Incompleteness of graph neural networks for points clouds in three dimensions

To cite this article: Sergey N Pozdnyakov and Michele Ceriotti 2022 *Mach. Learn.: Sci. Technol.* **3** 045020

View the article online for updates and enhancements.

You may also like

MACHINE
LEARNING
Science and Technology

PAPER

Incompleteness of graph neural networks for points clouds in three dimensions

Sergey N Pozdnyakov and Michele Ceriotti*

Laboratory of Computational Science and Modelling, Institute of Materials, Ecole Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland
* Author to whom any correspondence should be addressed.
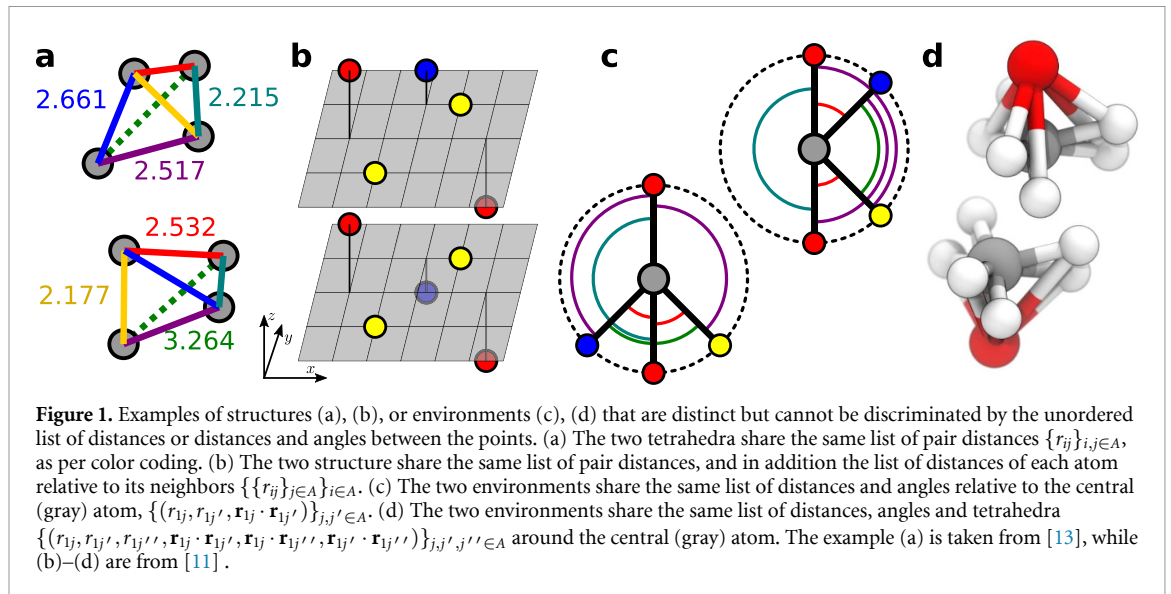
E-mail: michele.ceriotti@epfl.ch

Abstract

Graph neural networks (GNN) are very popular methods in machine learning and have been applied very successfully to the prediction of the properties of molecules and materials. First-order GNNs are well known to be incomplete, i.e. there exist graphs that are distinct but appear identical when seen through the lens of the GNN. More complicated schemes have thus been designed to increase their resolving power. Applications to molecules (and more generally, point clouds), however, add a geometric dimension to the problem. The most straightforward and prevalent approach to construct graph representation for molecules regards atoms as vertices in a graph and draws a bond between each pair of atoms within a chosen cutoff. Bonds can be decorated with the distance between atoms, and the resulting 'distance graph NNs' (dGNN) have empirically demonstrated excellent resolving power and are widely used in chemical ML, with all known indistinguishable configurations being resolved in the fully-connected limit, which is equivalent to infinite or sufficiently large cutoff. Here we present a counterexample that proves that dGNNs are not complete even for the restricted case of fully-connected graphs induced by 3D atom clouds. We construct pairs of distinct point clouds whose associated graphs are, for any cutoff radius, equivalent based on a first-order Weisfeiler-Lehman (WL) test. This class of degenerate structures includes chemically-plausible configurations, both for isolated structures and for infinite structures that are periodic in 1, 2, and 3 dimensions. The existence of indistinguishable configurations sets an ultimate limit to the expressive power of some of the well-established GNN architectures for atomistic machine learning. Models that explicitly use angular or directional information in the description of atomic environments can resolve this class of degeneracies.

1. Introduction

Point clouds can be used to provide an abstract description of shapes, and objects across different length scales [1, 2], and have been widely used in the construction of machine-learning models for computer vision [3], remote sensing [4] and autonomous driving [5]. A description in terms of an unordered set of points is relevant also at the atomic scale, where molecules and materials are most naturally characterized in terms of the position and nature of their atomic constituents. The list of Cartesian coordinates of points, however, does not reflect the fact that the properties associated with a given structure are usually invariant, or equivariant, to symmetry operations such as rigid translations, rotations, or permutation of the ordering of the points. In the context of atomistic simulations, the problem of describing a structure in a symmetry-adapted manner has been a central concern in early applications of machine-learning models [6–8], and has since given rise to the development of a large number of *representations* that attempt to characterize fully a structure while simultaneously fulfilling the requirements of symmetry [9].

In fact, deep connections are present between most of the existing frameworks [10], that differ in implementation details but can be understood as describing structures in terms of unordered lists of
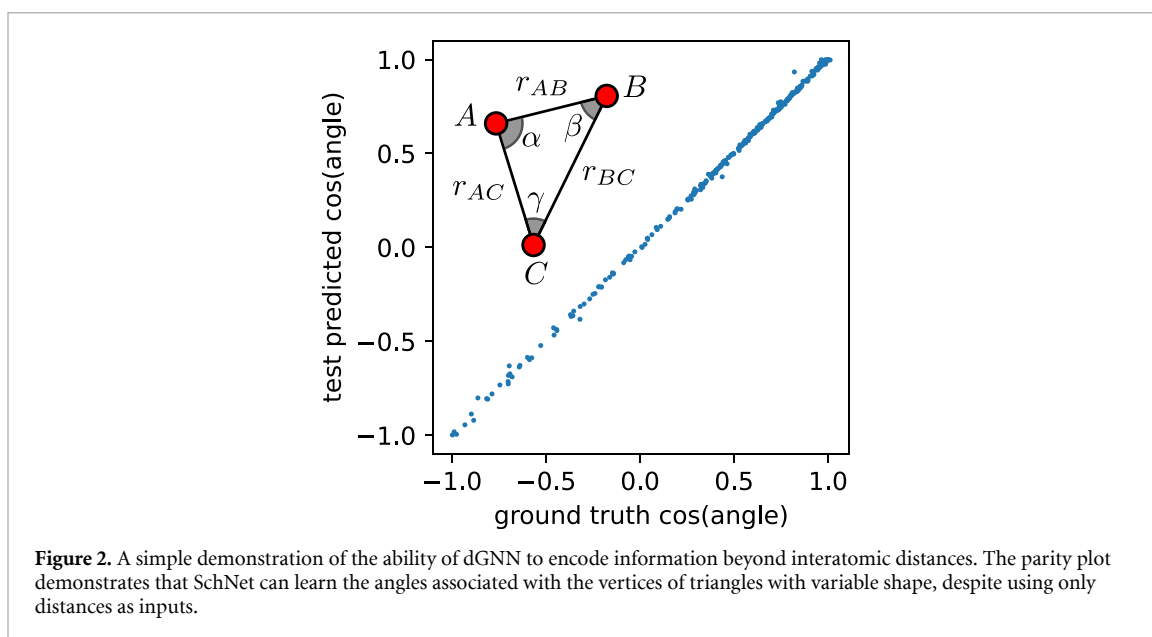
Figure 1. Examples of structures (a), (b), or environments (c), (d) that are distinct but cannot be discriminated by the unordered list of distances or distances and angles between the points. (a) The two tetrahedra share the same list of pair distances $\{r_{ij}\}_{i,j\in A}$, as per color coding. (b) The two structure share the same list of pair distances, and in addition the list of distances of each atom relative to its neighbors $\{\{r_{ij}\}_{j\in A}\}_{i\in A}$. (c) The two environments share the same list of distances and angles relative to the central (gray) atom, $\{(r_{1j}, r_{1j'}, \mathbf{r}_{1j} \cdot \mathbf{r}_{1j'})\}_{j,j'\in A}$. (d) The two environments share the same list of distances, angles and tetrahedra $\{(r_{1j}, r_{1j'}, r_{1j''}, \mathbf{r}_{1j} \cdot \mathbf{r}_{1j'}, \mathbf{r}_{1j} \cdot \mathbf{r}_{1j''}, \mathbf{r}_{1j'} \cdot \mathbf{r}_{1j''})\}_{j,j',j''\in A}$ around the central (gray) atom. The example (a) is taken from [13], while (b)–(d) are from [11].

distances, angles, tetrahedra, etc within atom-centered environments, corresponding to correlations between 2, 3, 4, ν neighbors of the central atom. This systematic study has revealed that—at least for low values of ν—atom-centered representations are *incomplete*, i.e. there are pairs of structures that are distinct, but contain environments that are indistinguishable based on the unordered list of distances and angles [11], affecting also the ability of the representation to resolve local deformations of certain structures [12].

