-DERIVED FROM ATOM CONNECTIVITY GRAPHS	68%	(ANTUNES ET AL., 2022)

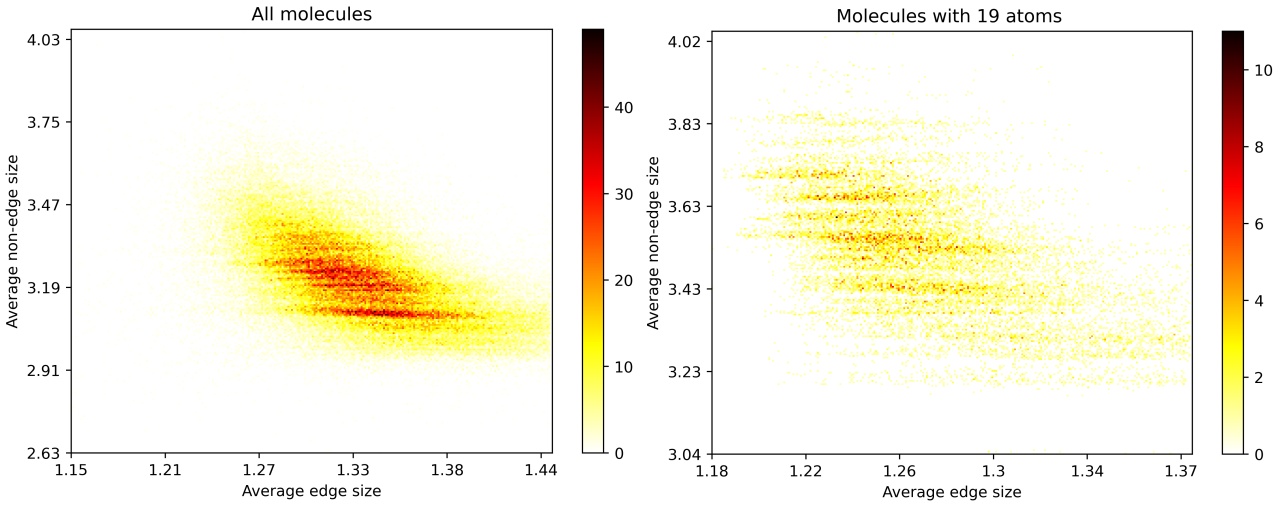

Figure 6. **Left**: the heatmap of all molecular graphs from QM9 in the simplest continuous invariants. **Right**: 18336 graphs with 19 atoms. The color indicates the number of molecules at every pixel.

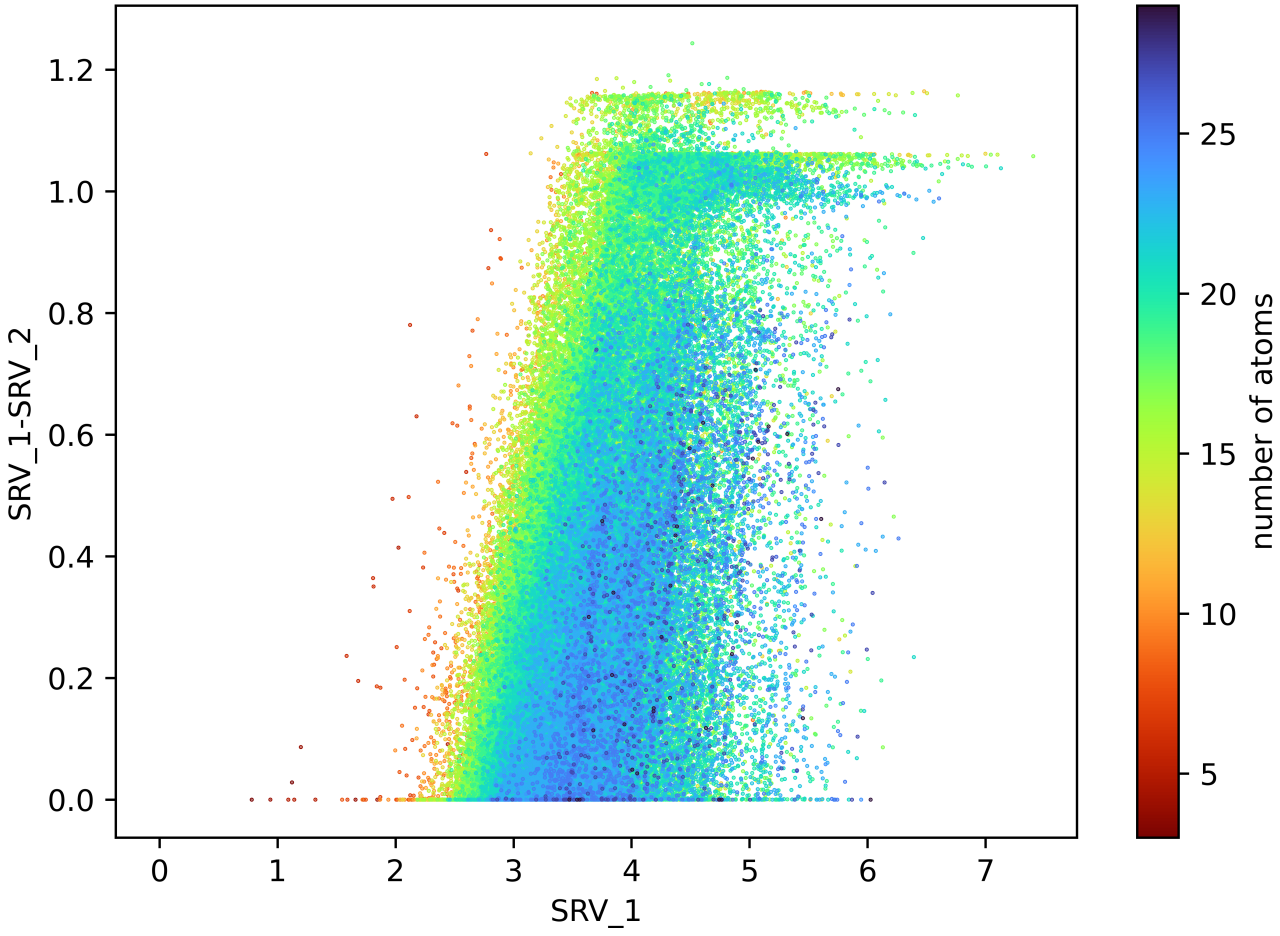

Figure 7. Every dot represents a molecular graph with the invariant coordinates $x = \mathrm{SRD}_1$, $y = \mathrm{SRD}_1 - \mathrm{SRD}_2$, all in Angstroms, where $1\text{Å} = 10^{-10}m \approx$ the smallest interatomic distance.

