CONDENSED MATTER PHYSICS

Structure motif–centric learning framework for inorganic crystalline systems

Huta R. Banjade[1†], Sandro Hauri[2†], Shanshan Zhang[2], Francesco Ricci[3], Weiyi Gong[1], Geoffroy Hautier[3,4], Slobodan Vucetic[2]*, Qimin Yan[1]*

Incorporation of physical principles in a machine learning (ML) architecture is a fundamental step toward the continued development of artificial intelligence for inorganic materials. As inspired by the Pauling's rule, we propose that structure motifs in inorganic crystals can serve as a central input to a machine learning framework. We demonstrated that the presence of structure motifs and their connections in a large set of crystalline compounds can be converted into unique vector representations using an unsupervised learning algorithm. To demonstrate the use of structure motif information, a motif-centric learning framework is created by combining motif information with the atom-based graph neural networks to form an atom-motif dual graph network (AMDNet), which is more accurate in predicting the electronic structures of metal oxides such as bandgaps. The work illustrates the route toward fundamental design of graph neural network learning architecture for complex materials by incorporating beyond-atom physical principles.

INTRODUCTION

Machine learning (ML) methods, in combination with massive material data, offer a promising route to accelerate the discovery and rational design of functional solid-state compounds by using a data-driven paradigm (*1*). Supervised learning has been effective in material property predictions, such as phase stability (*2*–*4*), crystal structure (*5*), effective potential for molecule dynamics simulations (*6*), and energy functionals for density functional theory–based simulations (*7*). With the recent progress in deep learning, ML has also been applied to inorganic crystal systems to learn from high-dimensional representations of crystal structures and to identify their complex correlations with materials properties. For instance, bandgaps of given classes of inorganic compounds have been predicted using deep learning (*8*), and ML has been applied on charge densities (*9*) and Hamiltonian data (*10*) to predict electronic properties. Recent development of graph convolutional network (GCN) (*11*, *12*), when combined with domain knowledge, offers a powerful tool to create an innovative representation of crystal structures for inorganic compounds. Within the GCN framework, any type of grid and atomic structure can be successfully modeled and analyzed. The flexible graph network structure endows these learning frameworks (*13*) a large room for improvement by considering more node/edge interactions in the crystal graphs (*14*).

Whether ML can efficiently approximate the unknown nonlinear map between input and output relies on an effective representation of solid-state compound systems that capture structure-property relationships that form the basis of many design rules for functional materials. In inorganic crystalline materials with unit cells that satisfy the periodic boundary condition, bonding environments determined by local and global symmetry are essential components for the understanding of complex material properties (*15*). As stated in the Pauling's first rule (*16*), a coordinated polyhedron of anions is formed about each cation in a compound, effectively creating structure motifs that behave as fundamental building blocks and are highly correlated with material properties.

Structure motifs in crystalline compounds play an essential role in determining the material properties in various scientific and technological applications. For instance, the identification of VO_4 functional motif enabled the discovery of 12 vanadate photoanode materials via high-throughput computations and combinatorial synthesis (*17*). In the field of complex oxide devices, MnO_6 octahedral motifs are correlated with small hole polarons that limit electrical conductivity (*18*). In battery cathodes for energy storage, high ion mobility is explained by the local bonding environment of a multivalent ion (*19*). V^{4+} ion-related motifs and the connections between these motifs are found to be important determining factors for the selective oxidation of hydrocarbons (*20*–*22*). The presence of MO_4 tetrahedra (M as Si or Al) can be used to identify the most promising synthetic candidates from the pool of hypothetical zeolites (*23*). When designing novel battery materials, it is found that the changing coordination pattern of a migrating ion can be used as a descriptor of ion mobility (*24*, *25*).

