Article

Leveraging Persistent Homology Features for Accurate Defect Formation Energy Predictions via Graph Neural Networks

Zhenyao Fang* and Qimin Yan*

Cite This: *Chem. Mater.* 2025, 37, 1531−1540

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: In machine-learning-assisted high-throughput defect studies, a defect-aware latent representation of the supercell structure is crucial for the accurate prediction of defect properties. The performance of current graph neural network (GNN) models is limited due to the fact that defect properties depend strongly on the local atomic configurations near the defect sites and due to the oversmoothing problem of GNN. Herein, we demonstrate that persistent homology features, which encode the topological information on the local chemical environment around each atomic site, can characterize the structural information on defects. Using the dataset containing a wide spectrum of O-based perovskites with all available vacancies as an example, we show that incorporating the persistent homology features, along with proper choices of graph pooling operations, significantly increases the prediction accuracy, with the MAE reduced by 55%. Those features can be easily integrated into the state-of-the-art GNN models, including the graph Transformer network and the equivariant neural network, and universally improve their performance. Besides, our model also overcomes the convergence issue with respect to the supercell size that was present in previous GNN models. Furthermore, using the datasets of defective $BaTiO_3$ with multiple substitutions and multiple vacancies as examples, our GNN model can also predict the defect−defect interactions accurately. These results suggest that persistent homology features can effectively improve the performance of machine learning models and assist the accelerated discovery of functional defects for technological applications.
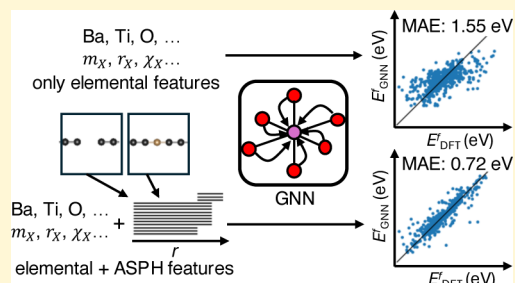
INTRODUCTION

Understanding the effect of point defects in materials has been a central area of study in materials science, since defects can significantly influence the physical and chemical properties of materials (such as reducing the bulk modulus and inducing structural phase transitions[1−,3] and introducing in-gap defect states which can further affect their performance in optoelectronic devices and catalysts.[4−6] One of the key quantities in defect chemistry is the defect formation energy,[7,8] which can affect various defect properties, such as the favorable type of defects and the location of defects in a given host material, the defect concentration at certain temperature, and defect migration pathways near surfaces or grain boundaries.[9,10] Therefore, fast and accurate predictions of defect formation energies are crucial to understanding the defect-related phenomenon in materials, enabling high-throughput screening of host material with desired defect properties.

To evaluate the defect formation energies, first-principles calculations based on the density functional theory (DFT) are often employed.[8,11,12] However, this method often involves large supercell calculations that are computationally expensive. Therefore, various machine learning methods, along with physics-inspired descriptors (such as the formation enthalpy, the energy above hull, the band gap, and the crystal reduction potential), were proposed to accelerate the predictions of defect formation energies.[13−17] Although these machine learning

models have achieved great accuracy in predicting defect formation energies, they lack the generalizability to accommodate a wide spectrum of materials because they rely on human-selected features, which are designed specifically for certain types of materials. Besides, due to the model architecture, most reported (shallow) machine learning models cannot incorporate the local structural information and thus cannot evaluate the defect formation energies for different symmetry-inequivalent atomic sites with structural relaxations, restricting their applications in predicting defect formation energies in complex materials that can potentially host multiple defects.

Recent years, graph neural networks (GNNs) have shown remarkable success in predicting various material properties, including the formation enthalpy, band gap, and bulk modulus.[18−20] In a GNN model, the input crystal structure is first converted into a graph, where the nodes and edges represent the atoms and bonds in the unit cell.[21] The graph then passes through several convolution layers, where the node information

exchanges between adjacent nodes and aggregate together, allowing the GNN model to learn the local chemical environment of each atom and to extract latent representations of the whole crystal structure, which will be used to make predictions on material properties. Until recently, few GNN models have been applied to predict defect formation energies.[22]−[24] As first attempts to use GNNs to predict defect related properties, these models incorporated various levels of chemical information, including the bond angle which may help detect the disrupted chemical bonding network near the vacancies,[24] but their prediction accuracy is currently limited for a diverse set of materials. The reasons for the poor performance of GNNs on predicting defect formation energies are 2-fold. First, defect formation energies highly depend on the atomic configurations near the defective sites and are independent of the atoms far from the defective sites, and are thus local properties of the crystal, contrary to the formation enthalpy or the band gap that are global quantities of the whole crystal. Second, GNNs suffer from the oversmoothing problem.[25] When passing through consecutive convolution layers, features between adjacent nodes will exchange and aggregate, and finally converge to constant values in a deep GNN model. This could lead to incorrect predictions especially when the prediction target only depends on the local atomic configuration. Therefore, in order for GNNs to accurately predict the defect formation energies, the structural information about the defects, such as the distance to the defect site or the number of nearby defects, must be encoded explicitly in the features of each node. This approach amplifies the differences in node features between the nodes near and far from the defect site, which can potentially reduce the oversmoothing effect of GNN and direct the focus of the neural network toward the nodes near the defects.

Based on these motivations, we propose to incorporate the persistent homology features into GNN models to predict the defect formation energy accurately. Persistent homology features are generated from the topological invariants of the homology group associated with the surrounding atomic configurations near each atom.[26,27] These features have already been applied in topological data analysis and in predicting (global) properties of molecules and materials, such as the formation enthalpy.[28,29] In this work, we demonstrate that persistent homology features encode the local structural information related to defects and can significantly increase the prediction accuracy of defect formation energies in complex materials. We show from several toy models that persistent homology features contain such structural information as the type and the size of defects, the number of defects around each atom (within a cutoff radius), and the distance to the defects from each atom. To verify the efficacy of persistent homology features, we train various GNN models on the dataset containing around 7700 O-based perovskite defective structures with monovacancies to predict their neutral defect formation energies. We find that introducing the persistent homology features significantly increases the performance of GNN models, with the mean absolute error (MAE) reduced by 55%. Besides, our model with the persistent homology features and the global max pooling layers can overcome the convergence issue with respect to the supercell size reported previously.[24] Furthermore, we also apply the GNN model to datasets of a prototypical cubic perovskite $BaTiO_3$[30] with multiple substitutions and multiple vacancies. We show that the persistent homology features also help capturing the defect−defect interaction energies in those data sets. These results suggest the crucial role of persistent

homology features in extracting local structural information near defects and in predicting physical and chemical properties related to defects in complex materials using machine learning or deep learning methods.