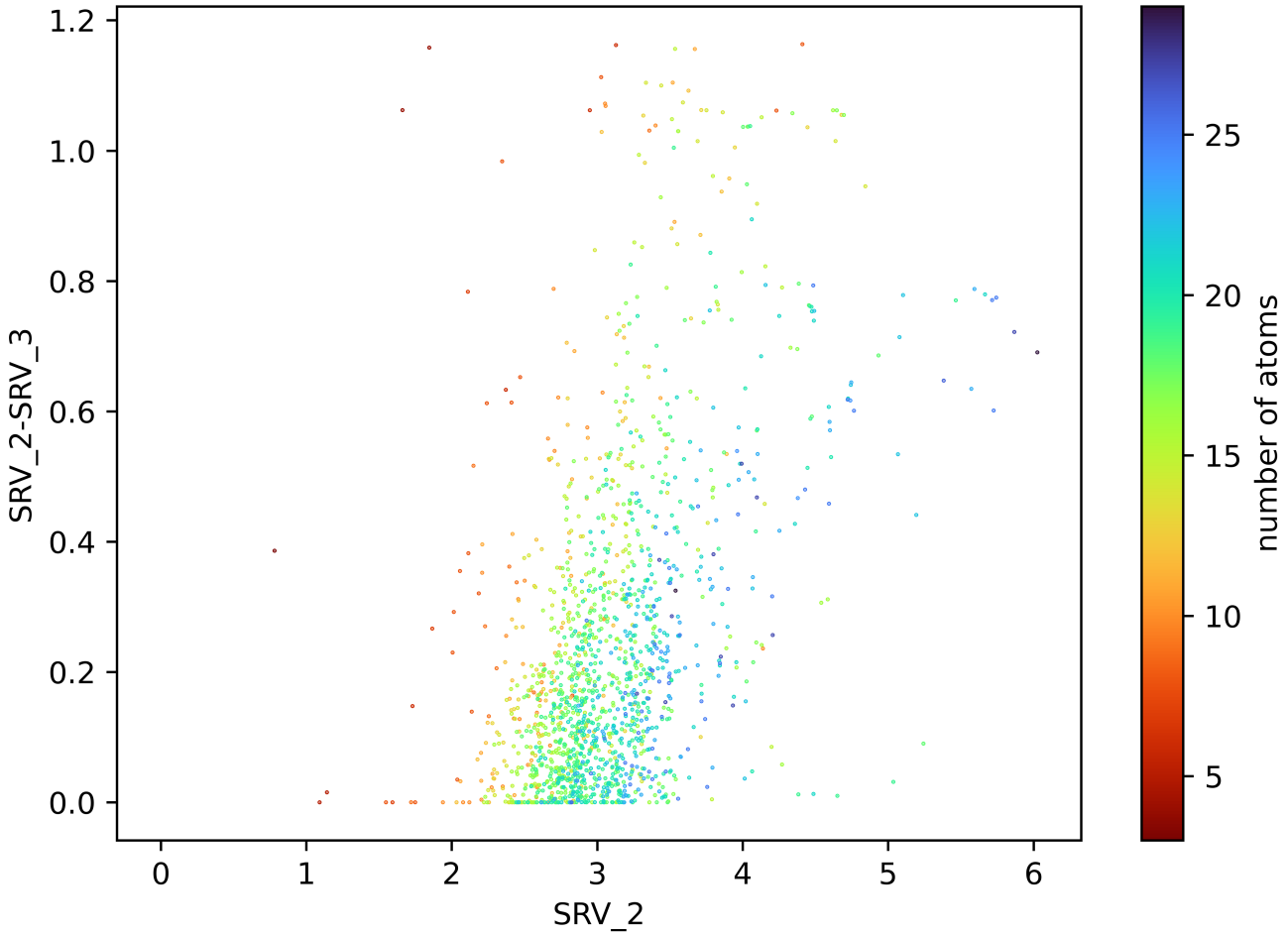

Figure 8. The projection of QM9 to $x = \mathrm{SRD}_1$, $y = \mathrm{SRD}_2 - \mathrm{SRD}_3$.

B. Invariants and metrics on Euclidean graphs in any dimension $n \geq 2$

This section extends all new concepts and results from sections 3 and 4 to any dimension $n \geq 2$. Any n vectors $p_1, \ldots, p_n \in \mathbb{R}^n$ can be written as columns in the $n \times n$ matrix whose determinant has $\mathrm{sign}(p_1, \ldots, p_n)$, which is ± 1 or 0 (if $p_1, \ldots, p_n$ are linearly dependent).

Definition B.1 (Centered Representation $\mathrm{CR}(G; A)$ of a graph with a sequence $A \subset V(G)$)**.** *Let $G \subset \mathbb{R}^n$ be a graph on m unordered points with the center of mass $O(G) = 0$. For any $1 \leq h \leq n$, fix a* base sequence *A of ordered vertices $p_1, \ldots, p_h \in V(G)$. If $h = n$, let $\mathrm{sign}(A)$ be the sign of the $n \times n$ determinant on the vectors $p_1, \ldots, p_n$, else $\mathrm{sign}(A) = 0$. Let $D(A)$ be the matrix of signed distances between the ordered points $0 = p_0, p_1, \ldots, p_h$. The matrix $R(G; A)$ has $m - h$ unordered columns, one for each vertex $q \in V(G) - A$, consisting of $h + 1$ distances $d(q, p_i)$ for $i = 0, \ldots, h$, where $p_0 = 0$. The* Centered Representation $\mathrm{CR}(G; A)$ *is the triple* $[\mathrm{sign}(A), D(A), R(G; A)]$.

Definition B.2 (Nested Centered Distribution $\mathrm{NCD}(G; h)$ of order h)**.** *Let $G \subset \mathbb{R}^n$ be any Euclidean graph on m unordered vertices and the center of mass at the origin $0 \in \mathbb{R}^n$. Fix an* order $1 \leq h \leq n$.

(a) *For any $h - 1$ distinct ordered vertices $p_1, \ldots, p_{h-1} \in V(G)$, the* Centered Distribution $\mathrm{CD}^{(h)}_{h-1}(G; p_1, \ldots, p_{h-1})$ *of index $h - 1$ is the unordered set of Centered Representation $\mathrm{CR}(G; p_1, \ldots, p_h)$ from Definition B.1 for all $p_h \in V(G) - \{p_1, \ldots, p_{h-1}\}$.*

(b) *Now we will iteratively decrement an integer k from $h - 1$ down to 1 and define $\mathrm{CD}^{(h)}_{k-2}$ of index $h - 2$, and so on until $\mathrm{CD}^{(h)}_1$ of index $k = 1$. For the initial $k = h - 1$, we use $\mathrm{CD}^{(h)}_k = \mathrm{CD}^{(h)}_{k-1}$ defined in part (a) above. For any $k - 1$ distinct*

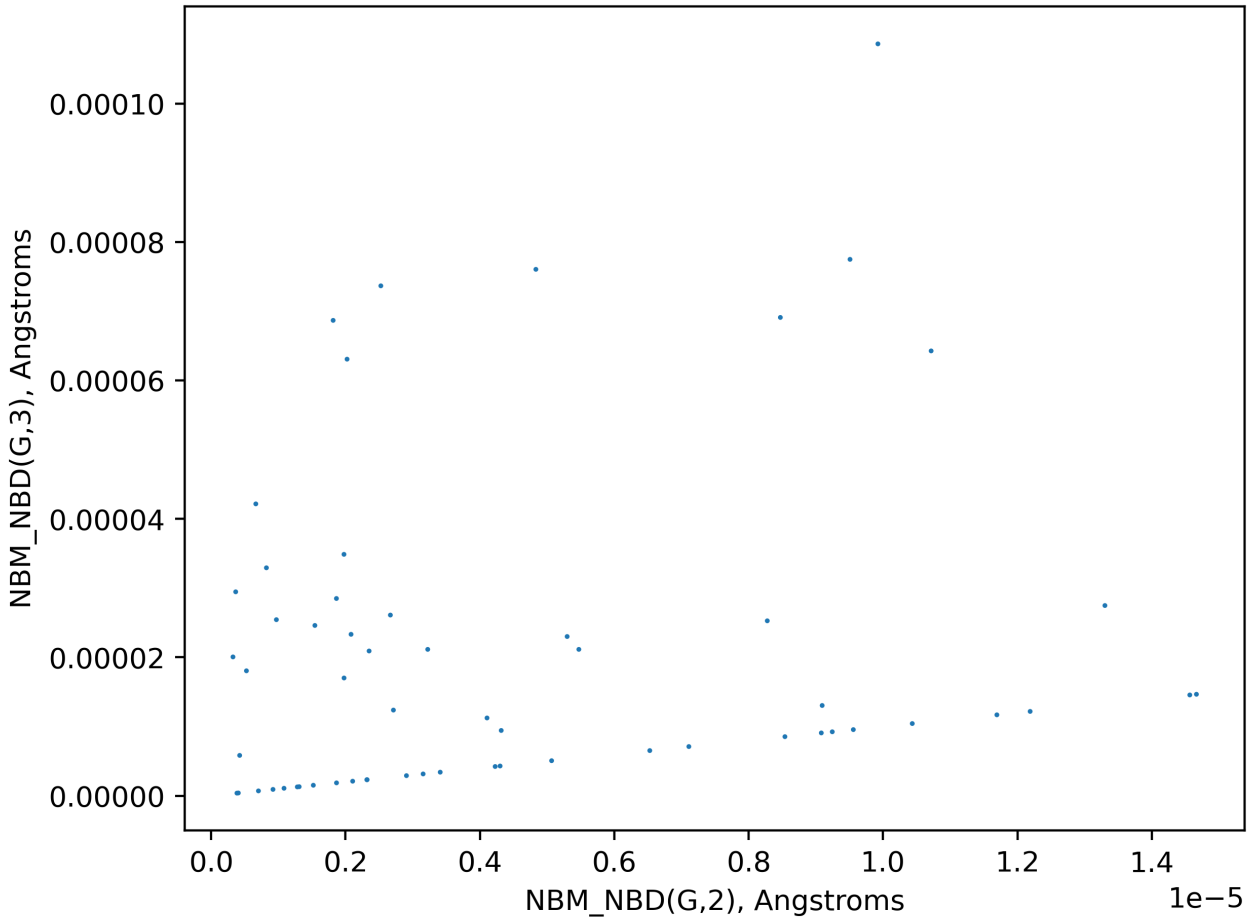

Figure 9. Each dot is a comparison of molecular graphs from QM9 by the distances on the progressively stronger invariants: $\mathrm{NCD}(G;2)$ vs $\mathrm{NCD}(G;3)$.