This issue is also closely related to classical problems in invariant theory [14], that aim to determine under which conditions two point clouds can be unequivocally identified by the unordered list of distances, or distances and angles. If a set of features cannot discriminate two structures, any metric [15, 16] built on those features will not be able to distinguish them. Furthermore any model built on those features—no matter how sophisticated—will be unable to learn their properties or to classify them in different categories, and will be limited in its expressive power. The actual impact of these degeneracies on machine-learning models of atomic and materials properties is mitigated by the fact that structures that contain problematic environments usually also contain others that are not degenerate, and so a model that *simultaneously* uses information on all centers can still distinguish the structures [17]. For instance, all structures in figure 1 except for (b) can be discriminated by knowledge of the set of unordered lists of distances $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ around each point (i.e. taking $\{\{r_{ij}\}_{j\in A}\}_{i\in A}$). Still, there is evidence that the incompleteness of atom-density representations does affect the ultimate performance of the machine-learning models built on them [11]. This is therefore an issue of practical, and not only theoretical, relevance, which is one of the factors driving a transition towards systematically-improvable features with higher correlation order [10, 18–20].

In the broader context of models based on point clouds, graph neural networks (GNN) have been used extensively to describe the relative arrangement of points, mapping the cloud onto a graph [3, 21]. In the case of atomistic models, most well-established GNN frameworks treat atoms as the vertices of the graph, labeled by their chemical identity, and the edges are associated with the distances between them [22–24]. Even though this class of 'distance graph' NN (dGNN) only uses information on the distances between points, the way this is combined in the subsequent steps of the network allows for very flexible models with considerable descriptive power. One simple but compelling example is given in figure 2. By applying a dGNN to a set of triangular configurations, one can predict the angles even though the network uses only distance information. As we shall see, this is not generally true: for configurations with many points it is not always possible to infer angular information using a dGNN.

It is well-known that for general discrete graphs it is possible to build pairs of items that are distinct, and yet indistinguishable by most traditional GNN [25]. This has triggered the development of higher-order graph networks [26, 27]. Examples of actual chemical structures that cannot be distinguished by a dGNN can also be obtained if one considers a finite cutoff [28, 29]. To the best of our knowledge, however, no examples have been shown of 3D point clouds whose *fully-connected*, distance-labeled graphs cannot be resolved by a dGNN. In fact, with a sufficiently long cutoff in the construction of the graph, dGNNs can discriminate between all examples of environments that are degenerate under $\nu = 1, 2, 3$ atom-centered descriptors shown in figure 1, as well as those in [28, 29], suggesting a higher descriptive power than atom-centered descriptors.

Figure 2. A simple demonstration of the ability of dGNN to encode information beyond interatomic distances. The parity plot demonstrates that SchNet can learn the angles associated with the vertices of triangles with variable shape, despite using only distances as inputs.

Here we present a class of point clouds involving pairs of structures that are not distinguishable by dGNN, irrespective of the cutoff chosen in the construction of the molecular graph. We show that this construction includes also configurations that correspond to realistic chemical structures, and demonstrate how for these configurations there is a limit to the accuracy that can be reached by this class of models.

2. Weisfeiler-Lehman test for geometric GNNs

The Weisfeiler-Lehman (WL) test is a well-known graph-theoretical procedure [30, 31] that provides a sufficient condition for two graphs being distinct, and that has been shown to be equivalent to an assessment of the resolving power of several classes of GNNs [25, 32, 33]. Here we use a version of the test that incorporates explicitly the distances between nodes in the construction of the edge identifiers, and is similar in spirit to the construction of graph kernels in [34]. Given a set of n nodes with labels, $\{l_i\}_{i=1...n}$, and distances between them and their neighbors $\{r_{ij}\}$, we construct fingerprints of each node as the multiset of label/distance pairs, supplemented with a label for the node

$$h_i = \text{hash}(l_i, \{\{(l_j, r_{ij})\}\}_{j=1...n}). \tag{1}$$

The hash function assigns a unique identifier to distinct multisets. The values of the hashes are then used to re-label the nodes, and the procedure is iterated: the process of encoding in the node labels the structure of the neighborhood and then iterating the procedure has a clear analogy with message passing constructs. Returning to the example in figure 2, it should now be clear how, for a triangle, a dGNN can learn angular information. After two iterations, the node descriptors for each of the vertices contain information not only on the two distances to the first neighbors (e.g. for vertex A, r_{AB} and r_{AC}), but also on the fact that the neighbors have a further neighbor at a certain distance (e.g. r_{BC}), which is all the information needed to reconstruct the angles. Usually (but not always) the iterative procedure converges to fixed values of the vertex labels; if the multisets of hashes that characterize two graphs are identical, the graphs cannot be distinguished by this distance-decorated WL test, and will not be discriminated by a geometric dGNN that uses only point labels and pair distances to characterize graph neighborhoods.

 Figure 3 shows an example, adapted from [25], of a pair of configurations whose distance-decorated graphs cannot be distinguished by a WL test *provided that the graph only includes first-neighbor distances*. Several analogous examples have been shown for different types of CNNs [28, 29], where the use of a finite cutoff in the construction of the molecular graph affects the resolving power of the network. As shown in the figure, however, increasing the cutoff distance to include further nodes in the definition of the graph is usually sufficient to distinguish the structures. In what follows, we will discuss a counterexample that cannot be resolved by simply using a bigger cutoff: a family of structures that fail the WL test even when considering their fully-connected molecular graph.
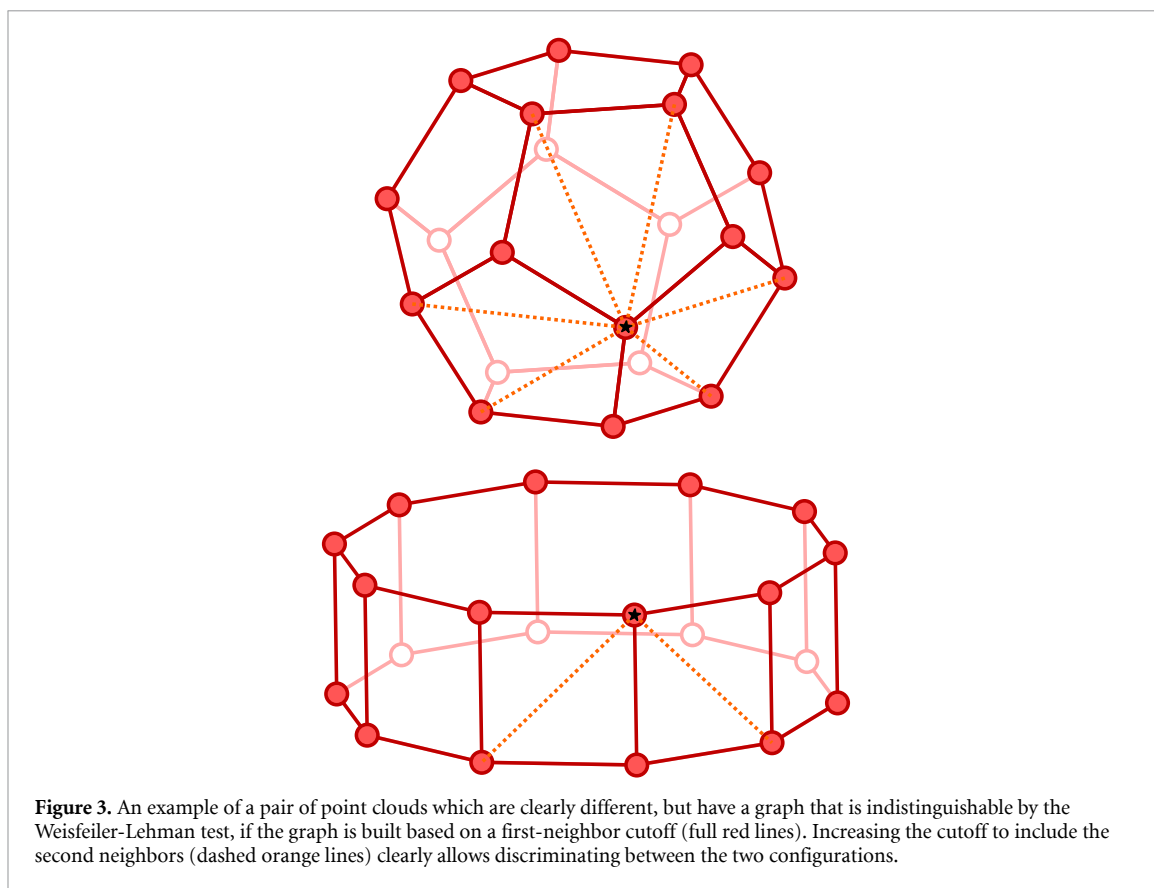
Figure 3. An example of a pair of point clouds which are clearly different, but have a graph that is indistinguishable by the Weisfeiler-Lehman test, if the graph is built based on a first-neighbor cutoff (full red lines). Increasing the cutoff to include the second neighbors (dashed orange lines) clearly allows discriminating between the two configurations.
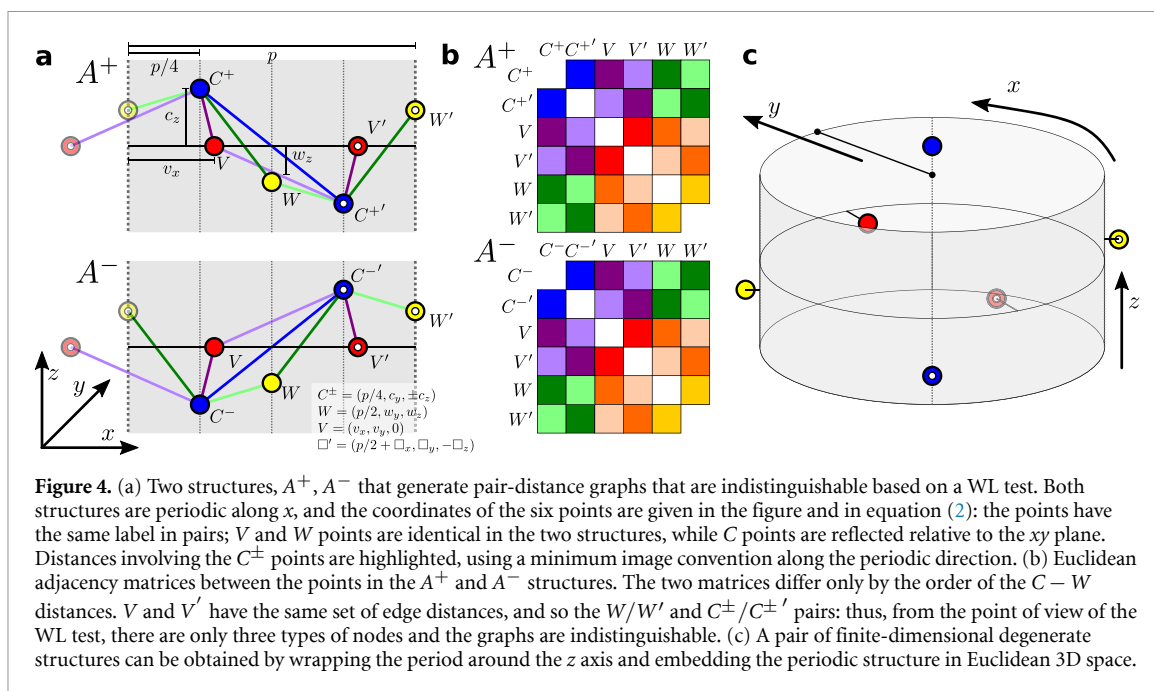
3. A counterexample