Governing the structure-property relationship, structure motifs or coordination environments can be viewed as effective structural descriptors for crystals. The efforts for identification of local coordination environments initially focused on structure types (*26*, *27*) or preferential coordination numbers (*28*) based on simple rules (*29*). Very recently, owing to the development of data-driven approaches, systematic and robust approaches to automatically identify local environments have been developed (*30*, *31*), which motivated the use of structure motif information for material design in a data-driven paradigm. For instance, structure motif information has been used to define crystal structure similarity (*32*) for all the compounds in the Materials Project database (*33*). A recent work comprehensively evaluated the validity and suggested the limited predictive power of the Pauling rules (*34*). Recent analysis and the dataset of local environment and connectivity (*30*) provide a

[1]Department of Physics, Temple University, Philadelphia, PA 19122, USA. [2]Department of Computer and Information Science, Temple University, Philadelphia, PA 19122, USA. [3]Institute of Condensed Matter and Nanoscience (IMCN), Université catholique de Louvain (UCLouvain), Chemin étoiles 8, bte L7.03.01, Louvain-la-Neuve 1348, Belgium. [4]Thayer School of Engineering, Dartmouth College, Hanover, NH 03755, USA.
*Corresponding author. Email: qiminyan@temple.edu (Q.Y.); vucetic@temple.edu (S.V.)
†These authors contributed equally to this work.

novel set of material information that can serve as essential input for ML techniques in materials science.

In this work, we propose to incorporate structure motif information in an ML framework. We show that the presence of structure motifs and their connections extracted from a material structure database can be used by unsupervised learning algorithms to define unique representations in a high-dimensional space. The dimension reduction process reveals strong clustering effects, representing the neighborhood properties of metal elements in the periodic table. By combining the motif information with graph convolutional neural networks, we develop a motif-centric deep learning architecture called the atom-motif dual graph neural network (AMDNet), whose accuracy surpasses that of the state-of-the-art atom-based graph network MatErials Graph Network (MEGNet) (*12*) for the prediction of electronic structures of inorganic crystalline materials.

RESULTS

Structure motifs clustering by unsupervised learning

In a recent work (*35*), it is shown that an unsupervised learning algorithm Atom2Vec can learn high-dimensional vector representations of atoms that encode basic properties of atoms by using an extensive database of chemical formulas. Clustering of atoms in the vector space classifies them into groups consistent with the periodic table. Furthermore, it is possible to use vector representations of atoms to calculate the similarities among materials and make property predictions. In this work, we will enhance the previous development by demonstrating that structure motifs encoded in crystal structures reveal useful information about structural properties and electronic structures of crystalline systems.

We focus on binary and ternary metal oxides that constitute a vast and diverse material space where crystal structures are well characterized by local environments through cation-oxygen coordination. The material set includes 22,606 complex oxides in the Materials Project database (*33*). We extract the structure motif information using the local environment identification method developed by Waroquiers *et al.* (*30*) as implemented in the pymatgen code (*36*), following the definition of structure motifs or coordination environments by the International Union of Crystallography (*37*) and International Union of Pure and Applied Chemistry (*38*) as listed in *30*.

We identify the connections between a motif and its neighboring motifs based on the number of oxygen atoms shared by those motifs. Three different types of connectivity may exist, from which we identify the connections as corner sharing (if only one atom is shared), edge sharing (if two atoms are shared), and face sharing (if three or more than three atoms are shared). Details on the motif type and connectivity are included in the notes S1 and S2. The motif environment is defined by the neighboring motifs and the type of connection a motif has. By iterating through all the structures in the dataset, motif-environment pairs are identified, and the motif environment matrix is generated. Details on the motif environment matrix are included in note S3.

Next, we propose the learning algorithm that is able to take advantage of the above motif data collection process and convert each row of the motif environment matrix effectively into a high-dimensional vector that represents a unique structure motif. To create the vector representations for structure motifs, we treat motifs as the basic building blocks and study their presence and motif-wise environment in 22,606 oxide crystal structures extracted from the Materials Project database. Figure 1 shows the high-level representation of the workflow used in the unsupervised learning algorithm. Material properties, such as orbital interactions within a crystal, are known to be related to bond lengths and bonding angles. We extract the following quantities to represent motif connections: (i) the distance between the cation center of a motif M_1 and its neighboring motif center (M_2) and (ii) the M_1-O-M_2 bonding angles for those oxygen atoms shared by the two motifs. The extracted motif connection information will be an essential input for the learning process using GCN as described below.