ordered vertices $p_1,\dots,p_{k-1}\in V(G)$*, the* Centered Distribution $\mathrm{CD}^{(h)}_{k-1}(G;p_1,\dots,p_{k-1})$ *of index* $k-1$ *is the unordered collection of* $\mathrm{CD}^{(h)}_{k}(C;p_1,\dots,p_k)$ *of index* k *for all vertices* $p_k\in V(G)-\{p_1,\dots,p_{k-1}\}$.

(c) *The* Nested Centered Distribution $\mathrm{NCD}(G;h)$ *of order* h *is the unordered collection of* $\mathrm{CD}^{(h)}_{1}(G;p_1)$ *of index 1 for all vertices* $p_1\in V(G)$*. For the order* $h=n$*, the* mirror image $\overline{\mathrm{NCD}}(G;n)$ *is obtained from* $\mathrm{NCD}(G;n)$ *by reversing* $\mathrm{sign}(p_1,\dots,p_n)$ *of* $n\times n$ *determinants in all* $\mathrm{CR};p_1,\dots,p_n)$.

If a sequence $0\cup A=(p_0,p_1,\dots,p_n)\subset\mathbb{R}^n$ degenerates to a lower dimensional subspace, i.e. the vectors $p_1,\dots,p_n$ become linearly dependent, then $\mathrm{sign}(A)$ of discontinuously changes. To guarantee the Lipschitz continuity, we multiply these signs by the strength σ below, while the volume $\mathrm{vol}(0\cup A)$ of the simplex on $0\cup A$ is not Lipschitz continuous.

Definition B.3 (*strength* $\sigma(C)$)**.** *For any sequence* C *of* $n+1$ *ordered points* $p_0,\dots,p_n\in\mathbb{R}^n$*, the* half-perimeter $p(C)=\dfrac{1}{2}\sum\limits_{1\le i<j\le n}|p_i-p_j|$ *is the half-sum of pairwise distances between points of* C*. Let* $\mathrm{vol}(C)$ *denote the volume of the* n*-dimensional simplex on* C*. The* strength *of the simplex* C *is* $\sigma(C)=\dfrac{\mathrm{vol}^2(C)}{p^{2n-1}(C)}$.

In dimension $n=1$, for any pair $C=\{p_0,p_1\}\subset\mathbb{R}$, the volume $\mathrm{vol}(C)$ is the length $|p_0-p_1|$, the half-perimeter distance

$p(C)$ is the half-distance $\frac{1}{2}|p_0 - p_1|$, so the strength is $\sigma(C) = \dfrac{\mathrm{vol}^2(C)}{p(C)} = 2|p_0 - p_1|$.

Lemma B.4 (Theorem 4.4 in (Widdowson & Kurlin, 2023)). *Let B be obtained from a sequence $A \subset \mathbb{R}^n$ of n points by perturbing every point within its ε-neighborhood. Then $|\sigma(A) - \sigma(B)| \le 2\varepsilon\lambda_n$ for a constant λ_n, where $\lambda_1 = 2$, $\lambda_2 = 2\sqrt{3}$, $\lambda_3 \approx 0.43$.*

Definition B.5 (max metric M_∞ on CRs). *Let Euclidean graphs $G, F \subset \mathbb{R}^n$ on m unordered vertices have base sequences A, B of $h \le n$ vertices. Consider the $m - h$ columns of $R(G; A)$ as a cloud of $m - h$ unordered points in $\mathbb{R}^h$, also for $R(F; B)$. The* max *metric $M_\infty(\mathrm{CR}(G; A), \mathrm{CR}(F; B))$ is the maximum of $\dfrac{2}{\lambda_n}|\mathrm{sign}(A)\sigma(0 \cup A) - \mathrm{sign}(B)\sigma(0 \cup B)|$, $L_\infty(D(A), D(B))$, and the bottleneck distance $W_\infty(R(G; A), R(F; B))$, where all signs are zeros for $h < n$.*

In Definition B.5, λ_n is the Lipschitz constant of σ from Lemma 4.2.

Definition B.6 (Nested Bottleneck Metric NBM on NCDs). *Let $G, F \subset \mathbb{R}^n$ be any Euclidean graphs on m unordered vertices. For any ordered vertices $p_1 \dots, p_{h-1} \in V(G)$ and $q_1 \dots, q_{h-1} \in V(F)$, the complete bipartite graph $\Gamma(G; p_1, \dots, p_{h-1}; F; q_1, \dots, q_{h-1})$ has $m - h + 1$ white vertices and $m - h + 1$ black vertices representing $\mathrm{CR}(G; p_1, \dots, p_h)$ and $\mathrm{CR}(F; q_1, \dots, q_h)$ for all $m - h + 1$ vertices $p_h \in V(G) - \{p_1, \dots, p_{h-1}\}$ and $q_h \in V(F) - \{q_1, \dots, q_{h-1}\}$, respectively. Set the* weight $w(e)$ *of an edge e joining the vertices represented by $\mathrm{CR}(G; p_1, \dots, p_h)$, $\mathrm{CR}(F; q_1, \dots, q_h)$ as the max metric M_∞ between these distributions, see Definition B.5. Then Definition 4.4 gives the bottleneck matching distance $\mathrm{BMD}(\Gamma(G; p_1, \dots, p_{h-1}; F; q_1, \dots, q_{h-1}))$.*

For any integer $1 \le i < h$ and ordered vertices $p_1 \dots, p_{i-1} \in V(G)$ and $q_1 \dots, q_{i-1} \in V(F)$, the complete bipartite graph $\Gamma(G; p_1, \dots, p_{i-1}; F; q_1, \dots, q_{i-1})$ has $m - i + 1$ white vertices and $m - i + 1$ black vertices representing $\mathrm{CD}_i^{(h)}(G; p_1, \dots, p_i)$ and $\mathrm{CD}_i^{(h)}(F; q_1, \dots, q_i)$ for all $m - i + 1$ variable vertices $p_i \in V(G) - \{p_1, \dots, p_{i-1}\}$ and $q_i \in V(F) - \{q_1, \dots, q_{i-1}\}$, respectively. Set the weight $w(e)$ *of an edge e joining the vertices represented by $\mathrm{CD}_i^{(h)}(G; p_1, \dots, p_i)$ and $\mathrm{CD}_i^{(h)}(F; q_1, \dots, q_i)$ as the previously computed distance $\mathrm{BMD}(\Gamma(G; p_1, \dots, p_i; F; q_1, \dots, q_i))$ for a smaller number i of fixed vertices. Then Definition 4.4 gives the bottleneck matching distance $\mathrm{BMD}(\Gamma(G; p_1, \dots, p_{i-1}; F; q_1, \dots, q_{i-1}))$. For $i = 1$, the graph $\Gamma(G, F)$ has $m + m$ vertices representing $\mathrm{CD}_1(G; p_1)$, $\mathrm{CD}_1^{(h)}(F; q_1)$ for all $p_1 \in V(G)$ and $q_1 \in V(F)$. The* Nested Bottleneck Metric $\mathrm{NBM}(\mathrm{NCD}(G; h), \mathrm{NCD}(F; h))$ *is the Bottleneck Matching Distance $\mathrm{BMD}(\Gamma(G, F))$.*

C. Metrics on graphs and their continuity under perturbations

This appendix verifies the axioms and Lipschitz continuity for all auxiliary metrics in section 4.