Consider the construction in figure 4(a). Six points, with labels that are identical in pairs, are arranged following the pattern in the figure, forming two structures, A^+ and A^- with

$$
\begin{aligned}
C^\pm &= (p/4, c_y, \pm c_z) \\
W &= (p/2, w_y, w_z) \\
V &= (v_x, v_y, 0) \\
\square' &= (p/2 + \square_x, \square_y, -\square_z)
\end{aligned}
\tag{2}
$$

where the last definition indicates the relation between plain and primed points. The structures are periodic along the x axis, with a period p, and have open boundary conditions along y and z. After one iteration of the modified WL procedure discussed in section 2, one sees that the unordered set of distances for V/V', W/W', C/C' pairs are identical, so that the plain and primed points will receive the same hash value, and the two graphs cannot be discriminated by the test despite being different (figure 4(b)). The essential ingredient to induce a degeneracy is the swapping of the distances between C and W points, i.e. that $d(C^+, W) = d(C^-, W')$ and $d(C^+, W') = d(C^-, W)$, which changes the geometry of the structures but does not affect the outcome of the WL test, that relies only on the unordered set of distances. This swap, which can be clearly seen by plotting the distance matrix, is a consequence of the regular spacing of the points along the x axis and the periodic boundary conditions. Distances between the C points and the W points to their left in the A^+ structure correspond to distances between the C points and the W points to their right in the A^- configuration. A more thorough discussion of the role of periodicity is also given in appendix A, which deals with the 3D-periodic case.

It is important to stress that even though we only show minimum-image distances in the figure, the graphs generated by the periodic structure are degenerate even when considering a fully-connected graph, including all distances between periodic replicas. Thus, the effectiveness of the counterexample is independent on the details of the definition of the graph neighborhoods, and it equally affects cases in which one only considers the neighbors within a fixed cutoff distance, or just the first k neighbors. The WL procedure is a very powerful test to differentiate graphs based on first-order connectivity information. Failure in distinguishing graphs based on the WL test indicates that the pairs are equivalent for a broad class of pair-distance GNNs [22], that includes SchNet [23], convolutional-networks molecular fingerprints [35],

Figure 4. (a) Two structures, A^+, A^- that generate pair-distance graphs that are indistinguishable based on a WL test. Both structures are periodic along x, and the coordinates of the six points are given in the figure and in equation (2): the points have the same label in pairs; V and W points are identical in the two structures, while C points are reflected relative to the xy plane. Distances involving the $C^\pm$ points are highlighted, using a minimum image convention along the periodic direction. (b) Euclidean adjacency matrices between the points in the A^+ and A^- structures. The two matrices differ only by the order of the $C - W$ distances. V and V' have the same set of edge distances, and so the W/W' and $C^\pm/C^{\pm'}$ pairs: thus, from the point of view of the WL test, there are only three types of nodes and the graphs are indistinguishable. (c) A pair of finite-dimensional degenerate structures can be obtained by wrapping the period around the z axis and embedding the periodic structure in Euclidean 3D space.
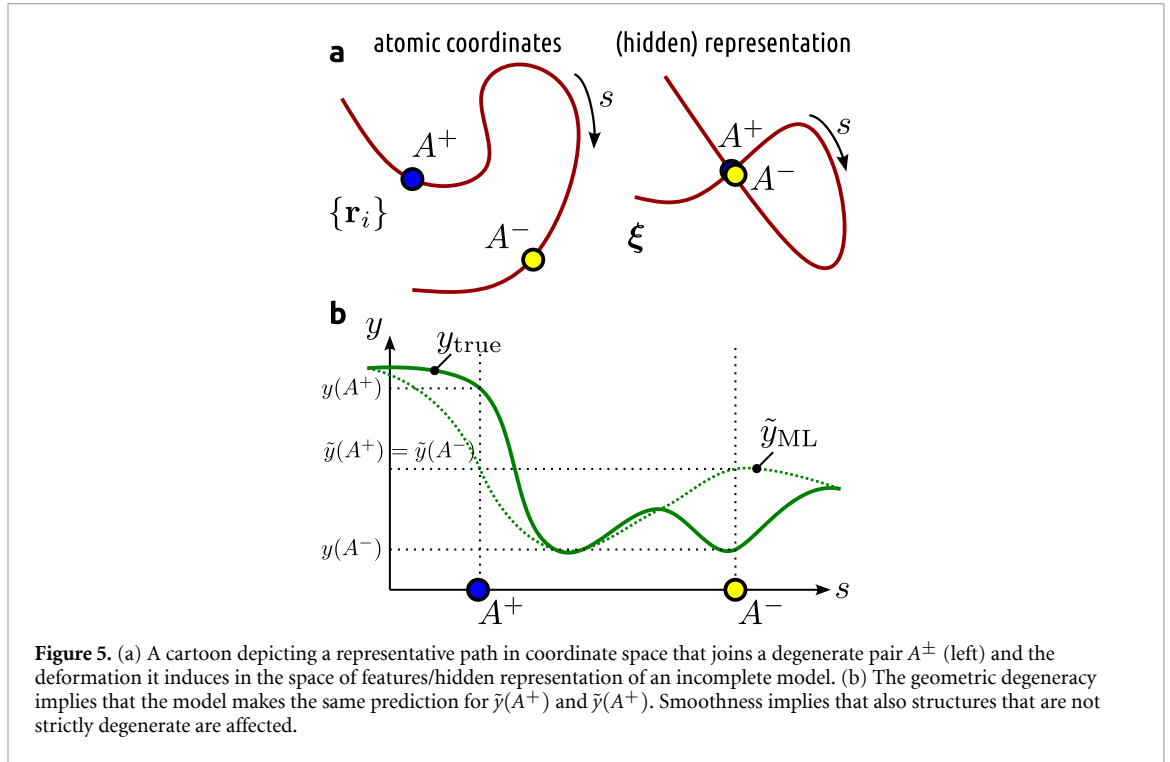
molecular graphs convolutions [36], graph networks [37], and many others, which from a graph theoretical perspective form a hierarchy of tests, all less powerful than the WL analysis [38]. As a special case, the $A^\pm$ structures in figure 4(a) are also degenerate with respect to the *full set* of neighbor-distances multisets, which means they are an example for which $\nu = 1$ atom-centered features cannot discriminate *globally* between the two structures. In this sense, this is a a much more difficult case than the degenerate tetrahedra in figure 1(a), which have the same set of distances but can be easily distinguished when considering the triplets of distances associated with the vertices, and of that in figure 1(b), which is globally undistinguishable by $\nu = 1$ representations, but can be resolved by a dGNN.