Our aim is to identify patterns and clustering information for these high-dimensional motif vectors that, in turn, influence the complex material properties of oxide compounds. By using various linear and nonlinear transformations, dimension reduction algorithms serve this purpose by creating a low-dimensional representation (called embedding) that best preserves the overall variance of the original dataset. To demonstrate the clustering of the motif vectors, we visualize the high dimensional data by using the t-distributed stochastic neighbor embedding (t-SNE) (*39*), a recently developed nonlinear dimensionality reduction technique. Before the t-SNE, we apply singular value decomposition (*40*) to project the original high-dimensional representation of materials to 60 dimensions, corresponding to the largest 60 singular values. The detailed procedure for t-SNE is presented in note S4.

Figure 2 shows the projected motif vector data in two dimensions obtained through the t-SNE process, where different motif types are represented by different colors. We observe that there exist distinct clusters based on the motif types. First, detailed analysis of those clusters shows that the chemical properties of the elements forming the motifs play an important role in the formation of clusters. For instance, all the Lanthanide-based motifs formed different clusters on the basis of motif type (cluster 1 in Fig. 2 and cluster 9 in fig. S3 in note S4). Yttrium-based motifs always stay close to Lanthanide-based ones, as the chemical properties of Yttrium are known to be similar to Lanthanides. In addition, motifs associated with Zn and Mg always cluster together, which is consistent with the fact that Zn is chemically similar to Mg because both of them exhibit only one normal oxidation state ($^{+}2$) and that their ions (Zn^{2+} and Mg^{2+}) are similar in size.

As shown in Fig. 2, cluster 1 contains cubic motifs associated with Lanthanides, while the cuboctahedral motifs associated primarily with main group elements appear in cluster 2. The clustering of motifs, determined by elements as described above, is in accordance with the grouping pattern in the periodic table, although no information about the periodic table was used in the vectorization process. Octahedral motifs associated mostly with the transition metal elements occur together in cluster 3, while the tetrahedral and square planer motifs associated with transition metal elements are located but well separated in cluster 4. This motif cluster separation reveals that the vectorization process based on the matrix environment matrix is able to capture both local bonding environment information and elemental information. Additional motif clusters in Fig. 2 are presented in fig. S3 in note S4. These findings, achieved by unsupervised learning, strongly support our intuition that structure motifs can serve as essential input for crystalline compounds that carry both elemental and structural information.

Incorporation of motif information in graph neural networks

As above atomic–level building blocks of crystals, structure motifs and motif-wise interactions within a crystal strongly influence the

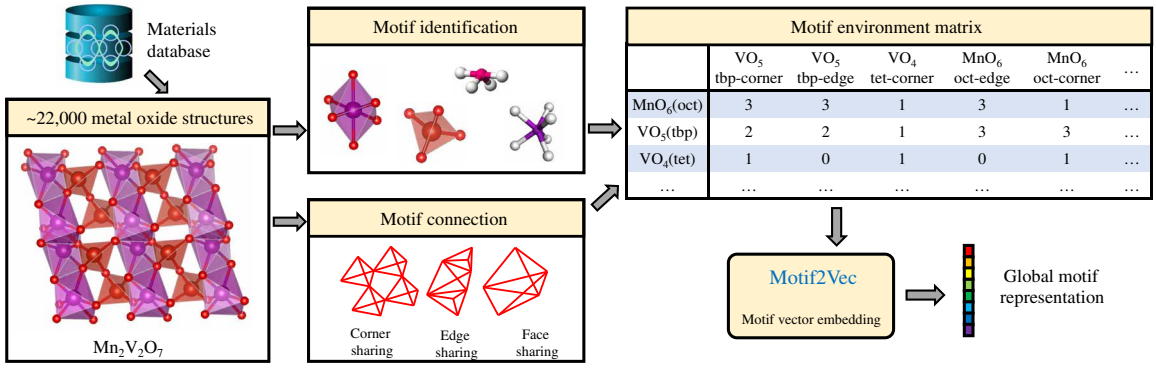

Fig. 1. Extraction of structure motif information in inorganic crystalline compounds (metal oxides) and the generation of global motif representations using the motif environment matrix.