■ RESULTS

Persistent Homology Features and Defects. A simplicial complex X is formed by gluing "standard" geometric objects (simplices) of various dimensions (such as points, line segments, triangles, tetrahedrons) together. One of its homological properties is the d-dimensional Betti number $\beta_d(X)$, defined as the rank of the homology group $H_d(X)$. Any nontrivial element in the homology group $H_d(X)$ corresponds to some d-dimensional cycle (cyclic simplices) that is not simultaneously the boundary of any $(d + 1)$-dimensional chain, namely a d-dimensional "hole"; for example, a circle is a one-dimensional cycle and encloses a hole, whereas the boundary of a disk, though also being a one-dimensional cycle, does not enclose a hole since it is the boundary of the two-dimensional disk. Therefore, the d-Betti numbers characterize the number of d-dimensional holes in the simplicial complex and reflect the connectivity of the simplicial complex (see Supporting Information I for definitions).[31] In the context of topological data analysis, any set of data points can be converted into a simplicial complex by connecting the points within some cutoff radius together. The cutoff radius is allowed to vary (to an upper bound) so that the underlying pattern behind the given set of data points can be revealed through the birth, the death, and the persistence of Betti numbers (see Supporting Information I for more discussions).,[26,32,33]

In chemical systems such as molecules or materials, the atomic configuration near each atom can be naturally regarded as a point cloud, and the chemical bonding network as a simplicial complex, allowing for the calculations of persistent homology features. Previous work introduced atom-specific persistent homology (ASPH) features as a topological representation of the crystal structure.[29] For each base atom and each chemical species in the unit cell, we generate a point cloud which centers at the base atom and includes all other atoms of the given chemical species within the given upper bound cutoff radius. The persistent homology features are then calculated for each point cloud; since crystalline materials are three-dimensional, only 0-, 1-, and 2-Betti numbers are relevant. The birth, death, and persistence of these Betti number features are further characterized by five statistical quantities, including the minimum, maximum, mean, standard deviation, and the (weighted) sum, resulting in 35 statistical representations of the persistent homology features for each atom in the unit cell (for 0-Betti numbers, only its death will be considered). An example is shown in Figure 1A, where the crossed atom represents the vacancy in the two-dimensional lattice. The point cloud centered at the red atom includes the vacancy site, while that centered at the black atom excludes the vacancy site. The 0- and 1-Betti numbers are calculated for these two point clouds and their statistical quantities will be collected as the persistent homology features for these two atoms, respectively.

Since persistent homology features contain information about holes within a point cloud, we conjecture that these features can unveil the structural changes induced by defects in materials. In this work we examine two of the most common defect types, namely vacancies and substitutions. Vacancies can be regarded as holes in a material and thus be detected by persistent homology features, while substitutions, as we will demonstrate
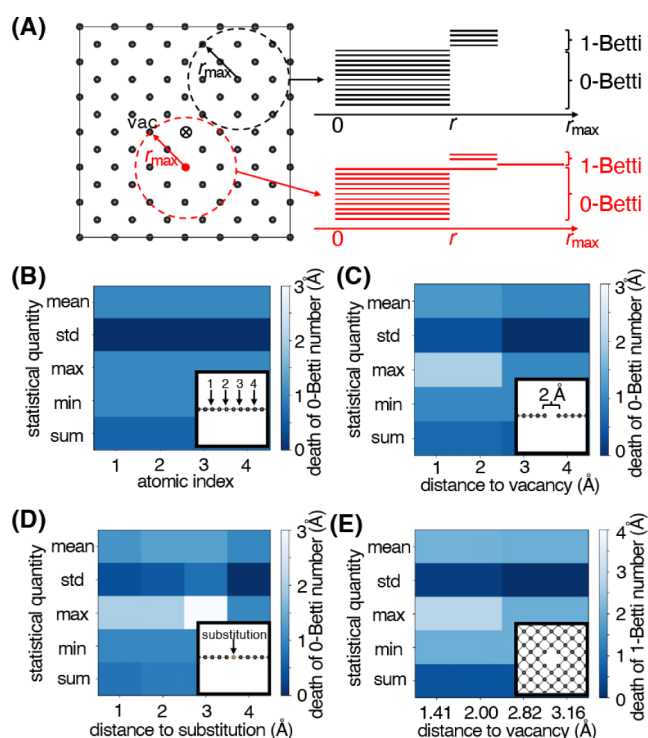
Figure 1. (A) A schematic plot of the calculation procedure of atoms-specific persistent homology features. (B) The calculated statistical quantity of the death of 0-Betti number features for the one-dimensional pristine toy model. (C, D) Same as (B), but for the toy model with one vacancy (C) and with a substitution (D) respectively. (E) The calculated statistical quantity of the death of the 1-Betti number features for the two-dimensional toy model with one vacancy.

numbers reflect the connected components in the simplicial complex. Therefore, all atoms on this one-dimensional chain have the same features, with both the minimum and the maximum statistical quantity to be 1 Å, which corresponds to the interatomic distance. We further consider the case with one vacancy in the chain, as shown in Figure 1C. For atoms that are at least 3 Å far away from the vacancy, the point cloud around them does not contain the vacancy site. Thus, the maximum death of their 0-Betti number features is 1 Å, above which all atoms in the point cloud are connected with each other and form one trivial connected component. On the other hand, for atoms less than 3 Å away from the vacancy site, the maximum death of their 0-Betti number features increases to 2 Å which is required for the two connected components to the left and right of the vacancy to connect with each other. Therefore, the 0-Betti number features can help distinguish the atoms close to or far from the defect site, as determined by the upper bound cutoff radius.

Apart from vacancies, we also argue that ASPH features are sensitive to substitutions. By definition, for each atom in the unit cell, the point cloud which will be used to calculate persistent homology features is constructed for each chemical species. Therefore, when an atom is substituted by an exotic species, the point cloud for this species will contain only two points (the base atom and the substituted atom), and the critical distance at which these two points are connected is exactly the distance to the substitution site. This is confirmed by the above toy model with one substitution, as shown in Figure 1D. The calculated features clearly reflect the distance to the substitution from each atom.

Additionally, in the two-dimensional case, not only the 0-Betti numbers, but also the 1-Betti numbers are relevant. Here we calculate the 1-Betti number properties for a two-dimensional lattice with one vacancy, shown in Figure 1E. In this case, the death of the 1-Betti number indicates the critical cutoff radius at which the hole formed by the vacancy is filled, with the critical cutoff radius to be the diagonal length of the square. More discussions on other toy models, including two distant vacancies and multiple element cases, can be found in Supporting Information I.