This construction produces a pair of continuous manifolds of dimension 7 (and co-dimension 8, discarding translations), and can be further generalized in several different ways. Arbitrarily many pairs of V and W atoms, possibly with different labels, can be added to the two structures without breaking the degeneracy. Each pair increases the dimension of the manifold by two and the codimension by four. Structures of any size and complexity can be generated with this construction, even though the increase in co-dimension suggests that they become less 'dense'. One should keep in mind that the presence of degenerate configurations affects the accuracy and numerical stability of ML models even for other structures [39, 40]. From the point of view of the input features (or more broadly, the hidden representation of a deep-learning framework), an overlap of two structures that should be distinct determines a distortion that brings close together structures that should be far apart (figure 5(a)). From the point of view of predicting properties, e.g. the potential energy of a molecule, the smoothness of the approximation is a key requirement to ensure stability and transferability of the model, which is fulfilled by all practical implementations. Thus, small deformations of a molecule should result in small changes of the target, and so a degenerate model will predict similar values for all structures that are close to a pair of degenerate configurations, even though they are not exactly degenerate (figure 5(b)).

As we discuss in appendix A, periodic boundary conditions can be added to the y and z axis without lifting the degeneracy of the pairs. Intriguingly, it is also possible to 'fold' an arbitrary number P of repeat units of the structure around the z axis, to obtain a pair of finite structures embedded in 3D Euclidean space (figure 4(c)). If the coordinates in the x-periodic structure are $\{(x_i, y_i, z_i)\}$, the position of the finite-size structure are

$$
\begin{aligned}
r_i &= Pp/2\pi + y_i \\
\theta_i &= 2\pi x_i/pP \\
z_i &= z_i
\end{aligned}
\tag{3}
$$

in a cylindrical coordinate system. This coordinate transformation works because it maps equal distances in the periodic structures to equal Euclidean distances in the finite configurations, and so it does not affect the nature of the graphs, and their signature when subject to the WL test.

Figure 5. (a) A cartoon depicting a representative path in coordinate space that joins a degenerate pair $A^{\pm}$ (left) and the deformation it induces in the space of features/hidden representation of an incomplete model. (b) The geometric degeneracy implies that the model makes the same prediction for $\tilde{y}(A^+)$ and $\tilde{y}(A^+)$. Smoothness implies that also structures that are not strictly degenerate are affected.

4. Beyond distance-based GNNs

All of these structures are easily distinguished by incorporating information on the angles, so that any scheme that contains at least 3-body descriptors, such as the smooth overlap of atomic positions (SOAP) features [13], or atom-centered symmetry functions [6], are not affected by this counterexample. As a consequence, higher-order GNNs [26], as well as many other deep learning approaches that incorporate angular information (such as DimeNET [41] GemNET [42], REANN [43], ALIGNN [44], etc) are also immune to these counterexamples, even though one cannot exclude that 3D counterexamples may be found also for some of these architectures, given that it is known that general graphs that defy higher-order versions of the WL test exist. Frameworks that can generate systematically arbitrary high-order correlation features [10] (such as MTP [18], ACE [19], NICE [20], etc) can be shown to be provide a complete description of interatomic correlations [9, 45] and are therefore, in the appropriate limit, symmetry-adapted universal approximators [46]. Similarly, equivariant neural networks (such as Tensor Field networks [27], Cormorant [47], etc) incorporate high-order correlation information by combining information on the interatomic distance *vectors*, constraining the functional form of the messages in a way that preserves equivariance, and achieve completeness in a very similar way as for the high-order correlation features [48].

The emergence of models which, in theory, have universal interpolating properties does not make the search for this kind of pathological configurations less important. For example, they can help verify whether the practical implementation of a ML scheme is consistent with its theoretical properties. Take for instance the universal scalar framework of [49]. A universal approximator for vectorial functions of the coordinates of n_{atoms} particles, that are equivariant in O(3) and invariant with respect to atom index permutations is proposed in the form (equation 11 in the reference)

$$\sum_{t=1}^{n_{\text{atoms}}} f(\mathbf{x}_t, \mathbf{x}_1, \dots \mathbf{x}_{t-1}, \mathbf{x}_{t+1}, \dots, \mathbf{x}_{n_{\text{atoms}}})\mathbf{x}_t, \tag{4}$$

where f is an arbitrary scalar function that is O(3) invariant, and invariant to the permutation of its arguments past the first. The crux of problem, however, is how to implement in practice such a universal symmetric scalar function. The example that is provided in section 7 of [49] uses a functional form that is suitable to learn a restricted class of targets, but is *not* a universal approximator. The proof for the full functional form is rather cumbersome, and we report it in appendix B, but the key problem is that the scalar functions are written in terms of the multi-set of the elements of the Gram matrix, e.g. $f(\{\mathbf{x}_i \cdot \mathbf{x}_j\}_{i,j=1}^{n_{\text{atoms}}})$, and therefore cannot distinguish between the degenerate structures we introduce here, and neither between the planar configurations in figure 1(c), that only differ by the order of the entries of their Gram matrix.

Furthermore, provably universal equivariant frameworks are such in the limit in which they generate high-order correlations—either explicitly or as a consequence of the stacking of equivariant layers [48]. It is an interesting, and open, question whether a given order suffices to guarantee complete resolving power. There is no reason to believe that the present family of degenerate structures is the *only* one that could be realized, based on a similar idea of generating configurations in which some distance pairs are swapped. It remains to be seen if a combination of the ideas we introduce here with those that underlie the counterexamples [11] for atom-centered descriptors of order $\nu = 2, 3$ could allow constructing configurations that are indistinguishable to convolutional schemes that also use angular (and possibly dihedral) information.

5. Significance for chemical ML

The construction in figure 4 might seem somewhat contrived, but it is not difficult to use it to generate point clouds that correspond to realistic targets for machine learning. Consider for example the case of predicting the stability of molecular structures. Figure 6 demonstrates three pairs of degenerate configurations corresponding to a water tetramer, that we obtained based on the expression for a finite structure (3) with two repeat units. Structure A_1^+, in particular, is a mildly distorted version of the ground state configuration of $(H_2O)_4$ [51], that we label as A_0. The energies of these structures (which we computed at the B3LYP [52] level, using PySCF [53]) are not absurd: the cohesive energies relative to four isolated H_2O molecules are $E_1^+ = -0.92$ eV, $E_1^- = 2.43$ eV, $E_2^+ = 0.39$ eV, $E_2^- = 2.07$ eV. These structures are part of a continuous manifold of degenerate pairs: as shown in figure 7, simply modifying the parameters in the construction (2) around these specific values generates hundreds of configurations with energies comparable with the dissociation energy of the tetramer. In particular, one can also obtain pairs for which both structures are below the cluster dissociation energy: we consider as an example the $A_3^\pm$ structures, which have $E_3^+ = -0.65$ eV, $E_3^- = -0.04$ eV.

Thus, in the best-case scenario a distance-based GNN can only predict the energies of $A_{1,2,3}^\pm$ with a root mean square error (RMSE) of ≈ 1.1 eV. For reference, the cohesive energy of the ground-state structure is $E_0 = -1.45$ eV, and the typical thermal energy at room temperature is 0.025 eV. Water oligomers have been extensively studied as a way to understand the properties of the hydrogen bond network in water [54, 55], that in turn influence the behavior of all chemical and biochemical phenomena that take place in solution. From a modeling point of view, an accurate estimation of the energetics of two and three-molecule clusters is important because it underlies the construction of very accurate interatomic models of the energetics of bulk water, such as the MB-pol potential [56, 57] that adds to a polarizable baseline that captures long-range interactions a series of short-range corrections based on fitting to high-end quantum calculations of dimers and trimers. These terms are then computed by summing over all pairs or triplets of water molecules in the system, so that even for a bulk configuration, one has to evaluate cluster energies.