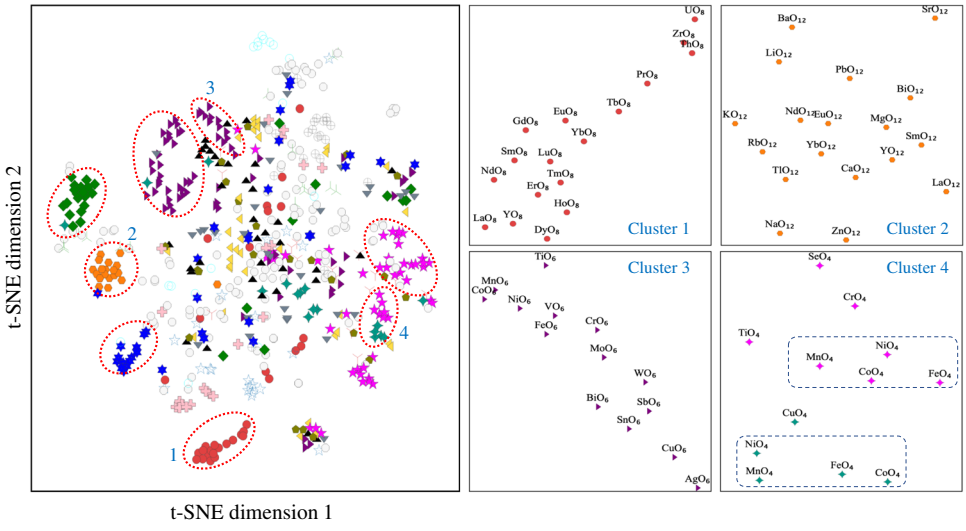

Fig. 2. The t-distributed stochastic neighbor embedding projection of motif vectors constructed by using the motif environment matrix. The motif clusters 1 to 4 are associated with various motif types including (1) cube, (2) cuboctahedron, (3) octahedron, and (4) a mixture of tetrahedron (in magenta) and square plane (in remnant). t-SNE, t-distributed stochastic neighbor embedding.

material properties. Structure motif information can be used as an essential input to a graph neural network (GNN) that predicts physical properties of materials. Following the standard notation used in the GNN framework (*41*), we represent an attributed graph as $G = (V, E)$, where $V = [\mathbf{v}_i]_{i=1,2,\ldots,N^v}$ is a set of nodes of cardinality N^v and $\mathbf{v}_i$ is the node attribute vector of the i^{th} node. $E = [(\mathbf{e}_k, r_k, s_k)]_{k=1,2,\ldots,N^e}$ is a set of edges of carnality N^e, where $\mathbf{e}_k$ is the attribute vector for edge k between nodes s_k and r_k. Several GNNs have been proposed (*11*, *12*, *14*) that formulate the task of predicting chemical properties of materials as learning a mapping $f(G:W) \rightarrow y$, where W is a set of learnable parameters and y is a target property.

Most of the graph networks applied to crystalline materials (*11*, *12*, *14*) are based on graphs on the atomic level G_0^{atom} as input for the network. These atomic graphs contain information about atoms (such as atomic number, electronegativity, and many others) and bonds. For instance, in the G_0^{atom} of atomic graph network MEGNet, $\mathbf{v}_i$ is a vector representing the i^{th} atom in a unit cell and is represented by the atomic number of the element. $\mathbf{e}_{ij}$ is a vector representing a bond between atom i and atom j.

In this work, to enable a learning architecture that synthesize both atom-level and motif-level graph representation of materials, we propose that AMDNet can be constructed to enhance the learning process and improve the prediction accuracy for electronic structure properties of metal oxides. We follow the procedure introduced in existing atomic graph networks (*12*, *42*) to represent the edges, where two atoms are connected if they are no more than 5 Å apart. We propose to represent the metal oxides as motif graphs G_0^{motif}, where each motif in a crystal is represented by a node $(v_i)_{G_0^{motif}}$ and each connection between two motifs is represented by an edge $(e_{ij})_{G_0^{motif}}$ as shown in Fig. 3. Motif graphs represent the same materials with higher granularity than atom graphs, but more comprehensive information can be encoded in each motif node, such as local distortions and site symmetries. The motif graph uses the same edge representation as in the atom graph, and the motif-motif edge distances are measured from the center atom of one motif to that of a neighboring motif.