Although we only considered the vacancy and substitution defect in the following numerical tests on perovskite dataset (see next section), we expect that ASPH features can also reflect information on other types of defect such as interstitial defect.
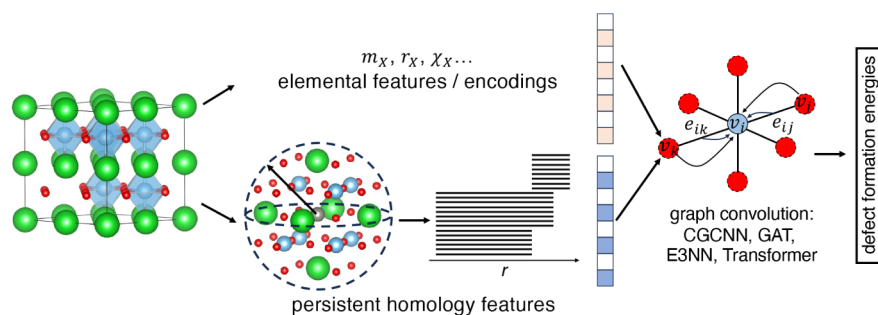
below, are also encoded in these ASPH features because the point cloud is constructed for each chemical species around the base atom.

To verify our hypothesis, we construct a toy model consisting of a one-dimensional atomic chain with equal interatomic spacing (1 Å). First, we consider the case without any defect, as shown in Figure 1B. Since this is a one-dimensional chain, only the 0-Betti number features are relevant. Therefore, we calculate the 0-Betti number features for all atoms with the upper bound cutoff radius to be 3 Å and generate the statistical quantities of them. As discussed in Supporting Information I, 0-Betti

Figure 2. A schematic plot of the GNN model based on the attention mechanism to predict the defect formation energies. The input defective structure will be converted into a crystal graph, where node features contain both the elemental features (such as the mass m_X, the radius r_X, and the electronegativity χ_X of element X) and the ASPH features, calculated by constructing the point cloud around each atom and tracking the birth/death/persistence of the homology features. The graph passes through several graph convolution layers, where features on adjacent nodes exchange and aggregate. The convoluted node features are gathered together into the global latent representation of the graph which will be used to predict the defect formation energy.

The additional atom and chemical bonding can be effectively captured when constructing the crystal graph, and if the interstitial atom belongs to an exotic chemical species, its unique contribution to the local structure will be explicitly reflected in the ASPH features, similar to the case of substitution defects. As shown in Supporting Information I, we constructed a two-dimensional toy model with one interstitial defect, and the statistical quantities are clearly distinct between atoms close to and far from the interstitial defect.

Finally, we anticipate that in three-dimensional complex materials with multiple elements and varying bond lengths and bond angles, 0-, 1-, and 2-Betti numbers are all nontrivial, and their statistical quantities can jointly capture the structural information near the defects, including the number of defects near each atom, the distance to those defects, and the size of the vacancies.

Vacancy Formation Energy in O-Based Perovskites. To numerically verify the capability of persistent homology features in capturing local structural information on defects and improving the prediction accuracy of defect-related properties in complex materials, we construct GNN models to predict the neutral defect formation energies for various datasets according to eq 1. These models can be easily adapted to predicting other defect properties such as charge transition levels and charged defect formation energies. For these properties, the defect charge state can be added to the global feature of the crystal and used to predict charged defect properties through the multilayer perceptron in the GNN model.[24]

Our GNN model features a typical model architecture for graph property prediction, as shown in Figure 2. First, the defective supercell structure is converted into a graph, where the nodes and the edges of the graph represent the atoms and the chemical bonds in the crystal. The node features carry the atomic information on each atom, including the elemental features and the ASPH features. In this work, we consider three types of elemental features: atomic number (only), elemental properties (including the atomic number, atomic mass, atomic radius, electron affinity, row and group numbers), and the one-hot encoding of the elemental type.[21] The complexity and representation capability progressively increase for these three different options; the atomic number feature has been used to predict configurational disorder related properties, which usually involve only few elements,[34] while the one-hot encoding feature was used to predict the formation enthalpies and the band gaps of a wide range of materials in Materials Project,[35] thus used in our work (benchmark results on elemental features can be found in Supporting Information II). On the other hand, the edge features represent the bonding properties, which in our case are simply the bond length expanded in a Gaussian function basis.[21] Note that in other types of higher-order GNN models, more bonding information on the crystal, such as the bond direction and the bond angle as well as the multiplet interactions, can also be included in the graph.[36,37]

The generated graph then passes through several graph convolution layers, where each node receives information from adjacent nodes and aggregate them together. Depending on the methods to aggregate and update node features, various convolution layers were proposed previously. In this work, we consider two convolution layers with the attention mechanism (the graph attention neural network (GAT)[38,39] and the Transformer network[40] and one without the attention mechanism (the crystal graph neural network (CGCNN).[21] In attention-based neural networks, the importance of an adjacent

node is expressed as the attention coefficient which is calculated from both the node features of the two nodes and edge features and reflects the similarity between the two node features (elemental properties and the persistent homology features) in the latent space. The node features are then updated by a linear combination of all adjacent node features weighted by the attention coefficients. The attention mechanism allows the GNN model to capture the chemical and structural distinctions between adjacent atoms and in general leads to better performance on various tasks.[41] In addition to those networks, we also consider the equivariant neural network (E3NN);[42] this network employs the E(3)-equivariant convolution operation which utilizes the spatial symmetries in crystals to make accurate predictions on material properties. In the following, the convolution layer type is considered a hyperparameter and optimized through Bayesian optimization (see Methods section), unless otherwise specified. Finally, after the graph convolution layers, all node features in the graph are gathered together through graph pooling layers to produce the latent representation of the whole crystal, which will be further passed to a multilayer perceptron network to obtain the defect formation energy. We will demonstrate below that the choices of node features, graph convolution layers, and pooling layers significantly affect the performance of GNN models; for the tasks related to predicting defect properties, the optimal choices of the model architecture are different from typical GNN models used to predict the formation enthalpies or band gaps of bulk materials.

Now we demonstrate the efficacy of persistent homology features using the dataset containing O-based perovskite defective structures. Here we consider all available mono-vacancies in these structures. To construct our dataset, we choose all perovskite structures with the chemical formula ABO_3 from Materials Project,[35] excluding those with cations after La in the periodic table. Most elements after La (such as the f-block elements, 6p and 6d elements) require further detailed investigations on the optimal $U(J)$ values that highly depend on individual perovskite structures within the DFT + U framework. Besides, excluding structures with those elements in our dataset does not affect the generalizability of our GNN models and the overall conclusion of our work. Next we use the pymatgen-analysis-defects package[43,44] to generate the super-cells to calculate the defect formation energies for each perovskite structure. The supercells are constructed as nearly cubic cells with the minimum (maximum) lattice constants along each direction to be 10 Å (25 Å), so that the interactions between defects and their periodic images along each Cartesian direction are minimal. Structures whose supercells cannot be constructed into the nearly cubic cell within the size range are discarded. Using the generated supercells, we enumerate all available symmetry-inequivalent atomic sites and generate the corresponding vacancy structures (including both cation and anion vacancies). To obtain the defect formation energies, we performed single-shot DFT calculations (without atomic relaxations) on the defective structures, according to eq 1. While we did not explicitly perform geometry optimizations when constructing the dataset, GNN models have the natural ability to capture geometric changes during relaxation processes through the edge properties, and since ASPH features can be easily incorporated into various kinds of GNN models that can capture the structural changes effectively, the lack of atomic relaxations does not affect our conclusion qualitatively.