In this context, the example we present is very relevant, in view of the growing interest in incorporating explicitly four-molecule terms [58]. With this application in mind, we chose the data set used in training the MB-pol potential [56, 57] to train models of the cluster cohesive energy, selecting 10 000 water dimers and 5000 water trimers by farthest-point sampling (FPS) [59], combined with 2000 water tetramers configurations (FPS selected from a high-temperature Hamiltonian replica exchange simulation with a confining potential, performed using i-PI [60] and the q-TIP4P/f forcefield [61]). Energies of all structures were then computed with the same B3LYP setup as for the degenerate structures.

We show a comparison between a simple kernel model based on SOAP features [13] and the SchNet framework [24], a GNN that has been very successfully applied to several chemical ML problems [62–66]. In SchNet [24], each atomic environment A_i is associated with a feature vector $\xi(A_i)$, that is initialized to values that depend only on the chemical nature of the *i*th atom. These are updated with on-site operations, and with interaction blocks that combine information on each of the neighbors according to

$$\xi_q^{k+1}(A_i) = \sum_{j \in A_i} \xi_q^k(A_j) W_q(r_{ij}), \tag{5}$$

where $W_q(r)$ indicates a set of distance-dependent continuous filters, and $\xi_q^k(A_j)$ indicates the features of the neighbors from the previous iteration of graph convolution. This functional form incorporates only the type of information associated with the distance-decorated WL test: the attributes of the neighbors, and scalar attributes of the interatomic separation, that are combined in a permutation-invariant manner. SOAP features, on the other hand, contain information on the angular relations between neighbors. Even though we evaluate them by first computing an expansion of the neighbor density on spherical harmonics, and then
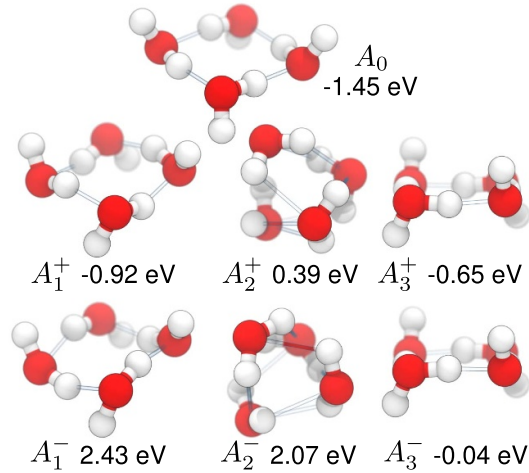
Figure 6. A few configurations of a water tetramer that will be used as benchmarks. A_0 is the ground state structure for the tetramer [50]. $A_1^\pm$, $A_2^\pm$ and $A_3^\pm$ are three pairs of structures that correspond to two periods of the 3D motif of figure 4(c). The geometries of the $A_{1,2}^+$ structures have been obtained by a constrained minimization of the energy, while $A_3^\pm$ have been obtained by minimizing simultaneously the energies of both degenerate structures.
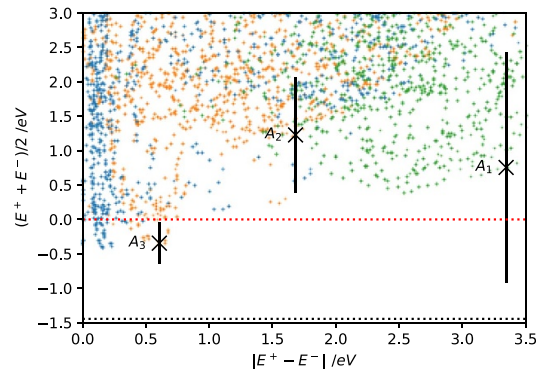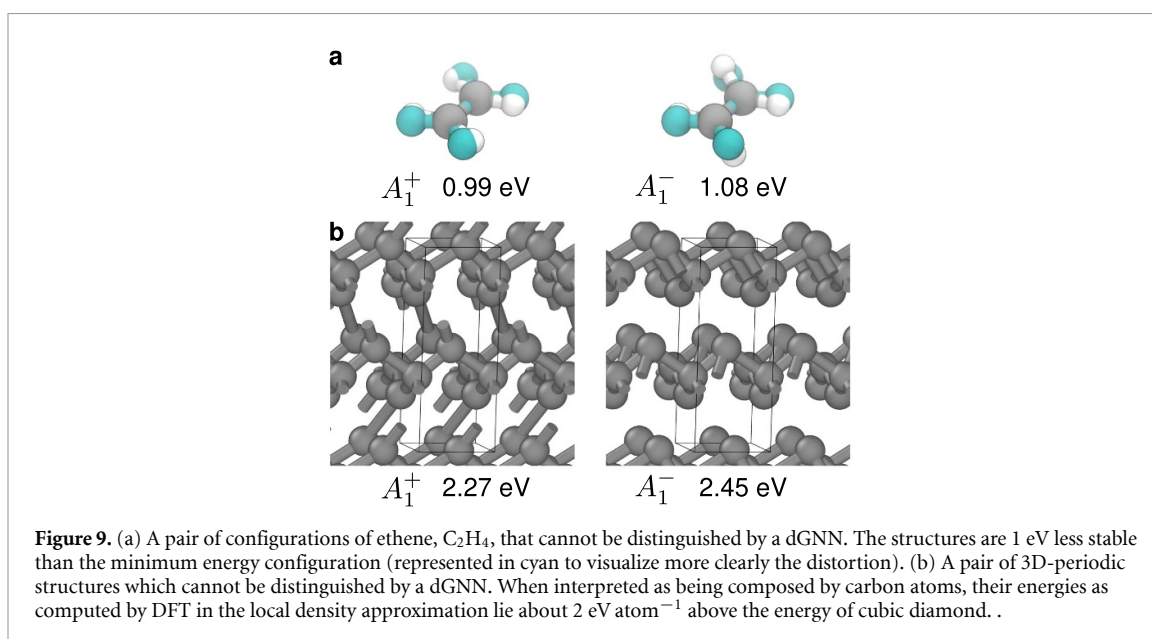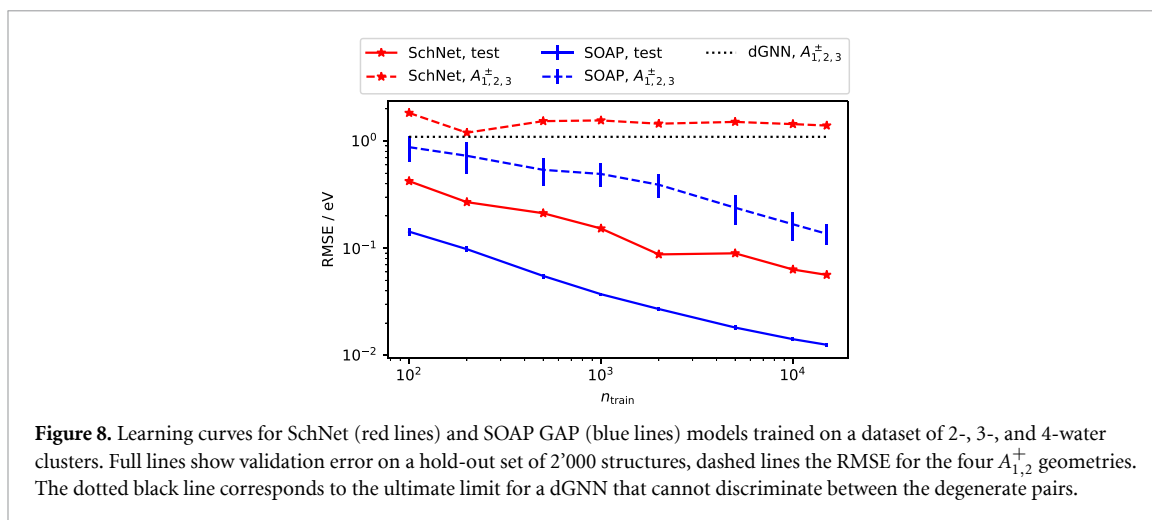

Figure 7. Scatter plot showing the energies (relative to the fully dissociated 4 H_2O geometry) of several hundreds of cluster geometries, obtained by finite changes of the parameters of the three structures shown in figure 6. For each pair of degenerate conformers, a point is shown indicating the mean and the spread of the two energies. Points corresponding to $A_{1,2,3}$ are shown as black crosses, and a bar indicates the energies of the two structures in the pair. The dissociation limit and the energy of the ground-state structure are shown as red and black dashed lines.

contracting the expansion coefficients to extract the rotationally-invariant components [13], SOAP features can also be expressed in a form that highlights their dependence on neighbor–neighbor angles

$$\xi_q(A_i) = \sum_{j,j' \in A_i} W_q(r_{ij}, r_{ij'}, \mathbf{r}_{ij} \cdot \mathbf{r}_{ij'}). \qquad (6)$$

We combine these features to build polynomial kernels $k(A_i, A_{i'}) = (\boldsymbol{\xi}(A_i) \cdot \boldsymbol{\xi}(A_{i'}))^4$, that are then used for kernel ridge regression. We emphasize that we chose SchNet as a widespread GNN model, but *any* distance-based GNN that has a discriminating power equivalent to (or lower than) a first-order WL test would exhibit the same problem. Similarly, we use an SOAP-based Gaussian Approximation Potential as a simple and well-understood scheme that incorporates explicitly angular information, but essentially any framework that does so would be capable of resolving the $A^\pm$ degeneracies.