Lemma C.1 (metric axioms for the bottleneck matching distance BMD). *Let S, Q be any unordered distributions of the same number of objects with a base metric d. Define the complete bipartite graph $\Gamma(S, Q)$ whose every edge e joining objects $R_S \in S$ and $R_Q \in Q$ has the weight $w(e) = d(R_S, R_Q)$. Then the bottleneck matching distance $\mathrm{BMD}(\Gamma(S, Q))$ from Definition 4.4 satisfies all metric axioms on such unordered distributions.*

Proof of Lemma C.1. The coincidence axiom means that $\mathrm{NBM}(S, Q) = 0$ if and only if the weighted distributions S, Q are equal in the sense that there is a bijection $g : S \to Q$ so that $d(g(R), R) = 0$ for any $R \in S$.

Indeed, if the weighted distributions S, Q can be matched by a bijection, we get a vertex matching E of $\Gamma(S, Q)$ whose all edges have weights $w(e) = 0$. Definition 4.4 implies that $\mathrm{BMD}(\Gamma(S, Q)) = 0$ as required.

Conversely, if $\mathrm{BMD}(\Gamma(S, Q)) = 0$, there is a vertex matching E in $\Gamma(S, Q)$ with all $w(e) = 0$. This matching E defines a required bijection $S \to Q$. The symmetry $\mathrm{BMD}(\Gamma(S, Q)) = \mathrm{BMD}(\Gamma(Q, S))$ follows from Definition 4.4 and the symmetry of the base metric d.

To prove the triangle inequality

$$\mathrm{BMD}(\Gamma(S, Q)) + \mathrm{BMD}(\Gamma(Q, T)) \ge \mathrm{BMD}(\Gamma(S, T)),$$

let E_{SQ}, E_{QT} be optimal vertex matchings in the graphs $\Gamma(S, Q), \Gamma(Q, T)$, respectively, such that

$$\mathrm{BMD}(\Gamma(S, Q)) = W(E_{SQ}), \mathrm{BMD}(\Gamma(Q, T)) = W(E_{QT}),$$

see Definition 4.4. The composition $E_{SQ} \circ E_{QT}$ is a vertex matching in $\Gamma(S,T)$, so $W(E_{SQ} \circ E_{QT}) \geq \mathrm{BMD}(\Gamma(S,T))$. It suffices to prove that

$$W(E_{SQ}) + W(E_{QT}) \geq W(E_{SQ} \circ E_{QT}).$$

Let e_{ST} be an edge with a largest weight from $E_{SQ} \circ E_{QT}$, so $W(E_{SQ} \circ E_{QT}) = w(e_{ST})$. The edge e_{ST} can be considered the union of edges $e_{SQ} \in E_{SQ}, e_{QT} \in E_{QT}$.

By the triangle inequality for the base metric d,

$$w(e_{SQ}) + w(e_{QT}) \geq w(e_{ST}) = W(E_{SQ} \circ E_{QT})$$

implies that

$$W(E_{SQ}) + W(E_{QT}) \geq W(E_{SQ} \circ E_{QT})$$

because both terms on the left-hand side are maximized for all edges (not only e_{SQ}, e_{QT}) from E_{SQ}, E_{QT}. □

Definition C.2 below makes sense for any distributions $\{[R_1,w_1],\dots,[R_m,w_m]\}$, where $R_1,\dots,R_m$ are objects with a base metric d and weights $w_1,\dots,w_m \in [0,1]$. Each R_i can be CBR or CBD of any depth with a base metric M_∞ or BMD from Definitions B.5, B.6.

Definition C.2 (EMD)**.** *Let $S = \{[R_i(S), w_i(S)]\}_{i=1}^{m(S)}$ and $Q = \{[R_j(Q), w_j(Q)]\}_{j=1}^{m(Q)}$ be weighted distributions of objects $R_i(S), R_j(Q)$, which live in a space with a metric d. A* flow *from S to Q is an $m(S) \times m(Q)$ matrix whose element $f_{ij} \in [0,1]$ represents a* partial flow *from $R_i(S)$ to $R_j(Q)$. The* Earth Mover's Distance *is the minimum* cost $\mathrm{EMD}(S,Q) = \sum\limits_{i=1}^{m(S)} \sum\limits_{j=1}^{m(Q)} f_{ij} d(R_i(S), R_j(Q))$ *for variable 'flows' $f_{ij} \in [0,1]$ subject to the conditions $\sum\limits_{j=1}^{m(Q)} f_{ij} \leq w_i(S)$ for $i = 1,\dots,m(S)$, $\sum\limits_{i=1}^{m(S)} f_{ij} \leq w_j(Q)$ for $j = 1,\dots,m(Q)$, and $\sum\limits_{i=1}^{m(S)} \sum\limits_{j=1}^{m(Q)} f_{ij} = 1$.*

The first condition $\sum\limits_{j=1}^{m(Q)} f_{ij} \leq w_i(S)$ means that not more than the weight $w_i(S)$ of $R_i(S)$ 'flows' into all $R_j(Q)$ via 'flows' f_{ij}, $j = 1,\dots,m(Q)$. The second condition $\sum\limits_{i=1}^{m(S)} f_{ij} = w_j(Q)$ means that all 'flows' f_{ij} from $R_i(S)$ for $i = 1,\dots,m(S)$ 'flow' into $R_j(Q)$ up to the maximum weight $w_j(Q)$. The last condition $\sum\limits_{i=1}^{m(S)} \sum\limits_{j=1}^{m(Q)} f_{ij} = 1$ forces to 'flow' all rows $R_i(S)$ to all rows $R_j(Q)$.

The EMD satisfies all metric axioms, see the appendix in (Rubner et al., 2000), needs $O(m^3 \log m)$ time for distributions of a maximum size m and is approximated in $O(m)$ time, see (Shirdhonkar & Jacobs, 2008; Sato et al., 2020).

Definition C.2 can be adapted for the EMD between NDDs by (1) replacing the bottleneck distance W_∞ in Definition B.5 with EMD between clouds of equally weighted points, and (2) replacing $\mathrm{BMD}(\Gamma)$ for a bipartite graph Γ with $\mathrm{EMD}(\Gamma)$ between the unordered sets (of potentially different sizes) of BDDs with weights on all white vertices and BDDs on all black vertices.

The Lipschitz continuity of NDD and EMD in Theorem D.1(c) needs Lemmas C.3, C.4, D.9.

If a metric graph G lives in an ambient metric space X, a natural perturbation of G is a shift of every vertex of G up to ε in the metric of X. Then the distance $d(p,q)$ between any vertices p,q of G changes by at most 2ε.

We will prove the continuity in more general settings by only assuming that $d(p,q)$ changes by at most 2ε for any $p,q \in V(G)$ without requiring an ambient space X.

Lemma C.3 (Lipschitz continuity of BMD)**.** *Let Γ be a complete bipartite graph with a vertex matching E such that any $e \in E$ has a weight $w(e) \leq \varepsilon$. Then $\mathrm{BMD}(\Gamma) \leq \varepsilon$.*

Proof of Lemma C.3. By Definition 4.4, the given matching E has the weight $W(E) = \max\limits_{e \in E} w(e) \leq \varepsilon$. Since $\mathrm{BMD}(\Gamma) = \min\limits_{E} W(E)$ is minimized for all vertex matchings, we get $\mathrm{BMD}(\Gamma) \leq \varepsilon$. □

Lemma C.4 (Lipschitz continuity of EMD)**.** *In Definition C.2, let distributions S, Q have a bijection $R_i(S) \leftrightarrow R_i(Q)$ between equally weighted objects such that $d(R_i(S), R_i(Q)) \leq \varepsilon$ for all $i = 1, \dots, m$, where $m = m(S) = m(Q)$. Then* $\mathrm{EMD}(S, Q) \leq \varepsilon$.