Using the above procedure, our dataset contains around 7700 defective structures derived from around 1100 host (pristine) structures. As shown in Figure 3A, this dataset contains host
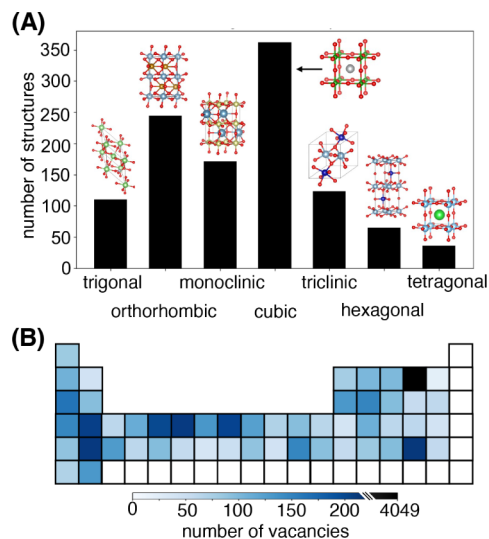

Figure 3. (A) The number of host structures of each crystal family in our constructed O-based perovskite dataset. The insets are example materials from each crystal family. (B) The number of vacancies of each elemental type in our dataset. Vacancies of all elemental types before La, except for noble gas elements, are present in our dataset.

structures from all crystal families, thus representing a comprehensive array of crystal structures. Besides, our dataset contains vacancies of all elemental types (before La), and the numbers of each type of vacancy are shown in Figure 3B; some of the most common elemental types are also summarized in Table 1. The various atomic and chemical bonding configurations make it challenging for GNN models to learn the structural information and predict the defect formation energies.

Table 1. Number of Vacancies of the Most Common Elemental Types in the Constructed O-Based Perovskite Dataset

vacancy type	O	Sr	Te	Cr	Ca	Fe	V
number	4049	216	203	178	173	171	161

First we present benchmark calculations on our constructed dataset using two state-of-the-art GNN models, namely the CGCNN model[21] and the higher-order model atomistic line GNN.[45] The latter model also incorporates the bonding angle using the line graph methods, allowing the model to capture the disrupted chemical bonding network near the defect sites and to make more accurate predictions about defect-related properties.[24] Both models used only the elemental features as the node features and global mean pooling layers. After training and testing on our dataset, the MAEs of these two GNN models are 1.86 and 1.83 eV, respectively.

The poor performance of these models demonstrates the challenge of our constructed dataset, but more importantly, the uniqueness of defect property prediction tasks. First, defect property calculations often require large supercell structures. Since GNN models involve graph pooling operations on all nodes and gather all node features as the global representation of the graph, the distinct node features of atoms near the defect site,

generated by a disrupted chemical bonding network (as is the case in atomistic line GNN), are averaged out by the node features of the numerous atoms located far from the defects in the supercell. Therefore, the performance of GNN models generally worsens using larger supercells for defect calculations, and this trend was reported previously.[24] Second, GNN models suffer from the oversmoothing problems.[25] Since the node features are exchanged between adjacent nodes for each convolution layer, the node features will exponentially converge to constant values in a deep GNN model. In our scenario, this problem will also average out the node features and make the distinct features carried by atoms near the defect site invisible to the model. Due to the above two issues, current GNN models cannot well capture the presence of defects and predict their properties on a dataset with limited number of data points, and additional defect-aware features must be added to the GNN models to better describe the structural information of defects.

Although several chemical descriptors were proposed to capture the local chemical environment in materials, they are not readily suitable to represent the local structural changes due to defects.[46−48] For example, the Coulomb matrix descriptors capture the local structure through the electrostatic interactions between nuclei, but they cannot distinguish the presence of nearby vacancies inherently since it relies on pairwise nuclear interactions. Another widely used descriptor is the smooth overlap of atomic positions. However, this method suffers from the summation of all atoms within the cutoff radius which can dilute the defect contribution to the whole crystal graph.

To overcome these issues, we choose to explicitly include the local chemical environment information associated with the defect site, represented as the ASPH features, into each node feature of the GNN model. As demonstrated above, the ASPH features are specifically sensitive to the presence of defects, and not dependent on any hyperparameter other than the cutoff radius. Therefore, these additional features can amplify the differences in node features between atoms located close and far from the defect sites for a diverse dataset. Besides, since ASPH features are derived from the statistical quantities of the persistent homology features, these features are not averaged out by other atoms that do not contribute to the defect properties or by the oversmoothing effect (along with proper choices of graph pooling layers), therefore enhancing the prediction accuracy of the GNN model.

For each defective structure, we generate the ASPH features (35 statistical quantities of the birth/death/persistence of Betti numbers) for each atom in the supercell.[29] When generating the features, the upper bound cutoff radius can affect the final performance of the GNN model. With larger upper bound cutoff radius, structural information from farther defects can be included, but computations will also become more expensive. In Supporting Information II, we present convergence test results on the upper bound cutoff radius, and decide to use 10 Å as the upper bound cutoff radius to generate the persistent homology features for the whole dataset. Besides, the generated homology features could be correlated with each other. To extract a concise feature vector and remove the redundant features, we preprocess the calculated ASPH features by the principal component analysis method, which reduces the dimensionality of features through linear transformations and concentrates the information in features into the principal components.[49] As shown in Supporting Information II, the MAE reduces as more principal components are included into the node features and reaches convergence at around 6 principal components; the sum of the

explained variance percentage of the first 6 principal components is 99.5%, suggesting that those principal components already contain most of the information from the original 35 persistent homology features in our dataset.

In Figure 4, we show the performance of GNN models (with global max pooling layers, see discussions below), including
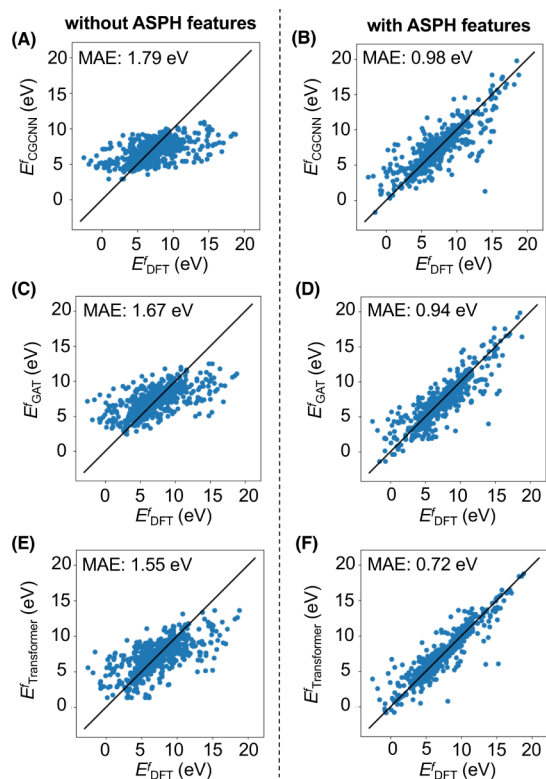

Figure 4. Comparison of the performance of GNN models (using global max pooling layers) on the O-based perovskite dataset without and with ASPH features. (A, B) CGCNN networks. (C, D) GAT networks. (E, F) Transformer networks.