In the motif graph, a combination of atom-level and motif-level information is encoded in each node. We adopt the atom-level node

representations by combining two existing approaches to form a 103-dimensional vector that uses the information of atoms within the motif. The first 86 dimensions represent the fractional encoding of the atoms proposed by Meredig *et al.* (*43*), and the next 17 dimensions are for physical properties proposed by Ward *et al.* (*44*). On the other hand, we define the motif fingerprints by order parameters (of dimension 61), which describe the numerical measure of the local environment around an atom relative to a target standard motif (*31*, *45*). These 61-dimension vectors are then concatenated with 103-dimensional atom-based feature to form the final 164-dimensional vector. Detailed descriptions about various types of order parameters and methods to compute these parameters are presented in the work by Zimmermann *et al.* (*31*). All the structural information used to construct the motif graph—including extended connectivity, angle, distance, and order parameters for each motif—is computed by using the python package robocrystallography (*46*) combined with the pymatgen code. By combining atomic-level and motif-level information, we use a 164-dimensional vector to represent each motif in the graph.

AMDNet

A high-level illustration of our proposed AMDNet architecture is shown in Fig. 4. To incorporate the motif information acquired above into the graph network learning framework, the central concept in the proposed architecture is to generate both motif graphs and atom graphs representing the same compounds, with different cardinality of edges and nodes, and combine the representation information before making predictions.

For each material, we generate an atom graph and a motif graph (Fig. 4). We adopt the convolution structure of the MEGNet proposed by Chen *et al.* (*12*) when constructing the atom-level graph network. The choice of graph network structure is only for a benchmark purpose, and many other types of crystal graph convolution networks could be used to take advantage of the motif-level graph information (*11*, *14*). As a preliminary test, we use the same architecture as that for the atom graphs in MEGNet to generate G_0^{motif} by using the 164-dimensional atom-motif-mixed vector input for the nodes in the network. Edges in G_0^{motif} are defined as the distances between the center atoms of any two motifs. Note that MEGNet can be interpreted as a neural network that encodes the whole crystal graph input to a low-dimensional vector of dimension 16, upon which a final single-value prediction is made. Taking advantage of this fixed-dimension representation of any MEGNet graph convolution network, we can effectively combine the information from motif and atom dual graphs by concatenating the two low-dimensional representations generated from motif graph and atom graph, respectively. This concatenated vector is then fed to a small feed-forward neural network for single-value predictions. More details are presented in Materials and Methods.

We use 22,606 binary and ternary metal oxides from the Materials Project database to evaluate the effectiveness of our proposed model and focus on the prediction of bandgaps which is one of the complex electronic structure problems. Metal oxides are a class of solid-state compounds that are challenging for both ab initio quantum simulations and ML in general, which is verified by our experiments on different datasets as presented in note S5. For the purpose of comparison, we create a motif graph network model, MNet, which use motif graphs (G_0^{motif}) as the only input to the network. Table 1 shows a comparison between MEGNet, MNet, and our proposed AMDNet on the prediction accuracy of bandgaps, formation energy, and the metal (compounds with bandgaps less than 0.2 eV in the Materials Project database) versus nonmetal (compounds with bandgaps 0.2 eV or greater) classification for all the metal oxides in our dataset. Additional data and discussions are provided in note S5.

The results show that, given the same training and test data, AMDNet shows its superiority in the bandgap prediction task compared to the state-of-the-art baseline model. The motif graph representation MNet performs worse than MEGNet, which is expected because it uses a much smaller graph representation. The combination of atom and motif graph AMDNet outperforms MEGNet on the bandgap prediction task, which illustrates that the motif representations enhance the effective learning of material properties. Figure 4B shows the comparison of the predicted bandgaps on the test dataset with the actual bandgaps. Bandgaps of a large portion of the compounds are clustered close to the diagonal, indicating a