We use standard hyperparameters for both SchNet and the SOAP model (example scripts, full model parameters and the training and test sets are provided in the Supplementary data). As shown in figure 8, the difference in behavior between the two schemes is qualitative. Both models can reduce monotonically the validation error on the *n*-mers dataset, but only the model that incorporates angular information can tell the $A^\pm$ pairs apart, and can bring the errors for $A_{1,2,3}^\pm$ below the theoretical limit that corresponds to predicting for both structures the mean of E^+ and E^-. The figure also shows that the two models exhibit very different performances, and that the SOAP-based kernel model yields a much larger error for $A_{1,2,3}^\pm$ than for the test set, which does contain structures with a similar energy range. It is difficult to conclusively determine the

Figure 8. Learning curves for SchNet (red lines) and SOAP GAP (blue lines) models trained on a dataset of 2-, 3-, and 4-water clusters. Full lines show validation error on a hold-out set of 2'000 structures, dashed lines the RMSE for the four $A_{1,2}^{+}$ geometries. The dotted black line corresponds to the ultimate limit for a dGNN that cannot discriminate between the degenerate pairs.

Figure 9. (a) A pair of configurations of ethene, C_2H_4, that cannot be distinguished by a dGNN. The structures are 1 eV less stable than the minimum energy configuration (represented in cyan to visualize more clearly the distortion). (b) A pair of 3D-periodic structures which cannot be distinguished by a dGNN. When interpreted as being composed by carbon atoms, their energies as computed by DFT in the local density approximation lie about 2 eV atom^{-1} above the energy of cubic diamond. .

reason for these observations, that depend at least in part on the detailed setup of the two models, which we have deliberately not optimized, because the qualitative observation of the failure of dGNNs is independent on parameters or implementation details. One may hypothesize that structures that are only distinguishable by angular information may be more challenging for a SOAP model—because of the higher complexity of a three-body potential compared to pair terms. As for the dGNNs, even in the absence of structures that are *exactly* degenerate, one can expect that the lack of resolving power is reflected in a slower convergence. Structures that lie in the vicinity of a singularity can be distinguished by the dGNN, but only by sacrificing the smoothness, and hence the data-efficiency, of the approximation (see also figure 5). This argument provides a plausible, although not conclusive, explanation for the improvement in performance that is observed when comparing dGNNs with frameworks that incorporate angular or directional information [67], and is consistent with similar considerations that can be made to explain the saturation of learning curves for SOAP-based models applied to a dataset containing structures close to two-neighbors degeneracies (figure 1(c)) [11].

We conclude by reiterating that this is just one of many possible examples that can be realized using our construction. Figure 9 shows a pair of ethene configurations that cannot be distinguished by a WL test. What is more, it also shows a pair of bulk carbon structures that are indistinguishable. The existence of 3D periodic structures that are indistinguishable even when using an infinite distance cutoff provides a clear demonstration that the degeneracies are not simply explained by the presence of a small number of interatomic distances. Rather, they are associated with the existence of periodicity and near-symmetries, so that the distances that decorate the fully-connected graph contain redundant information that is not sufficient to distinguish A^{+} from A^{-}. We emphasize that we provide these structures as examples of

configurations which defy dGNNs, and as a test to verify whether a given model can resolve the counterexamples we discuss here. There is at least one continuous manifold of degenerate geometries, and it is likely that a more thorough search, including the use of different chemical species, would lead to conformers that are even closer in energy to the thermodynamically-stable conformations and phases.

6. Conclusions

The idea of combining a characterization of point clouds in terms of graphs, where points are vertices and the connectivity is determined by point-to-point distances, with graph-convolution architectures has been extraordinarily successful in a broad range of applications of geometric machine learning, and in particular in the construction of deep models of chemical structures, that lend themselves naturally to such description. Despite their simplicity, and despite the fact that general graphs that are indistinguishable based on first-order GNN are known to exist, these frameworks have remarkable descriptive power for actual molecular structures, that correspond to a special class of fully-connected graphs with edges decorated by the distances between the atoms. For example, they are capable of discriminating between configurations that are degenerate to widely used atom-centered representations. The previously known cases in which they fail to distinguish two structures depended on the choice of a small, finite cutoff in the definition of the molecular graph. The construction we present here, that generates pairs of geometries that are indistinguishable when seen through the lens of a dGNN, regardless of the chosen cutoff, provides a counterexample that sets a limit to the accuracy that can be attained when regressing properties associated with the 3D point cloud. For the specific case of chemical machine learning, we show a concrete case, relevant to the chemistry of water clusters, demonstrating that the counterexample can have direct repercussions on practical applications, such as the training of many-body potentials for water and aqueous systems.

Atom-centered descriptors that rely on correlations with at least two neighbors, such as SOAP [13], Behler-Parrinello symmetry functions [68], etc – as well as high-order graph convolution schemes and equivariant neural networks—are immune to this family of counterexamples, although low-order correlations are affected by other types of degeneracies. Atom-centered descriptors of higher order, or equivariant network architectures, can in principle be made complete, often resulting in improved performance on chemical learning tasks. Determining rigorously the *minimal* amount of information that must be incorporated in geometric machine learning to guarantee that an architecture can resolve arbitrary point clouds is a fascinating topic that touches upon several open problems in signal processing [69] and invariant theory [14]. We hope that our data set will be used to test existing frameworks to determine empirically their descriptive power, and that our construction could be taken as a basis to break some of the angle-dependent schemes and to determine the simplest architecture that is not affected by this, or other, degeneracies.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgments

Appendix A. Extension of the counterexample to 2D and 3D periodic structures

One of the reasons why the pairs of configurations (A^+, A^-) we introduce in this work are indistinguishable to a WL test is the fact that each atomic environment in structure A^+ has a corresponding environment in A^- with the same set of neighbor distances. Introducing additional neighbors (as one does when considering periodicity along y and/or z) increases the size of the graph and the number of interatomic distances. This additional information could break the degeneracy of one or more environments.

In general, the equality of a pair of distances does not ensure that the distances corresponding to periodic replicas of the atoms involved will be the same: $\|(\Delta x_1, \Delta y_1, \Delta z_1)\|^2 = \|(\Delta x_2, \Delta y_2, \Delta z_2)\|^2$ does not guarantee that
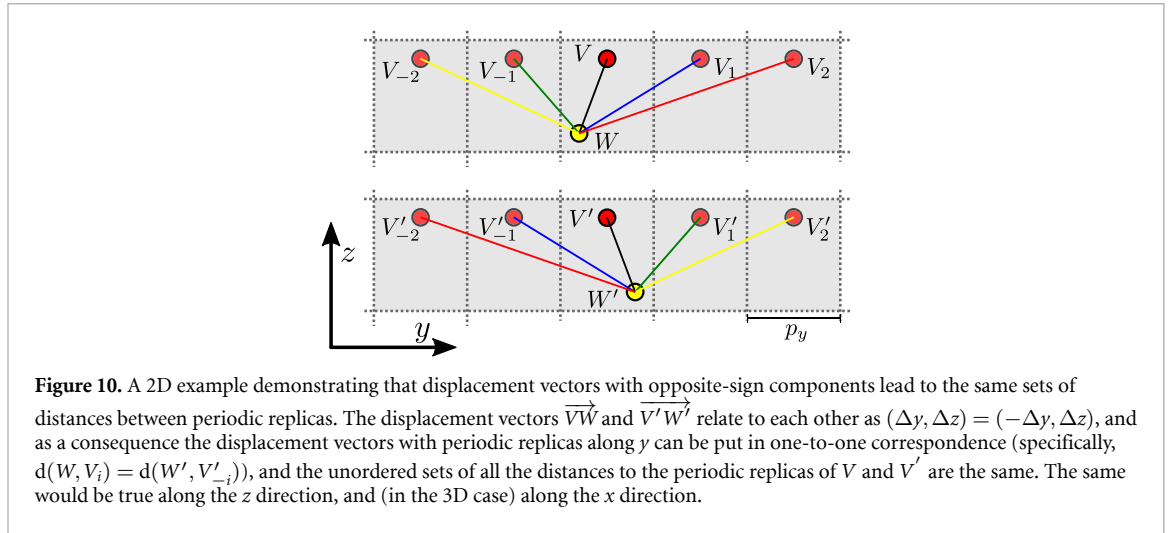
Figure 10. A 2D example demonstrating that displacement vectors with opposite-sign components lead to the same sets of distances between periodic replicas. The displacement vectors $\overrightarrow{VW}$ and $\overrightarrow{V'W'}$ relate to each other as $(\Delta y, \Delta z) = (-\Delta y, \Delta z)$, and as a consequence the displacement vectors with periodic replicas along y can be put in one-to-one correspondence (specifically, $\mathrm{d}(W, V_i) = \mathrm{d}(W', V'_{-i})$), and the unordered sets of all the distances to the periodic replicas of V and V' are the same. The same would be true along the z direction, and (in the 3D case) along the x direction.