CGCNN, GAT, Transformer networks, on our constructed O-based perovskite dataset. By including the ASPH features, the MAE decreases for all GNN networks, suggesting that ASPH features can consistently improve the prediction accuracy on defect-related tasks. Since those features can be easily incorporated into the node features, we expect that the improvement will also be observed for other GNN models. Furthermore, among all networks, the Transformer network performs the best, with the MAE reduced by 55% from 1.55 to 0.72 eV. This agrees with our previous findings in the configurationally disordered materials.[34] In Transformer networks, the features of the central node and adjacent nodes are treated as queries and keys, respectively. The overlap of the queries and keys reflects the similarity between the two nodes, determined by not only the elemental properties but also the chemical environment incorporated by ASPH features. Since the ASPH features entail the defect-related structure information such as the number of nearby defects and the distance to those defects, the attention mechanism allows the network to effectively extract the chemical information of defects and make accurate predictions on the defect properties. In addition to the above networks, we also trained the equivariant neural network (E3NN) on the dataset, and similar to other networks, introducing the ASPH features reduces the MAE from 1.94 to

1.27 eV (see Supporting Information II). On our dataset, E3NN network performs worse than the Transformer network, possibly due to the fact that both the node features and the prediction target are invariant quantities and also due to the diverse chemical bonding network in the host materials in our dataset. Nevertheless, we note that ASPH features can still improve the performance of the E3NN network.

Furthermore, in addition to the node features, we note that graph pooling layers will also affect the performance of GNN models on defect-related tasks. Previous GNN models for material property predictions usually use global mean (sum) pooling layers to collect all node features, which calculates the mean (sum) of those features as the global representation of the graph. However, this method includes the contributions from all node features and thus averages out the distinct features related to defects, as discussed above. Therefore, instead of global mean (sum) pooling, we choose to use global max pooling layers to extract the maximum features over all nodes, which will direct the focus of our GNN models to the most distinct features of atoms near the defect sites. The MAEs using different pooling operations are summarized in Table 2 (more discussions can be

Table 2. Comparison of MAE (eV) of the Transformer Models using Different global Pooling Layers on the O-Based Perovskite Dataset without and with ASPH Features

global pooling operations	max	sum	mean	min-max
without ASPH features	1.55	1.75	1.74	1.79
with ASPH features	0.72	0.90	0.86	0.85

found in Supporting Information II). Those results suggest that global max pooling layers are effective in extracting the defect information from the large supercell structures.

More importantly, the global max pooling layers, along with the ASPH features, can help overcome the previously reported limitation where the performance of GNN models is not convergent with respect to the size of the defective supercells.[24] Here, we demonstrate that our GNN model which uses global max pooling layers can produce converged MAE with respect to the supercell size. Since the crystal systems of our host (pristine) perovskites are diverse and the defective supercells are constructed nearly cubic, making it computationally expensive to increase the periodicity in all directions, we choose to focus only on host perovskites in the cubic and the tetragonal crystal system, where the transformation matrix to the defective supercell is diagonal, so that increasing the periodicity along Cartesian directions does not change the overall shape of the supercell significantly. Based on the constructed defective supercells (which already contain hundreds of atoms), we consecutively increase the size of the supercell only in the direction along which the lattice constants are the smallest, as shown in Figure 5A, while keeping the number of vacancies fixed, assuming that the defect formation energies from DFT methods remain constant. As the size of the supercell increases, the number of atoms within each supercell increases drastically, as shown in Figure 5B. In panel (C), the MAE of the Transformer GNN model using the global mean pooling layers increases gradually as the size of the supercell increases due to the "trivial" atoms irrelevant to the defect property. On the other hand the MAEs of our current model using the global max pooling layers and the ASPH features are convergent with respect to the supercell size (though the MAE is larger than that trained on the whole dataset, possibly due to less amount of
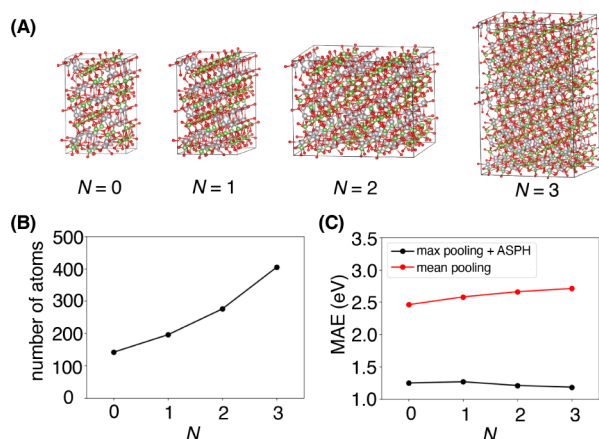
Figure 5. (A) Illustration of the defective supercell with increasing sizes, where $N = 0$ is the original defective supercell, and $N > 0$ represents the supercell by repeating the $N - 1$ cell along the directions along which the lattice constant is the smallest. (B) The average number of atoms of the defective supercells of different sizes. (C) The MAEs of the Transformer GNN model using global max pooling layers and ASPH features (black line), and using global mean pooling layers and without ASPH features (red line) on the dataset of defective supercells of different sizes.

Figure 6. (A) The comparison between the defect formation energies from the DFT method and the GNN model on the testing set of the BaTiO$_3$ dataset with multiple substitutions. The inset is a typical structure in the dataset, with the following color code: light green: Ba; dark green: Sr; gray: Ca; yellow: Zr; cyan: Ti; red: O. (B) Same as (A), but for the dataset of BaTiO$_3$ with multiple vacancies.

data), demonstrating that our GNN model can also adapt to the dilute defect limit.