Proof of Lemma C.4. In Definition C.2, choose partial flows $f_{ij} = \dfrac{1}{m}$ for $i = j$, otherwise $f_{ij} = 0$. Then $\mathrm{EMD}(S, Q) \leq \sum\limits_{i=1}^{m} \sum\limits_{j=1}^{m} f_{ij} d(R_i(S), R_j(Q)) = \sum\limits_{i=1}^{m} \dfrac{1}{m} d(R_i(S), R_i(Q)) \leq \dfrac{1}{m} \sum\limits_{i=1}^{m} \varepsilon = \varepsilon$. □

D. Proofs for Euclidean graphs from section 3

This appendix rigorously proves all parts of Theorem D.1.

Theorem D.1 (NCD solves Problem 1.1)**.** ***(a)*** *The Nested Centered Distribution* $\mathrm{NCD}(G; h)$ *in Definition B.2 is invariant under any rigid motion for all Euclidean graph G on m unordered vertices and, for a fixed dimension n, can be computed in time $O(n^2 m^{h+1})$ with space $O(n^2 m^{h+1})$ for any order $1 \leq h \leq n$.*

(b) $\mathrm{NCD}(G; 2)$ *is a complete invariant of all graphs $G \subset \mathbb{R}^2$ under rigid motion from the group* $\mathrm{SE}(n)$ *in any dimension $n \geq 1$.*

(c) *Perturbing each vertex of a graph $G \subset \mathbb{R}^n$ within its ε-neighborhood changes* $\mathrm{NCD}(G; h)$ *up to 2ε in both metrics* NBM *and* EMD *for any order $1 \leq h \leq n$.*

(d) *For any graphs $G, F \subset \mathbb{R}^n$ on m unordered vertices, the metrics* NBM *and* EMD *between the invariants* $\mathrm{NCD}(G; h)$ *and* $\mathrm{NCD}(F; h)$ *from Definition B.6 can be computed in time $O(m^{2h+1.5} \log^{h+1} m)$ with space $O(n^2 m^{2h+1} \log^{h-1} m)$ for any order $1 \leq h \leq n$.*

The *affine dimension* $0 \leq \mathrm{aff}(A) \leq n$ of a cloud $A = \{p_1, \dots, p_m\} \subset \mathbb{R}^n$ is the maximum dimension of the vector space generated by all inter-point vectors $p_i - p_j$, $i, j \in \{1, \dots, m\}$. Then $\mathrm{aff}(A)$ is an isometry invariant and is independent of an order of points of A. Any cloud A of 2 distinct points has $\mathrm{aff}(A) = 1$. Any cloud A of 3 points that are not in the same straight line has $\mathrm{aff}(A) = 2$.

Lemma D.2 provides a simple criterion for a matrix to be realizable by squared distances of a point cloud in $\mathbb{R}^n$.

Lemma D.2 (realization of distances)**.** ***(a)*** *A symmetric $m \times m$ matrix of $s_{ij} \geq 0$ with $s_{ii} = 0$ is realizable as a matrix of squared distances between points $p_0 = 0, p_1, \dots, p_{m-1} \in \mathbb{R}^n$* if and only if *the $(m-1) \times (m-1)$ matrix $g_{ij} = \dfrac{s_{0i} + s_{0j} - s_{ij}}{2}$ has only non-negative eigenvalues.*

(b) *If the condition in (a) holds,* $\mathrm{aff}(0, p_1, \dots, p_{m-1})$ *equals the number $k \leq m - 1 \leq n$ of positive eigenvalues. Also in this case, $g_{ij} = p_i \cdot p_j$ define the* Gram matrix GM *of the vectors $p_1, \dots, p_{m-1} \in \mathbb{R}^n$, which are uniquely determined in time $O(m^3)$ up to an orthogonal map in $\mathbb{R}^n$.*

Proof of Lemma D.2. **(a)** We extend Theorem 1 from (Dekster & Wilker, 1987) to the case $m < n + 1$ and justify the reconstruction of $p_1, \dots, p_{m-1}$ in time $O(m^3)$ uniquely in $\mathbb{R}^n$ up to an orthogonal map from $\mathrm{O}(n)$.

The part *only if* $\Rightarrow$. Let a symmetric matrix S consist of squared distances between points $p_0 = 0, p_1, \dots, p_{m-1} \in \mathbb{R}^n$. For $i, j = 1, \dots, m-1$, the matrix with the elements

$$g_{ij} = \frac{s_{0i} + s_{0j} - s_{ij}}{2} = \frac{p_i^2 + p_j^2 - |p_i - p_j|^2}{2} = p_i \cdot p_j$$

is the Gram matrix, which can be written as $\mathrm{GM} = P^T P$, where the columns of the $n \times (m-1)$ matrix P are the vectors $p_1, \dots, p_{m-1}$. For any vector $v \in \mathbb{R}^{m-1}$, we have

$$0 \leq |Pv|^2 = (Pv)^T (Pv) = v^T (P^T P) v = v^T \mathrm{GM} v.$$

Since the quadratic form $v^T \mathrm{GM} v \geq 0$ for any $v \in \mathbb{R}^{m-1}$, the matrix GM is positive semi-definite meaning that GM has only non-negative eigenvalues, see Theorem 7.2.7 in (Horn & Johnson, 2012).

The part *if* $\Leftarrow$. For any positive semi-definite matrix GM, there is an orthogonal matrix Q such that $Q^T \mathrm{GM} Q = D$ is the diagonal matrix, whose $m-1$ diagonal elements are non-negative eigenvalues of GM. The diagonal matrix $\sqrt{D}$ consists of the square roots of eigenvalues of GM.

(b) The number of positive eigenvalues of GM equals the dimension $k = \mathrm{aff}(\{0, p_1, \dots, p_{m-1}\})$ of the subspace in $\mathbb{R}^n$ linearly spanned by $p_1, \dots, p_{m-1}$. We may assume that all $k \le n$ positive eigenvalues of GM correspond to the first k coordinates of $\mathbb{R}^n$. Since $Q^T = Q^{-1}$, the given matrix $\mathrm{GM} = QDQ^T = (Q\sqrt{D})(Q\sqrt{D})^T$ becomes the Gram matrix of the columns of $Q\sqrt{D}$. These columns become the reconstructed vectors $p_1, \dots, p_{m-1} \in \mathbb{R}^n$.

If there is another diagonalization $\tilde{Q}^T \mathrm{GM} \tilde{Q} = \tilde{D}$ for $\tilde{Q} \in \mathrm{O}(n)$, then $\tilde{D}$ differs from D by a permutation of eigenvalues, which is realized by an orthogonal map, so we set $\tilde{D} = D$. Then $\mathrm{GM} = \tilde{Q} D \tilde{Q}^T = (\tilde{Q}\sqrt{D})(\tilde{Q}\sqrt{D})^T$ is the Gram matrix of the columns of $\tilde{Q}\sqrt{D}$.

The new columns differ from the previously reconstructed vectors $p_1, \dots, p_{m-1} \in \mathbb{R}^n$ by the orthogonal map $Q\tilde{Q}^T$. Hence the reconstruction is unique up to $\mathrm{O}(n)$-transformations. Computing eigenvectors $p_1, \dots, p_{m-1}$ requires a diagonalization of GM in time $O(m^3)$ (**?**)section 11.5]press2007numerical. □

Though Lemma D.2 gives a two-sided criterion for realizability of distances by points $p_1, \dots, p_m \in \mathbb{R}^n$, the space of distance matrices is highly singular and cannot be easily sampled. Even $m = 4$ points in $\mathbb{R}^2$ have 6 distances that should satisfy a polynomial equation saying that the tetrahedron with these 6 edge lengths has volume 0. So a randomly sampled matrix of potential distances for $m > n+1$ is unlikely to be realizable by a cloud of m ordered points in $\mathbb{R}^n$.

Chapter 3 in (Liberti & Lavor, 2017) discusses realizations of a complete graph given by a distance matrix in $\mathbb{R}^n$. Lemma D.3(a) and later results hold for all clouds including degenerate ones, e.g. for 3 points in a straight line.