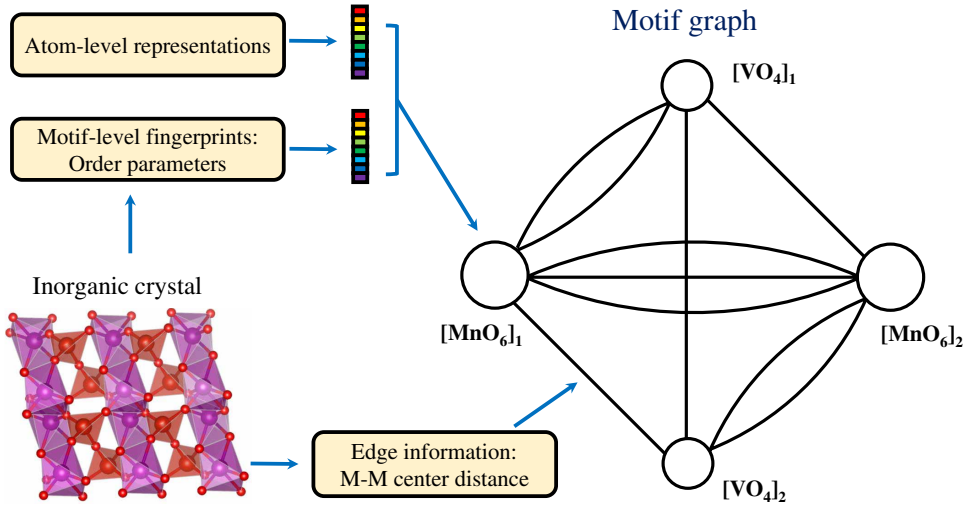

Fig. 3. Construction of a motif graph based on both atom-level and motif-level information encoded in an inorganic crystal.

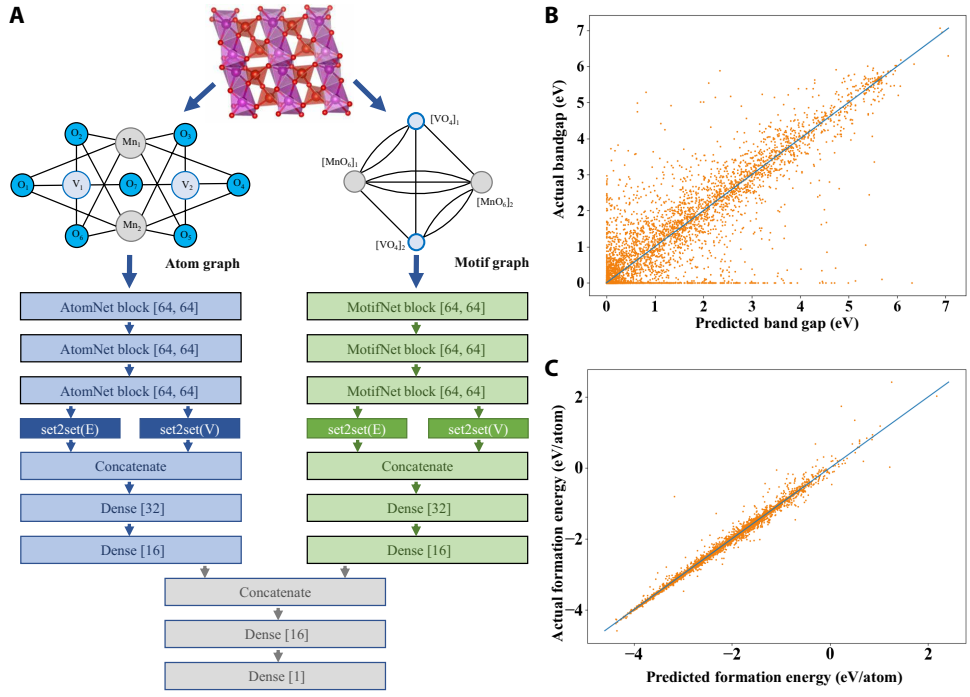

Fig. 4. AMDNet architecture and materials property predictions. (**A**) Demonstration of the learning architecture of the proposed atom-motif dual graph network (AMDNet) for the effective learning of electronic structures and other material properties of inorganic crystalline materials. (**B**) Comparison of predicted and actual bandgaps [from density functional theory (DFT) calculations] and (**C**) comparison of predicted and actual formation energies (from DFT calculations) in the test dataset with 4515 compounds.