Finally, we narrow our dataset down to structures with only O-vacancies. For the whole dataset, the accurate prediction of the vacancy formation energies requires information about the local chemical environment around the vacancy as well as the elemental type of the vacancy. As we focus on only O-vacancies, the prediction task is simpler than that on the whole dataset since the elemental type information on the vacancy is no longer needed. However, most of the host materials in our dataset still possess multiple symmetry-inequivalent O-sites with varying vacancy formation energies, making it challenging for previously proposed machine learning and GNN models to tackle. With our GNN model with ASPH features and global max pooling layers, the MAE is 0.45 eV, which is comparable to most previous machine learning and deep learning models predicting oxygen vacancy formation energies in oxides.[50−52]

Multi-defect Interaction Energy in BaTiO$_3$. As discussed above in the toy models, persistent homology features contain information not only about vacancies, but also about substitution defects. Therefore, we anticipate that those features can also improve the prediction performance for structures with substitutions. Here we use the cubic perovskite BaTiO$_3$ as an example and consider the defect formation energies of multiple substitutions. We construct the dataset containing 1000 defective BaTiO$_3$ in a $3 \times 3 \times 3$ supercell, and randomly replace at most 8 Ba (out of 27) atoms by Ca or Sr and at most 4 Ti (out of 27) atoms replaced by Zr; those substitutions are common in BaTiO$_3$.[53] By training our GNN model on this dataset, we find that after including the persistent homology features, the MAE decreases slightly from 0.07 to 0.03 eV, and the comparison between defect formation energies of DFT and GNN methods are shown in Figure 6a. We note that substitutional defects do not change the overall chemical bonding network of BaTiO$_3$, but only replace atoms by other chemical species randomly, similar to configurationally disordered materials.[34] In this case, even without the persistent homology features, our GNN model can already capture the
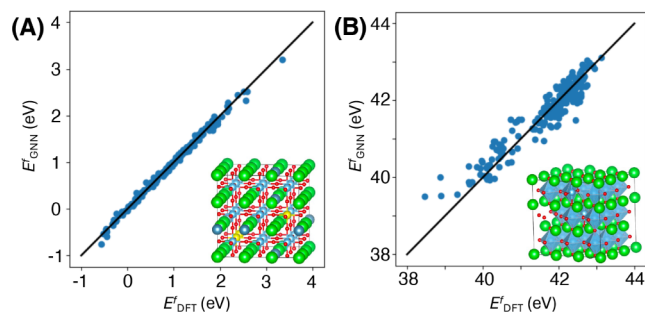
elemental swaps and make accurate predictions on the formation energies, and adding the persistent homology features only slightly increases the accuracy.

Furthermore, we consider BaTiO$_3$ with multiple vacancies. The above discussions revealed that persistent homology features can also capture the number of vacancies around each atom. Therefore, persistent homology features can also benefit predictions with multiple vacancies. We construct another data set containing 1000 defective BaTiO$_3$ in a $3 \times 3 \times 3$ supercell, and randomly remove (exactly) 1 Ba, 1 Ti, and 3 O atoms. Since we constrained the host material and also the number of vacancies, the energy fluctuations of different defect configurations arise from the vacancy−vacancy interactions. Therefore, the prediction target in fact reflects the multivacancy interaction energy; however, for consistency, we still call it the defect formation energy and calculate it from eq 1. By including the ASPH feature, the MAE decreases from 0.55 to 0.43 eV, where the comparison between DFT and GNN formation energies is shown in Figure 6b. The worse performance on tasks related to vacancy defects than to substitutional defects is because vacancies disrupt the chemical bonding network in materials and affect the exchange of information indirectly through the adjacency matrix of graphs, making it more challenging to capture the vacancy formation energies than substitution formation energies. Besides, by the construction of ASPH features (Figure 1), those features are naturally more sensitive to substitutions as exotic chemical species, while vacancies are obliquely encoded in the birth and death of 1-Betti numbers and 2-Betti numbers. We believe more direct encoding of the vacancy information in the persistent homology features, along with advanced GNN models which capture the interrupted chemical bonding network due to defects (such as higher-order GNNs), could further improve the prediction accuracy on defect properties.

■ CONCLUSIONS

In this work, we reveal the relationship between persistent homology features, obtained from the topological invariants of the homology group of point clouds, and the defect structures including multiple vacancies and substitutions. Through several toy models, we demonstrate that persistent homology features contain the local structural information near each atom, such as the type and the number of defects, the distance to those defects from each atom, and the size of the defects. These persistent homology features can be easily incorporated into the node

features in any existing GNN models and improve the performance of GNN models in predicting defect-related properties.

We numerically verify the role of persistent homology features using the dataset of O-based perovskite materials, which covers a wide spectrum of crystal systems and vacancies of various elemental types. We construct and train the GNN models on the dataset to predict the neutral vacancy formation energies. The results demonstrate that including the persistent homology features, along with global max pooling layers, can significantly increase the prediction accuracy of GNN models with the MAE reduced by around 55%, outperforming previous machine learning and GNN models. Our model also produces MAE that does not depend on the supercell sizes, allowing for accurate predictions of defect properties in the low density limit.

Furthermore, we demonstrate that persistent homology features can help predict the defect−defect interaction energies using a dataset of the defective cubic perovskite $BaTiO_3$ with multiple substitutions and with multiple vacancies. These results justify the essential role of persistent homology features in the accurate predictions of defect properties, such as defect formation energies and charge transition levels, using machine learning and deep learning methods, and will benefit future studies on high-throughput screening and design of the combination of host materials and specific defects with desired properties and functionalities.

◼ METHODS

Density Functional Theory Calculations. First-principles calculations are carried out using the Vienna Ab initio Simulation Package[54,55] with projector augmented wave pseudopotentials.[56,57] We used the Perdew−Burke−Ernzerhof functional in the generalized gradient approximation for all calculations.[58] The kinetic energy cutoff of the plane-wave basis sets is 550 eV, and the energy convergence threshold is 10^{-6} eV. The Γ **k**-point sampling scheme is used for all pristine and defective supercell calculations, and for elemental energy calculations we used the **k**-point grid with the **k**-point density of $0.03(2\pi/\text{Å})$. We performed the spin-polarized calculations with the DFT + U scheme for *3d* elements, with the *U* values taken from Materials Project.[35] The DFT-D3 method was also used to take account of the van der Waals interactions[59,60] in all calculations.

In this work we only consider the neutral charge state for all defects, and therefore the defect formation energy for the defect M_C, following the Kröger-Vink notation, is defined as

$$E^f[M_C] = E[M_C] - E[\text{bulk}] + \sum_i n_i \mu_i \tag{1}$$

where n_i is the difference in the number of elements between the defective and the pristine structures ($n_i > 0$ for removed atoms and $n_i < 0$ for added atoms), and μ_i is the chemical potential of element *i*, which we assumed to be the energy of the most stable single crystal of this element (except for H, N, O, F, Cl, which we calculate their energies from corresponding gas molecules).[7,8]

Training GNN Models. To train our GNN model, our dataset is split into the training set, validation set, and testing set according to 60:20:20 ratio. The validation set is used to prevent overfitting on the training set, which usually occurs when the dataset contains few data. The MAE on the validation set is tracked during the training procedure, and we use the model

with the minimal MAE on the validation set as our optimal model for testing.

The hyperparameters are optimized based on the Bayesian optimization method as implemented in Optuna,[61] which calculates the expected improvement from a new set of hyperparameters using the tree-structured Parzen Estimator method,[62] and decides whether to accept or decline this set of hyperparameters. The set of hyperparameters used in this work includes the number of convolution layers, the number of hidden channels in each convolution layer, the number of attention heads (used only in GNN models with the attention mechanism), learning rate, weight decay, and batch size.