Any points $p_1, \dots, p_{n-1} \in A$ have $\mathrm{aff}(p_1, \dots, p_{n-1}) \le n-2$. For example, any two distinct points in $A \subset \mathbb{R}^3$ generate a straight line. In $\mathbb{R}^2$, any point $p_1 \ne O(A)$ forms a suitable $\{p_1\}$. In $\mathbb{R}^3$, one can choose any distinct points $p_1, p_2 \in A$ so that the infinite straight line via p_1, p_2 avoids $O(A)$.

If there are no such p_1, p_2, then $A \subset \mathbb{R}^3$ is contained in a straight line L, so $\mathrm{aff}(A) = 1$. In this degenerate case, the stronger condition $\mathrm{aff}(O(A) \cup \{p_1, \dots, p_{n-1}\}) = \mathrm{aff}(A)$ will help reconstruct $A \subset L$ by using any point $p_1 \ne O(A)$. The first step is to reconstruct any ordered sequence from its distance matrix in Lemma D.3(a).

Lemma D.3(a) holds for all degenerate clouds, e.g. for three points are in a straight line.

Lemma D.3 (reconstruction of ordered points)**.** ***(a)*** *Any sequence of ordered points $A = (p_1, \dots, p_m)$ in $\mathbb{R}^n$ can be reconstructed (uniquely up to isometry) from the matrix of the Euclidean distances $|p_i - p_j|$ in time $O(m^3)$. If all distances are divided by $R = \max_{i=1,\dots,m} |p_i|$, the reconstruction of $A \subset \mathbb{R}^n$ is unique up to isometry and uniform scaling.*

(b) *If $m \le n$, the uniqueness of reconstructions in part (a) holds if we replace isometry with rigid motion. Hence any $n-1$ ordered points $p_1, \dots, p_{n-1}$ can be uniquely reconstructed from all pairwise distances between $0, p_1, \dots, p_{n-1}$ up to $\mathrm{SO}(n)$ rotation around the origin $0 \in \mathbb{R}^n$.*

Proof of Lemma D.3. **(a)** By translation, we can put p_1 at the origin $0 \in \mathbb{R}^n$. Let GM be the $(m-1) \times (m-1)$ matrix $g_{ij} = \dfrac{p_i^2 + p_j^2 - |p_i - p_j|^2}{2} = p_i \cdot p_j$ constructed from squared distances between $p_1 = 0, \dots, p_m$ for $i, j = 2, \dots, m$. By Lemma D.2(b) if GM has $k \le n$ positive eigenvalues, then $p_1 = 0, \dots, p_m$ can be uniquely determined up to isometry in $\mathbb{R}^k \subset \mathbb{R}^n$ in time $O(m^3)$. If all distances are divided by the same radius R, the above construction guarantees uniqueness up to isometry and uniform scaling.

(b) If $m \le n$, any mirror image of $A \subset \mathbb{R}^n$ after a suitable rigid motion in $\mathbb{R}^n$ can be assumed to belong to an $(n-1)$-dimensional hyperspace $H \subset \mathbb{R}^n$, where they are matched by a mirror reflection $H \to H$ with respect to an $(n-2)$-dimensional subspace $S \subset H$. This reflection is realized by the $\mathrm{SO}(n)$ rotation through $180°$ around S. □

Lemma D.3(b) for $m = n = 3$ implies that any triangle is determined by its sides up to rigid motion in $\mathbb{R}^3$. For example, the sides $3, 4, 5$ define a right-angled triangle whose mirror images are not related by rigid motion inside a plane $H \subset \mathbb{R}^3$, but are matched by composing a suitable rigid motion in H and a $180°$ rotation of $\mathbb{R}^3$ around a line in H.

Lemma D.4 (time of determinant)**.** *Any $n \times n$ determinant can be computed in time $O(n^3)$ with space $O(n^3)$.*

***Proof of Lemma** D.4.* Any $n \times n$ determinant can be computed by Gaussian elimination in time $O(n^3)$ with space $O(n^3)$, see (Bunch & Hopcroft, 1974). The more recent theoretical estimate is $O(n^{2.373})$ by (Fisikopoulos & Penaranda, 2016). □

***Proof of Theorem** D.1(a).* Any rigid motion of $\mathbb{R}^n$ mapping a Euclidean graph $G \subset \mathbb{R}^n$ to another graph F is a bijection preserving distances and signs of determinants, and hence induces a bijection $\mathrm{CBR}(G; p_1, \dots, p_i) \to \mathrm{CBR}(F; q_1, \dots, q_i)$ for all $p_1, \dots, p_i \in V(G)$ and corresponding vertices $q_1, \dots, q_i \in V(F)$ for any $i = 1, \dots, h$, which implies a bijection $\mathrm{NCD}(G; h) \to \mathrm{NCD}(F; h)$. By Definition 3.5, if G has m unordered vertices, the $\mathrm{NCD}(G)$ consists of $m(m-1)\dots(m-h+1) = O(m^h)$ Centered Base Representations $\mathrm{CBR}(G; A)$ for all base sequences $A \subset V(G)$ of h ordered vertices.

Every $\mathrm{CBR}(G; A)$ consists of the three components $\mathrm{sign}(A), \mathrm{CD}(A), \mathrm{CR}(G; A)$. For $h = n$, $\mathrm{sign}(A)$ is the $n \times n$ determinant computable in time $O(n^3)$ with space $O(n^3)$ by Lemma D.4. The distance matrix $\mathrm{CD}(A)$ needs $O(h^2)$ time and $O(h^2)$ space. The $(h+1) \times (m-h)$ matrix $\mathrm{CR}(G; A)$ has $O(hm)$ distances, each computable in time $O(n)$. So $\mathrm{CBR}(G; A)$ can be computed in time $O(n^2 m)$ with space $O(n^3 + hm)$, where $n \le m$. Multiplying these complexities by the number $O(m^h)$ of base sequences gives the final time $O(n^2 m^{h+1})$ and space $O(n^3 + hm^{h+1})$ for $\mathrm{NCD}(G)$. □

The proof of Theorem D.1(b) will use the fact that any point in $\mathbb{R}^n$ is uniquely determined by $n+1$ distances to $n+1$ ordered points that affinely span $\mathbb{R}^n$, and also Lemma D.5.

Lemma D.5 (equal CBRs)**.** *Let a Euclidean graph $G \subset \mathbb{R}^n$ have the vertex set $V(G)$ with the center of mass at $p_0 = 0 \in \mathbb{R}^n$. Let $n-1$ ordered vertices $p_1, \dots, p_{n-1}$ linearly span an $(n-1)$-dimensional subspace $S \subset \mathbb{R}^n$. Let $G(p_1, \dots, p_{n-1})$ be the subgraph of G on the vertex set $V(G)$ and all edges of G at $p_1, \dots, p_{n-1}$. For any other vertex p, let $\mathrm{CBR}'(G; p_1, \dots, p_{n-1}, p)$ be obtained from the Centered Base Representation $\mathrm{BR}(G; p_1, \dots, p_{n-1}, p)$ by removing signs of distances from all vertices $q \in V(G) \setminus \{p_1, \dots, p_{n-1}, p\}$ to p. If $\mathrm{BR}'(G; p_1, \dots, p_{n-1}, p) = \mathrm{BR}'(G; p_1, \dots, p_{n-1}, p')$ for some vertices $p, p' \in V(G) \setminus \{p_1, \dots, p_{n-1}\}$, the mirror reflection with respect to S maps $G(p_1, \dots, p_{n-1})$ to itself and p to p'.*

***Proof of Lemma** D.5.* Under the reflection f_S of $\mathbb{R}^n$ with respect to the subspace $S \subset \mathbb{R}^n$, the vertices p, p' should be swapped because they have equal (signed) distances to the ordered points $p_0, \dots, p_{n-1} \in S$. The equality of given CBR′s means that $V' = V(G) \setminus \{p_1, \dots, p_{n-1}, p, p'\}$ bijectively maps to itself via $q \mapsto q'$ so that any matched q, q' have the same distances to the $n+1$ ordered points $p_0, \dots, p_{n-1}, p$ as to $p_0, \dots, p_{n-1}, p'$, respectively. Any point in $\mathbb{R}^n$ is determined by its distances to the n affinely independent points $p_0, \dots, p_{n-1}$ up to the mirror reflection f_S. Since f_S fixes $p_0, \dots, p_{n-1}$, the reflection f_S should swap q, q' in such pairs and all their edges, so we conclude that $f_S(G(p_1, \dots, p_{n-1})) = G(p_1, \dots, p_{n-1})$ and $f_S(p) = p'$. □

***Proof of Theorem** D.1(b).* The completeness is proved by reconstructing any Euclidean graph $G \subset \mathbb{R}^n$ from $\mathrm{NCD}(G; n)$ uniquely up to rigid motion.