$$\|(\Delta x_1 + n_x p_x, \Delta y_1 + n_y p_y, \Delta z_1 + n_z p_z)\|^2 \;=\; \|(\Delta x_2 + n_x p_x, \Delta y_2 + n_y p_y, \Delta z_2 + n_z p_z)\|^2 \tag{A1}$$

for all periodicities (p_x, p_y, p_z) and cell indices (n_x, n_y, n_z). A special case for which equation (A1) clearly holds is that in which not just the distances, but the full *distance vectors* are equal, $(\Delta x_1, \Delta y_1, \Delta z_1) = (\Delta x_2, \Delta y_2, \Delta z_2)$. This is the case for several distance pairs in the structures in figure 7 (e.g. the displacement vector $\overrightarrow{VC^+}$ equals $\overrightarrow{V'C^{-'}}$) but would not suffice, alone, to ensure that the fully-periodic structures are degenerate. We must consider an additional, more subtle case, in which the distance vectors have one or more components which are equal in magnitude, but have opposite signs, e.g. $\Delta z_1 = -\Delta z_2$. In this case, while it is not true that the distances corresponding to the same replica (n_x, n_y, n_z) will be equal, it will be possible to establish a 1-to-1 mapping between the replicas—in this case, $(n_x, n_y, n_z) \rightarrow (n_x, n_y, -n_z)$ (see figure 10), and so the unordered sets of distances will be equal.

One can then see that in the original 1D construction illustrated in figure 4(a), every time distances between pairs of points are the same (i.e. are marked by the same color in the figure), then the corresponding displacement vectors consist of the same Cartesian components with, possibly, opposite signs. According to the discussion above, this implies that if some distances within the original 1D counterexample are the same, then the unordered sets of distances associated with the replicas along the y and/or z axis are also the same. Thus, the WL test (which operates exclusively based on the unordered sets of distances) cannot discriminate between the 2D and 3D analogues of the 1D construction.

Appendix B. Counterexample for a 'universal approximator' implementation

Villar *et al* [49] proposes a general recipe to construct universal approximators on point clouds. As a specific example, the authors propose learning a function defined for a point cloud described by N 3D coordinates $\mathbf{r}_i$ and a scalar node property m_i (interpreted as a point mass). The goal is to learn a rotationally covariant and permutationally invariant tensorial function h:

$$h \;:\; (\mathbb{R} \times \mathbb{R}^3)^N \rightarrow \mathbb{R}^{3 \times 3}. \tag{B1}$$

The authors propose a functional form of the type

$$h((m_i, \mathbf{r}_i)_{i=1}^N) = \sum_{i=1}^N f_0 \left(\mathbf{r}_i^T \mathbf{r}_i, m_i, \{\mathbf{r}_k^T \mathbf{r}_l, m_k, m_l\}_{k,l \neq i} \right) \mathbf{r}_i \mathbf{r}_i^T$$

$$+ \sum_{i>j=1}^N f_1 \left(\mathbf{r}_i^T \mathbf{r}_j, m_i, m_j, \{\mathbf{r}_k^T \mathbf{r}_l, m_k, m_l\}_{k,l \neq i,j} \right) \mathbf{r}_i \mathbf{r}_j^T$$

$$+ f_2 \left(\{\mathbf{r}_i^T \mathbf{r}_j, m_i, m_j\}_{i,j=1}^N \right) \mathbf{1} \tag{B2}$$

where $\mathbf{1}$ indicates the identity matrix and $\{\ldots\}$ indicates unordered multisets. We are interested in determining whether this form provides a universal approximator for symmetric functions of the type (B1), and present a counterexample showing that this is not the case.

We use a simpler geometry than the general construction we present in the main text, but the counterexample is based on similar principles. This pair of structures is shown in figure 1(b), and is also a counterexample for two-body atom centered additive models. We chose this pair because the coordinates take integer values, and so it is simpler to show explicit values of the various quantities. The general case would also result in the impossibility of learning the targets. We also take all m_i equal to each other, so that the m_i can be dropped from the definition (B2). The coordinates of the first set of points are given by:

$$(\mathbf{r}_i^+)_{i=1}^5 = \begin{bmatrix} 1 & -1 & 2 & -2 & 0 \\ 1 & -1 & 0 & 0 & 1 \\ 0 & 0 & 2 & -2 & 1 \end{bmatrix}^{\mathrm{T}} \tag{B3}$$

while those of the second set of points are

$$(\mathbf{r}_i^-)_{i=1}^5 = \begin{bmatrix} 1 & -1 & 2 & -2 & 0 \\ 1 & -1 & 0 & 0 & 1 \\ 0 & 0 & 2 & -2 & -1 \end{bmatrix}^{\mathrm{T}}. \tag{B4}$$

We use a simple diagonal function as target,

$$h((\mathbf{r}_i)_{i=1}^5) = \sum_k \lambda_k^3 \, 1, \tag{B5}$$

which is constructed by summing over the eigenvalues λ_k of the Gram matrix of each structure, and is clearly rotationally covariant and invariant with respect to permutations of the points in the point cloud. The value of the function h for the two point clouds differ by

$$h((\mathbf{r}_i^+)_{i=1}^5) - h((\mathbf{r}_i^-)_{i=1}^5) = \begin{bmatrix} 192 & 0 & \\ 0 & 192 & 0 \\ 0 & 0 & 192 \end{bmatrix}. \tag{B6}$$

We decompose the difference between the predictions given by the form (B2) as:

$$h((\mathbf{r}_i^+)_{i=1}^5) - h((\mathbf{r}_i^-)_{i=1}^5) = \Delta_0 + \Delta_1 + \Delta_2, \tag{B7}$$

where each Δ_k corresponds to the difference between the values of the term that contains f_k. The entries of the Gram matrix of the two structures are the same (even though ordered differently, so that their eigenvalues differ), and so obviously Δ_2 has to be zero, given that f_2 is evaluated on identical sets. The expression for Δ_0 contains 10 terms, but most cancel out, leaving just

$$\Delta_0 = f_0(\zeta) \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 2 \\ 0 & 2 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 2\,f_0(\zeta) \\ 0 & 2f_0(\zeta) & 0 \end{bmatrix},$$
$$\zeta = (2, \{-8, -8, -2, -2, -2, -2, -2, -2, 2, 2, 2, 2, 2, 2, 8, 8\}). \tag{B8}$$

Similarly, the expression for Δ_1 contains 20 terms, but most cancel out, leaving an expression of the form

$$\Delta_1 = \begin{bmatrix} 0 & -4\,f_1(\kappa_1) + 4\,f_1(\kappa_2) & 2\,f_1(\kappa_3) - 2\,f_1(\kappa_4) \\ 0 & 0 & 2\,f_1(\kappa_3) - 2\,f_1(\kappa_4) \\ 0 & -4\,f_1(\kappa_1) + 4\,f_1(\kappa_2) & 0 \end{bmatrix}$$
$$\kappa_1 = (-2, \{-8, -8, -2, -2, -2, -2, -2, -2, -2, -1, -1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 8, 8\})$$
$$\kappa_2 = (2, \{-8, -8, -2, -2, -2, -2, -2, -2, -2, -1, -1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 8, 8\}) \tag{B9}$$
$$\kappa_3 = (1, \{-8, -8, -2, -2, -2, -2, -2, -2, -2, -1, -1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 8, 8\})$$
$$\kappa_4 = (-1, \{-8, -8, -2, -2, -2, -2, -2, -2, -2, -1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 8, 8\}).$$

Combining the different parts, we obtain that

$$h((\mathbf{r}_i^+)_{i=1}^5) - h((\mathbf{r}_i^-)_{i=1}^5) = \begin{bmatrix} 0 & 4(f_1(\kappa_2) - f_1(\kappa_1)) & 2(f_1(\kappa_3) - f_1(\kappa_4)) \\ 0 & 0 & 2(f_1(\kappa_3) - f_1(\kappa_4) + f_0(\zeta)) \\ 0 & 4(f_1(\kappa_2) - f_1(\kappa_1)) + 2f_0(\zeta) & 0 \end{bmatrix}, \qquad \text{(B10)}$$

which is incompatible with the diagonal form of the actual difference between the functions. This rather cumbersome derivation demonstrates that—even though the general form of scalar invariants discussed in [49] does provide a framework to build universal approximators—the specific form chosen as a practical example is not able to fit certain ground truth functional dependencies for certain types of 3D point clouds.