◼ ASSOCIATED CONTENT

Data Availability Statement

The datasets and codes used in this work are available on Github at https://github.com/qmatyanlab/Defect_GNN.

Ⓢ Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.chemmater.4c03028.

> Detailed definitions of persistent homology features and additional benchmark results on the GNN model architecture (PDF)

◼ AUTHOR INFORMATION

Corresponding Authors

Zhenyao Fang − *Department of Physics, Northeastern University, Boston, Massachusetts 02115, United States;* ⓞ orcid.org/0000-0001-6841-3934; Email: z.fang@ northeastern.edu

Qimin Yan − *Department of Physics, Northeastern University, Boston, Massachusetts 02115, United States;* Email: q.yan@ northeastern.edu

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.chemmater.4c03028

Author Contributions

Z.F. and Q.Y. conceived the experiments. Z.F. wrote the code for graph neural networks and performed data analysis. Q.Y. supervised the study. The manuscript was written through contributions from all authors, and all authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

◼ REFERENCES

(1) Li, C.; Fu, T.; Hu, H.; Peng, X. Mechanical properties and their sensitivity to point defects: (HfNbTaTiZr)C high-entropy carbide. *Phys. Rev. B* **2022**, *105*, 224102.

(2) Boström, H. L. B.; Kieslich, G. Influence of Metal Defects on the Mechanical Properties of ABX3 Perovskite-Type Metal-formate Frameworks. *J. Phys. Chem. C* **2021**, *125*, 1467−1471.

(3) Dastider, A. G.; Rasul, A.; Rahman, E.; Alam, M. K. Effect of vacancy defects on the electronic and mechanical properties of two-dimensional MoSi2N4. *RSC Adv.* **2023**, *13*, 5307−5316.

(4) Pan, X.; Yang, M.-Q.; Fu, X.; Zhang, N.; Xu, Y.-J. Defective TiO₂ with oxygen vacancies: synthesis, properties and photocatalytic applications. *Nanoscale* **2013**, *5*, 3601−3614.

(5) Yin, W.-J.; Yang, J.-H.; Kang, J.; Yan, Y.; Wei, S.-H. Halide perovskite materials for solar cells: a theoretical review. *J. Mater. Chem. A* **2015**, *3*, 8926−8942.

(6) Leem, Y.-C.; Fang, Z.; Lee, Y.-K.; Kim, N.-Y.; Kakekhani, A.; Liu, W.; Cho, S.-P.; Kim, C.; Wang, Y.; Ji, Z.; Patra, A.; Kronik, L.; Rappe, A. M.; Yim, S.-Y.; Agarwal, R. Optically Triggered Emergent Meso-structures in Monolayer WS2. *Nano Lett.* **2024**, *24*, 5436−5443.

(7) Van de Walle, C. G.; Neugebauer, J. First-principles calculations for defects and impurities: Applications to III-nitrides. *J. Appl. Phys.* **2004**, *95*, 3851−3879.

(8) Freysoldt, C.; Grabowski, B.; Hickel, T.; Neugebauer, J.; Kresse, G.; Janotti, A.; Van de Walle, C. G. First-principles calculations for point defects in solids. *Rev. Mod. Phys.* **2014**, *86*, 253−305.

(9) Oba, F.; Choi, M.; Togo, A.; Tanaka, I. Point defects in ZnO: An approach from first principles. *Sci. Technol. Adv. Mater.* **2011**, *12* (3), 034302.

(10) Pochet, P.; Caliste, D. Point defect diffusion in Si and SiGe revisited through atomistic simulations. *Mater. Sci. Semicond. Process* **2012**, *15*, 675−690.

(11) Lany, S.; Zunger, A. Assessment of correction methods for the band-gap problem and for finite-size effects in supercell defect calculations: Case studies for ZnO and GaAs. *Phys. Rev. B* **2008**, *78*, 235104.

(12) Freysoldt, C.; Neugebauer, J. First-principles calculations for charged defects at surfaces, interfaces, and two-dimensional materials in the presence of electric fields. *Phys. Rev. B* **2018**, *97*, 205425.

(13) Frey, N. C.; Akinwande, D.; Jariwala, D.; Shenoy, V. B. Machine Learning-Enabled Design of Point Defects in 2D Materials for Quantum and Neuromorphic Information Processing. *ACS Nano* **2020**, *14*, 13406−13417.

(14) Deml, A. M.; Holder, A. M.; O'Hayre, R. P.; Musgrave, C. B.; Stevanović, V. Intrinsic Material Properties Dictating Oxygen Vacancy Formation Energetics in Metal Oxides. *J. Phys. Chem. Lett.* **2015**, *6*, 1948−1953.

(15) Wexler, R. B.; Gautam, G. S.; Stechel, E. B.; Carter, E. A. Factors Governing Oxygen Vacancy Formation in Oxide Perovskites. *J. Am. Chem. Soc.* **2021**, *143*, 13212−13227.

(16) Park, J.; Xu, B.; Pan, J.; Zhang, D.; Lany, S.; Liu, X.; Luo, J.; Qi, Y. Accurate prediction of oxygen vacancy concentration with disordered A-site cations in high-entropy perovskite oxides. *Npj Comput. Mater.* **2023**, *9* (1), 29.

(17) Yan, Q.; Kar, S.; Chowdhury, S.; Bansil, A. The Case for a Defect Genome Initiative. *Adv. Mater.* **2024**, *36* (11), 2303098.

(18) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* **2019**, *31*, 3564−3572.

(19) Reiser, P.; Neubert, M.; Eberhard, A.; Torresi, L.; Zhou, C.; Shao, C.; Metni, H.; Van Hoesel, C.; Schopmans, H.; Sommer, T.; Friederich, P. Graph neural networks for materials science and chemistry. *Commun. Mater.* **2022**, *3* (1), 93.

(20) Fung, V.; Zhang, J.; Juarez, E.; Sumpter, B. G. Benchmarking graph neural networks for materials chemistry. *Npj Comput. Mater.* **2021**, *7* (1), 84.

(21) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120*, 145301.

(22) Kazeev, N.; Al-Maeeni, A. R.; Romanov, I.; Faleev, M.; Lukin, R.; Tormasov, A.; Neto, A. H. C.; Novoselov, K. S.; Huang, P.; Ustyuzhanin, A. Sparse representation for machine learning the properties of defects in 2D materials. *Npj Comput. Mater.* **2023**, *9* (1), 113.

(23) Xiang, X.; Soh, D.; Dunham, S. Exploration of Deep Learning Models for Accelerated Defect Property Predictions and Device Design of Cubic Semiconductor Crystals. *J. Phys. Chem. C* **2024**, *128*, 8821−8829.