We prove that any Euclidean graph $G \subset \mathbb{R}^n$ can be reconstructed from its Nested Distance Distribution $\mathrm{NCD}(G; n)$ by induction on the dimension n.

The inductive base $n = 1$ is Example 4.7. Assume that any graph G on m unordered vertices can be reconstructed in $\mathbb{R}^k$ in time $O(k^3 m)$ for any $k < n$. Below we prove the inductive step for the dimension $n > 1$. Start from any $\mathrm{CBR}(G; A) = [\mathrm{sign}(A), \mathrm{CD}(A), \mathrm{CR}(G; A)]$ from Definition 3.5, where A is a sequence of some n ordered (not yet geometrically fixed) vertices $p_0, \dots, p_n \in V(G)$. The first point p_0 is fixed at the origin $0 \in \mathbb{R}^n$ as usual by translation.

Lemma D.2(b) for the matrix $\mathrm{CD}(A)$ gives the number $k \le n$ of positive eigenvalues of the Gram matrix of the n vectors $p_1, \dots, p_n$ in time $O(n^3)$. If $\mathrm{aff}(A) = k < n$, we use the nested structure of $\mathrm{NCD}(G; n)$ to take another $\mathrm{CBR}(G; p_1, \dots, p_k, q, \dots, p_n)$ for a new vertex $q \in V(G) - A$. Check if $\mathrm{aff}(p_1, \dots, p_k, q) = k+1$ again by Lemma D.2(b) using the matrix $D(p_1, \dots, p_k, q)$. If the affine dimension has not increased, we take another CBR with the same points $p_1, \dots, p_k$ and a new $(k+1)$-st point from $V(G) - \{A \cup q\}$ and so on.

This search through Centered Base Representations involving the remaining vertices of G requires a maximum of $m - n - 1$ steps with $O(n^3)$ time for every computation of the affine dimension. Hence in time $O(n^3 m)$, we can find a Centered Base

Representation $\mathrm{CBR}(G;A)$ whose base sequence A affinely generates the subspace of dimension $k=\mathrm{aff}(V(G))$ in $\mathbb{R}^n$. If $k<n$, the proof follows from the inductive hypothesis for the smaller dimension k.

If $\mathrm{aff}(V(G))=n$, use the same notations for the fixed vertices $0=p_0,\dots,p_n$ that linearly generate $\mathbb{R}^n$. Lemma D.3(a) for $m=n+1$ and the distance matrix $D(A)$ allow us to reconstruct $n+1$ ordered points $0=p_0,\dots,p_n$ up to isometry in $\mathbb{R}^n$ in time $O(n^3)$. By Definition 3.5 every column of $\mathrm{CR}(G;A)$ contains Euclidean distances from the vertices $0=p_0,\dots,p_n\in\mathbb{R}^n$, which affinely generate $\mathbb{R}^n$, to another vertex $q\in V(G)-A$.

These $n+1$ distances uniquely determine the position of q in $\mathbb{R}^n$ whose coordinates can be found as follows. Each scalar product $q\cdot p_i$ can be computed as $|q|\cdot|p_i|\cos\angle(q,0,p_i)=\dfrac{|q|^2+|p_i|^2-|q-p_i|^2}{2}$ for $i=1,\dots,n$. On another hand, $q\cdot p_i$ is a linear combination of unknown coordinates of q with coefficients equal to the coordinates of p_i. One can find all coordinates of q in time $O(n^3)$ by solving the system of linear equations, where the $n\times n$ determinant on the linear basis $p_1,\dots,p_n$ is not zero. The total time is $O(n^3m)$.

Since all vertices $q\in V(G)-A$ are geometrically unique, they can be (arbitrarily) ordered, say $p_{n+1},\dots,p_m$, following $p_0,\dots,p_n$. The signs of distances in the matrix $\mathrm{CR}(G;A)$ also tell us about (present or absent) edges from $p_0,\dots,p_n$ to all other vertices $q\in V(G)-A$.

The nested structure of $\mathrm{NCD}(G;n)$ allows us to consider $m-n$ unordered Base Representations $\mathrm{CBR}(G;p_1,\dots,p_{n-1},p_j)$ for all vertices p_j with $j=n,\dots,m$. Every vertex $p_j\in V(G)$ is uniquely determined in $\mathbb{R}^n$ by the column of its signed distances to $p_0,\dots,p_n$ in the $(n+1)\times(m-n-1)$ matrix $R(G;p_0,\dots,p_n)$ for $j=n+1,\dots,m$.

By Lemma D.5, this distance list of p_j (without edges between p_j,p_k for $j,k>n$) suffices to identify one or maximum two Base Representations among all $m-n$ unordered CBRs with the fixed n points $p_0,\dots,p_{n-1}$ and variable n-th vertices. If there is a choice of two CBRs, we can take any of them for p_j. Indeed, choosing another vertex p_k, which should be mirror symmetric to p_j, will produce a mirror image of the reconstructed subgraph $G(p_1,\dots,p_n,p_j)$ by Lemma D.5.

The matrix $\mathrm{CR}(G;p_1,\dots,p_{n-1},p_j)$ from the found CBRs contains signs that determine the (present or absent) edges from p_j to all other vertices p_k for $k=n+1,\dots,m$.

To guarantee the uniqueness of $G\subset\mathbb{R}^n$ under rigid motion and not only under isometry, we additionally use $\mathrm{sign}(p_1,\dots,p_n)$ from CBR to fix an orientation of the simplex on $p_0,\dots,p_n$. □

The strength $\sigma(A)$ depends only on the distance matrix $D(A)$, we write $\sigma(A)$ for brevity. When the simplex on A degenerates, the strength $\sigma(A)$ vanishes and is Lipschitz continuous by Lemma 4.2, while the volume of the simplex on B is not Lipschitz continuous as shown below.

In $\mathbb{R}^2$, consider the triangle with two vertices fixed at $(\pm l,0)$ and one moving vertex $(0,t\varepsilon)$ for $t\in[-1,1]$. The signed area of the triangle changes from $-l\varepsilon$ (unbounded because l can be large for any fixed small ε) to 0 (when $t=0$ and the triangle degenerates), then to $l\varepsilon$ (when $t=1$). The area changes by $2l\varepsilon$ while only one vertex moves by 2ε, so the ratio of the area change over a point perturbation can be as large as a half-distance between given points.