ORCID iDs

Sergey N Pozdnyakov ⬤ https://orcid.org/0000-0001-5980-5813
Michele Ceriotti ⬤ https://orcid.org/0000-0003-2571-2832

References

[1] Gumhold S *et al* 2001 *IMR* pp 293–305
[2] Guo Y, Wang H, Hu Q, Liu H, Liu L and Bennamoun M 2021 *IEEE Trans. Pattern Anal. Mach. Intell.* **43** 4338
[3] Wu W, Qi Z and Fuxin L 2019 *2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)* (Long Beach, CA: IEEE) pp 9613–22
[4] Bello S A, Yu S, Wang C, Adam J M and Li J 2020 *Remote Sens.* **12** 1729
[5] Li Y, Ma L, Zhong Z, Liu F, Chapman M A, Cao D, Li J and Trans I E E E 2021 *Neural Netw. Learn. Syst.* **32** 3412
[6] Behler J and Parrinello M 2007 *Phys. Rev. Lett.* **98** 146401
[7] Bartók A P, Payne M C, Kondor R and Csányi G 2010 *Phys. Rev. Lett.* **104** 136403
[8] Rupp M, Tkatchenko A, Müller K-R and von Lilienfeld O A 2012 *Phys. Rev. Lett.* **108** 058301
[9] Musil F, Grisafi A, Bartók A P, Ortner C, Csányi G and Ceriotti M 2021 *Chem. Rev.* **121** 9759
[10] Willatt M J, Musil F and Ceriotti M 2019 *J. Chem. Phys.* **150** 154110
[11] Pozdnyakov S N, Willatt M J, Bartók A P, Ortner C, Csányi G and Ceriotti M 2020 *Phys. Rev. Lett.* **125** 166001
[12] Pozdnyakov S N, Zhang L, Ortner C, Csányi G and Ceriotti M 2021 *Open Res. Eur.* **1** 126
[13] Bartók A P, Kondor R and Csányi G 2013 *Phys. Rev.* B **87** 184115
[14] Boutin M and Kemper G 2004 *Adv. Appl. Math.* **32** 709
[15] Sadeghi A, Ghasemi S A, Schaefer B, Mohr S, Lill M A and Goedecker S 2013 *J. Chem. Phys.* **139** 184118
[16] Widdowson D, Mosca M M, Pulido A, Cooper A I and Kurlin V 2022 *Match* **87** 529
[17] von Lilienfeld O A, Ramakrishnan R, Rupp M and Knoll A 2015 *Int. J. Quantum Chem.* **115** 1084
[18] Shapeev A V 2016 *Multiscale Model. Simul.* **14** 1153
[19] Drautz R 2019 *Phys. Rev.* B **99** 014104
[20] Nigam J, Pozdnyakov S and Ceriotti M 2020 *J. Chem. Phys.* **153** 121101
[21] Wang Y, Sun Y, Liu Z, Sarma S E, Bronstein M M and Solomon J M 2019 *ACM Trans. Graph.* **38** 1
[22] Gilmer J, Schoenholz S S, Riley P F, Vinyals O and Dahl G E 2017 *Int. Conf. Machine Learning* pp 1263–72
[23] Schütt K T, Sauceda H E, Kindermans P-J, Tkatchenko A and Müller K-R 2018 *J. Chem. Phys.* **148** 241722
[24] Schütt K, Kindermans P-J, Felix H E S, Chmiela S, Tkatchenko A and Müller K-R 2017 *Neural Information Processing Systems*
[25] Sato R 2020 arXiv:2003.04078
[26] Morris C, Ritzert M, Fey M, Hamilton W L, Lenssen J E, Rattan G and Grohe M 2021 arXiv:181002244 [Cs Stat]
[27] Thomas N, Smidt T, Kearnes S, Yang L, Li L, Kohlhoff K and Riley P 2018 arXiv:1802.08219
[28] Zhang Y, Xia J and Jiang B 2021 *Phys. Rev. Lett.* **127** 156002
[29] Garg V K, Jegelka S and Jaakkola T 2020 arXiv:2002.06157
[30] Weisfeiler B and Leman A 1968 *NTI Ser.* **2** 12
[31] Douglas B L 2011 arXiv:1101.5211
[32] Shervashidze N and Borgwardt K M 2009 *Proc. 22nd Int. Conf. Neural Information Processing Systems (NIPS'09)* (Red Hook, NY: Curran Associates Inc.) pp 1660–8
[33] Kipf T N and Welling M 2016 arXiv:1609.02907
[34] Shervashidze N, Schweitzer P, van Leeuwen E J, Mehlhorn K and Borgwardt K M 2011 *J. Mach. Learn. Res.* **12** 2539
[35] Duvenaud D K, Maclaurin D, Iparraguirre J, Bombarell R, Hirzel T, Aspuru-Guzik A and Adams R P 2015 *Advances in Neural Information Processing Systems* pp 2224–32
[36] Kearnes S, McCloskey K, Berndl M, Pande V and Riley P 2016 *J. Comput. Aided Mol. Des.* **30** 595
[37] Chen C, Ye W, Zuo Y, Zheng C and Ong S P 2019 *Chem. Mater.* **31** 3564
[38] Xu K, Hu W, Leskovec J and Jegelka S 2019 *Int. Conf. Learning Representations*
[39] Parsaeifard B and Goedecker S 2022 *J. Chem. Phys.* **156** 034302
[40] Pozdnyakov S, Willatt M J, Bartok-Partay A, Ortner C, Csanyi G and Ceriotti M 2022 *J. Chem. Phys.* **157** 177101
[41] Gasteiger J, Groß J and Günnemann S 2020 *Int. Conf. Learning Representations*
[42] Klicpera J, Becker F and Günnemann S 2021 arXiv:2106.08903
[43] Zhang Y, Xia J and Jiang B 2022 *J. Chem. Phys.* **156** 114801
[44] Choudhary K and DeCost B 2021 *npj Comput. Mater.* **7** 185
[45] Sanchez J, Ducastelle F and Gratias D 1984 *Physica* A **128** 334
[46] Dym N and Maron H 2020 arXiv:2010.02449
[47] Anderson B, Hy T S and Kondor R 2019 *Neural Information Processing Systems* p 10
[48] Nigam J, Pozdnyakov S, Fraux G and Ceriotti M 2022 *J. Chem. Phys.* **156** 204115

[49] Villar S, Hogg D W, Storey-Fisher K, Yao W and Blum-Smith B 2021 *Advances in Neural Information Processing Systems* vol 34, ed M Ranzato, A Beygelzimer, Y Dauphin, P Liang and J W Vaughan (Red Hook, NY: Curran Associates, Inc.) pp 28848–63
[50] Wales D J and Walsh T R 1997 *J. Chem. Phys.* **106** 7193
[51] Maheshwary S, Patel N, Sathyamurthy N, Kulkarni A D and Gadre S R 2001 *J. Phys. Chem.* A **105** 10525
[52] Becke A D 1993 *J. Chem. Phys.* **98** 5648
[53] Sun Q *et al* 2020 *J. Chem. Phys.* **153** 024109
[54] Liu K, Cruzan J D and Saykally R J 1996 *Science* **271** 929
[55] Zwier T S 2004 *Science* **304** 1119
[56] Medders G R, Götz A W, Morales M A, Bajaj P and Paesani F 2015 *J. Chem. Phys.* **143** 104102
[57] Nguyen T T, Székely E, Imbalzano G, Behler J, Csányi G, Ceriotti M, Götz A W and Paesani F 2018 *J. Chem. Phys.* **148** 241725
[58] Heindel J P and Xantheas S S 2021 *J. Chem. Theory Comput.* **17** 7341
[59] Imbalzano G, Anelli A, Giofré D, Klees S, Behler J and Ceriotti M 2018 *J. Chem. Phys.* **148** 241730
[60] Kapil V *et al* 2019 *Comput. Phys. Commun.* **236** 214
[61] Habershon S, Markland T E and Manolopoulos D E 2009 *J. Chem. Phys.* **131** 24501
[62] Schütt K T, Gastegger M, Tkatchenko A, Müller K-R and Maurer R J 2019 *Nat. Commun.* **10** 5024
[63] Brorsen K R 2019 *J. Chem. Phys.* **150** 204104
[64] Dandu N, Ward L, Assary R S, Redfern P C, Narayanan B, Foster I T and Curtiss L A 2020 *J. Phys. Chem.* A **124** 5804
[65] Westermayr J, Gastegger M and Marquetand P 2020 *J. Phys. Chem. Lett.* **11** 3828
[66] Westermayr J and Maurer R J 2021 *Chem. Sci.* **12** 10755
[67] Schütt K, Unke O and Gastegger M 2021 *Proc. 38th Int. Conf. on Machine Learning (Proc. Machine Learning Research)* vol 139, ed M Meila and T Zhang (PMLR) pp 9377–88
[68] Behler J 2011 *J. Chem. Phys.* **134** 074106
[69] Kakarala R 2012 *J. Math. Imaging. Vis.* **44** 341