(24) Rahman, M. H.; Gollapalli, P.; Manganaris, P.; Yadav, S. K.; Pilania, G.; DeCost, B.; Choudhary, K.; Mannodi-Kanakkithodi, A. Accelerating defect predictions in semiconductors using graph neural networks. *APL Mach. Learn.* **2024**, *2* (1), 016122.

(25) Li, Q.; Han, Z.; Wu, X.-M. Deeper Insights Into Graph Convolutional Networks for Semi-Supervised Learning. *Proceedings Of The AAAI Conference On Artificial Intelligence*, AAAI, 2018, 32.

(26) Pun, C. S.; Lee, S. X.; Xia, K. Persistent-homology-based machine learning: a survey and a comparative study. *Artif. Intell. Rev.* **2022**, *55*, 5169−5213.

(27) Wu, K.; Zhao, Z.; Wang, R.; Wei, G.-W. TopP-S: Persistent homology-based multi-task deep neural networks for simultaneous predictions of partition coefficient and aqueous solubility. *J. Comput. Chem.* **2018**, *39*, 1444−1454.

(28) Cang, Z.; Wei, G.-W. TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PloS Comput. Biol.* **2017**, *13* (7), No. e1005690.

(29) Jiang, Y.; Chen, D.; Chen, X.; Li, T.; Wei, G.-W.; Pan, F. Topological representations of crystalline compounds for the machine-learning prediction of materials properties. *Npj Comput. Mater.* **2021**, *7* (1), 28.

(30) Roedel, J.; Jo, W.; Seifert, K. T. P.; Anton, E.-M.; Granzow, T.; Damjanovic, D. Perspective on the Development of Lead-free Piezoceramics. *J. Am. Ceram. Soc.* **2009**, *92*, 1153−1177.

(31) Dey, T. K.; Wang, Y. *Computational topology for data analysis*; Cambridge University Press, 2022.

(32) Aktas, M. E.; Akbas, E.; El Fatmaoui, A. Persistence homology of networks: methods and applications. *Appl. Netw. Sci.* **2019**, *4* (1), 61.

(33) Zia, A.; Khamis, A.; Nichols, J.; Tayab, U. B.; Hayder, Z.; Rolland, V.; Stone, E.; Petersson, L. Topological deep learning: A review of an emerging paradigm. *Artif. Intell. Rev.* **2024**, *57* (4), 77.

(34) Fang, Z.; Yan, Q. Towards accurate prediction of configurational disorder properties in materials using graph neural networks. *Npj Comput. Mater.* **2024**, *10* (1), 91.

(35) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1* (1), 011002.

(36) Morris, C.; Ritzert, M.; Fey, M.; Hamilton, W. L.; Lenssen, J. E.; Rattan, G.; Grohe, M.Weisfeiler and Leman Go Neural: Higher-Order Graph Neural Networks. Proceedings of the AAAI Conference on Artificial Intelligence*Proceedings Of The AAAI Conference On Artificial Intelligence*AAAI20193314602−4609

(37) Reiser, P.; Neubert, M.; Eberhard, A.; Torresi, L.; Zhou, C.; Shao, C.; Metni, H.; Van Hoesel, C.; Schopmans, H.; Sommer, T.; Friederich, P. Graph neural networks for materials science and chemistry. *Commun. Mater.* **2022**, *3* (1), 93.

(38) Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. *arXiv* **2018**.

(39) Brody, S.; Alon, U.; Yahav, E. How Attentive are Graph Attention Networks? *arXiv* **2022**.

(40) Shi, Y.; Huang, Z.; Feng, S.; Zhong, H.; Wang, W.; Sun, Y. Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence IJCAI-21*; International Joint Conferences on Artificial Intelligence Organization: 2021, pp 1548−1554. .

(41) Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. Graph neural networks: A review of methods and applications. *AI Open* **2020**, *1*, 57−81.

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

(43) Alkauskas, A.; Yan, Q.; Van de Walle, C. G. First-principles theory of nonradiative carrier capture via multiphonon emission. *Phys. Rev. B* **2014**, *90*, 075202.

(44) Freysoldt, C.; Neugebauer, J.; Van de Walle, C. G. Fully Ab Initio Finite-Size Corrections for Charged-Defect Supercell Calculations. *Phys. Rev. Lett.* **2009**, *102*, 016402.

(45) Choudhary, K.; DeCost, B. Atomistic Line Graph Neural Network for improved materials property predictions. *Npj Comput. Mater.* **2021**, *7* (1), 185.

(46) Jäger, M. O. J.; Morooka, E. V.; Federici Canova, F.; Himanen, L.; Foster, A. S. Machine learning hydrogen adsorption on nanoclusters through structural descriptors. *Npj Comput. Mater.* **2018**, *4* (1), 37.

(47) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754−13769.

(48) Himanen, L.; Jäger, M. O. J.; Morooka, E. V.; Canova, F. F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **2020**, *247*, 106949.

(49) Greenacre, M.; Groenen, P. J. F.; Hastie, T.; D'Enza, A. I.; Markos, A.; Tuzhilina, E. Publisher Correction: Principal component analysis. *Nat. Rev. Methods Primers* **2023**, *3* (1), 22.

(50) Wan, Z.; Wang, Q.-D.; Liu, D.; Liang, J. Data-driven machine learning model for the prediction of oxygen vacancy formation energy of metal oxide materials. *Phys. Chem. Chem. Phys.* **2021**, *23*, 15675−15684.

(51) Park, S.; Lee, N.; Park, J. O.; Park, J.; Heo, Y. S.; Lee, J. Exploring the Latent Chemical Space of Oxygen Vacancy Formation Energy by a Machine Learning Ensemble. *ACS Mater. Lett.* **2024**, *6*, 66−72.

(52) Baldassarri, B.; He, J.; Gopakumar, A.; Griesemer, S.; Salgado-Casanova, A. J. A.; Liu, T.-C.; Torrisi, S. B.; Wolverton, C. Oxygen Vacancy Formation Energy in Metal Oxides: High-Throughput Computational Studies and Machine-Learning Predictions. *Chem. Mater.* **2023**, *35*, 10619−10634.

(53) Zhao, C.; Huang, Y.; Wu, J. Multifunctional barium titanate ceramics via chemical modification tuning phase structure. *InfoMat* **2020**, *2*, 1163−1190.

(54) Kresse, G.; Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **1993**, *47*, 558−561.

(55) Kresse, G.; Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **1996**, *6*, 15−50.

(56) Kresse, G.; Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **1999**, *59*, 1758−1775.

(57) Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **1994**, *50*, 17953−17979.

(58) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.

(59) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104.

(60) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32*, 1456−1465.

(61) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-Generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*; ACM Digital Library: 2019, 2623−2631.

(62) Bergstra, J.; Bardenet, R.; Bengio, Y.; Kégl, B. Algorithms for Hyper-Parameter Optimization. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2011. vol: *24*.