Lemma D.6 (time of strength)**.** *For any base sequence A of n ordered points $p_1,\dots,p_n\in\mathbb{R}^n$, the strength $\sigma(A)$ can be computed in time $O(n^3)$.*

***Proof of Lemma** D.6.* The half-perimeter $p(A)$ is computable via all pairwise distances in time $O(n^2)$. The squared volume $\mathrm{vol}^2(A)$ can be expressed by the Cayley-Menger $(n+2)\times(n+2)$ determinant from (Sippl & Scheraga, 1986) in inter-point distances, which can be computed in time $O(n^3)$ by Lemma D.4. □

Lemma D.7 (axioms and time of M_∞ on CBRs)**.** *Let $G,F\subset\mathbb{R}^n$ be Euclidean graphs with m unordered vertices and base sequences $A\subset V(G)$ and $B\subset V(F)$ of $h\le n$ ordered vertices. The metric $M_\infty(\mathrm{CBR}(G;A),\mathrm{CBR}(F;B))$ from Definition B.5 satisfies all metric axioms and is computable in time $O(h^2+m^{1.5}\log^{h+1}m)$ with space $O(h^2+m\log^{h-1}m)$.*

***Proof of Lemma** D.7.* The metric axioms for M_∞ follow from the same axioms for the metrics L_∞ and W_∞ because the maximum of metrics is still a metric, see metric transforms in section 4.1 of (Deza & Deza, 2009). The first metric $\dfrac{2}{\lambda_n}|\mathrm{sign}(A)\sigma(A)-\mathrm{sign}(B)\sigma(B)|$ can be computed in time $O(n^3)$ by Lemma D.6. The metric $L_\infty(\mathrm{CD}(A),\mathrm{CD}(B))$

requires time $O(h^2)$ and space $O(h^2)$. The bottleneck distance $W_\infty(\mathrm{CR}(G;A)), \mathrm{CR}(F;B))$ between $(h+1)\times(m-h)$ matrices $\mathrm{CR}(G;A)$, $\mathrm{CR}(F;B)$ with unordered columns (considered as clouds of $m-h$ unordered points in $\mathbb{R}^{h+1}$) needs time $O(m^{1.5}\log^{h+1} m)$ and space $O(m\log^{h-1} m)$ by Theorem 6.5 in (Efrat et al., 2001). □

Lemma D.8 (metric axioms for NBM on NCDs)**.** *The Nested Bottleneck Metric* NBM *from Definition B.6 satisfies all metric axioms on Nested Distance Distributions.*

Proof of Lemma D.8. Induction on the depth $i = n, \dots, 1$. The inductive base $i = n$ follows from the metric axioms in Lemma D.7 for M_∞ in Definition B.5.

The inductive step from a depth i (between $1, n$) to the smaller value $i-1$ follows from Lemma C.1 and the metric axioms in the inductive hypothesis for the depth i. □

Lemma D.9 (Lipschitz continuity of M_∞)**.** *Let A be a base sequence of $1 \le h \le n$ ordered vertices in a Euclidean graph $G \subset \mathbb{R}^n$. Let B, F be obtained from A, G, respectively, by perturbing every vertex of G within its ε-neighborhood in $\mathbb{R}^n$. Then $\mathrm{CBR}(G;A)$ changes in M_∞ from Definition B.5 by at most 2ε, so $M_\infty(\mathrm{CBR}(G;A), \mathrm{CBR}(F;B)) \le 2\varepsilon$.*

Proof of Lemma D.9. Order all vertices of the graphs G, F so that every vertex $p_i \in V(G)$ has the same index as its perturbation $q_i \in V(F)$. The bijection $p_i \leftrightarrow q_i$ induces the bijections between the corresponding elements of the matrices $\mathrm{CD}(A) \leftrightarrow \mathrm{CD}(B)$ and $\mathrm{CR}(G;A) \leftrightarrow \mathrm{CR}(F;B)$, which all differ by at most 2ε. Lemma 4.2 implies that $\frac{2}{\lambda_n}|\mathrm{sign}(A)\sigma(A) - \mathrm{sign}(B)\sigma(B)| \le 2\varepsilon$ Since all three components of the max metric M_∞ in Definition B.5 have the upper bound 2ε, conclude that $M_\infty \le 2\varepsilon$. □

Definition C.2 can be adapted for the EMD between NCDs by (1) replacing the bottleneck distance W_∞ in Definition B.5 with EMD between clouds of equally weighted points, and (2) replacing $\mathrm{BMD}(\Gamma)$ for a bipartite graph Γ with $\mathrm{EMD}(\Gamma)$ between the unordered sets (of potentially different sizes) of CBDs with weights on all white vertices and CBDs on all black vertices.

Proof of Theorem D.1(c). We first prove the Lipschitz continuity of the metric NBM on NCDs. Order all vertices of the graphs G, F so that every $p_i \in V(G)$ has the same index as its ε-perturbation $q_i \in V(F)$. In Definition B.6, for any base sequence A of $p_1, \dots, p_h \in V(G)$, there is a base sequence B of vertices $q_1, \dots, q_h \in V(F)$, which are ε-perturbations of $p_1, \dots, p_h$, respectively, such that $M_\infty(\mathrm{CBR}(G;A), \mathrm{CBR}(F;B)) \le 2\varepsilon$ by Lemma D.9.

These distances M_∞ are weights of edges in the index-preserving vertex matching E of the complete bipartite graph $\Gamma(G; p_1, \dots, p_{h-1}; F; q_1, \dots, q_{h-1})$ for any $p_1, \dots, p_{h-1}$ and their ε-perturbations $q_1, \dots, q_{h-1}$. Then $\mathrm{BMD}(\Gamma(G; p_1, \dots, p_{h-1}; F; q_1, \dots, q_{h-1})) \le 2\varepsilon$ by Lemma C.3. Since this conclusion holds for all (choices of) $p_1, \dots, p_{h-1} \in V(G)$, we iteratively apply this argument for the bipartite graphs $\Gamma(G; p_1, \dots, p_{i-1}; F; q_1, \dots, q_{i-1})$ for $1 \le i < n$ and finally conclude that $\mathrm{NBM}(\mathrm{NCD}(G;h), \mathrm{NCD}(F;h)) \le 2\varepsilon$. The proof that $\mathrm{EMD}(\mathrm{NCD}(G;h), \mathrm{NCD}(F;h)) \le 2\varepsilon$ is similar by using Lemma C.4 instead of C.3. □

Proof of Theorem D.1(d). In Definition B.6, for any fixed $1 \le i \le h$ and ordered vertices $p_1 \dots, p_{i-1} \in V(G)$ and $q_1 \dots, q_{i-1} \in V(F)$, the complete bipartite graph $\Gamma(G; p_1, \dots, p_{i-1}; F; q_1, \dots, q_{i-1})$ has $V = 2(m-i+1) = O(m)$ vertices and $E = (m-i+1)^2 = O(m^2)$ edges.

For $i = h$, the weight $w(e)$ of each edge e equals M_∞, which needs time $O(m^{1.5}\log^{h+1} m)$ and space $O(m\log^{h-1} m)$ by Lemma D.7 for any $h \le n \le m$. For all $O(m^2)$ edges of $\Gamma(G; p_1, \dots, p_{h-1}; F; q_1, \dots, q_{h-1})$, the time is $O(m^{3.5}\log^{h+1} m)$, the space is $O(m^3\log^{h-1} m)$. The bottleneck matching distance BMD for such a graph is computed by (Hopcroft & Karp, 1973) in time $O(E\sqrt{V}) = O(m^{2.5})$, which is dominated by the time $O(m^{3.5}\log^{h+1} m)$ preparing the weighted graph.

For all $O(m^{2(h-1)})$ choices of ordered vertices $p_1, \dots, p_{h-1} \in V(G)$ and $q_1, \dots, q_{h-1} \in V(F)$, the Bottleneck Matching Distance for all graphs $\Gamma(G; p_1, \dots, p_{h-1}; F; q_1, \dots, q_{h-1})$ are found in time

$$O(m^{2(h-1)})O(m^{3.5}\log^{h+1} m) = O(m^{2h+1.5}\log^{h+1} m)$$

with space $O(m^{2h+1}\log^{h-1} m)$. For every next iteration $i = h-2,\dots,1$, the parameter i goes down by 1 every time. We can compute all distances $\mathrm{BMD}(\Gamma(G;p_1,\dots,p_{i-1};F;q_1,\dots,q_{i-1})$ in time

$$O(m^{2(i-1)})O(m^{3.5}\log^{h+1} m) = O(m^{2i+1.5}\log^{h+1} m).$$

The sum of all these times for $i = 1,\dots,h-1$ is still $O(m^{2h+1.5}\log^{h+1} m)$ from the first step.

All CBDs in Definition 3.5 have sizes at most m, which is the maximum number of points in the given clouds. The EMD between weighted distributions of a maximum size m can be computed in near-cubic time $O(m^3\log m)$, see (Fredman & Tarjan, 1987; Goldberg & Tarjan, 1987). Since this complexity is dominated by the time $O(m^{3.5}\log^{h+1} m)$ for computing $O(m^2)$ weights M_∞, each in time $O(m^{1.5}\log^{h+1} m)$ by Lemma D.7, the total time for the EMD is the same as for the NBM, similarly for space complexities □