Table 1. Performance comparison between various graph architectures for the learning and prediction of electronic bandgaps, formation energy per atom, and metal versus nonmetal classification accuracy for the metal oxides (trained on 18,091 compounds and tested on 4515 compounds). Both mean absolute error (MAE) and root mean square error (RMSE) are given for the purpose of comparison.

Model	Bandgap MAE/ RMSE (eV)	Formation energy MAE/ RMSE (eV/ atom)	Metal versus nonmetal classification accuracy
MEGNet (atom graph)	0.54 / 0.82	0.047 / 0.104	75.3%
MNet (motif graph)	0.64 / 1.03	0.121 / 0.236	74.7%
AMDNet (motif-atom dual graph)	0.44 / 0.78	0.047 / 0.100	82.1%

good performance of our model on the bandgap prediction task. In addition, our model shows superior performance in the metal versus nonmetal classification task. As shown in Table 1, the classification accuracy is 82.1% for AMDNet, while for MEGNet, it is only 75.3%. On the other hand, the formation energy prediction shows almost identical performance with MEGNet, indicating that atom graph alone is sufficient for the formation energy prediction task, which is considered a simpler task compared to the bandgap prediction task. The comparison between predicted (by AMDNet) and actual formation energies is shown in Fig. 4C, and the comparison of prediction accuracy given by various models is shown in Table 1. We also perform additional training and test using another state-of-the-art atom-based GNN model, crystal graph convolutional neural networks (CGCNN) (*11*), based on the same material dataset. AMDNet outperforms both CGCNN and MEGNet in general for all the three learning tasks (see note S6 and table S3 for details).

Note that the root mean square errors for bandgap predictions are larger than the mean absolute errors (see Table 1), which is demonstrated as the existence of outliers in Fig. 4B. Similar trend is observed using both AMDNet and MEGNet, indicating that this is originated from the complexity of material-property relationships for bandgaps of solid-state systems. Therefore, the deep learning of electronic band structure–related properties in solids naturally goes beyond atomic bonds and motifs. Novel and higher-level material information such as orbital interactions determined by local site symmetries (irreducible representations), which is out of the scope of this work, should be incorporated in the learning framework to improve the prediction performance. Despite this, we would like to emphasize that the inclusion of motif information in the AMDNet adds another tier of important material information that is helpful to distinguish the electronic structures of several representative set of oxide materials. The analysis of graph embeddings from motif neural networks in the AMDNet can capture the essential correlations between the motif types/connections in crystals and the electronic band structures of solid-state materials (see note S7 for details). Therefore, AMDNet serves as one of the initial efforts to incorporate higher-level material information in deep learning models for solid-state materials.

DISCUSSION

We demonstrate in this work how structure motifs in crystal structures can be combined with both unsupervised and supervised ML techniques to enhance the effective representation of solid-state material systems. Motif vectors learned from motif environments in 22,606 metal oxides using unsupervised learning effectively capture the motif similarities and their clustering properties. To enhance the learning of solid-state crystalline systems for complex electronic structures, structure motif and connection information are incorporated as essential input in an AMDNet model, which outperforms the state-of-the-art atom GNN model for the prediction of electronic bandgaps and metal versus nonmetal classification task. In addition, AMDNet model is able to predict formation energy in close agreement with the existing state-of-the-art atom graph–based models. Furthermore, AMDNet outperforms the existing atom-based model for the prediction of metal oxidation states in complex oxides such as Cr-based systems (see note S8 for details). This experiment, together with the outstanding performance of a recent motif information–enhanced shallow learning model for oxidation states in metal-organic frameworks (*47*), clearly demonstrates that structure motif is an essential layer of material information to achieve advanced learning models that can predict material properties beyond the state-of-the-art.

Although our work is limited to structure motifs that are identified in metal oxides by using domain knowledge, the general applicability of the model can be realized. First of all, the automatic motif identification approach (*31*) we used in this work was applied in the original work to identify the structure motifs in all the inorganic compounds in the Materials Project database, including diverse classes of materials that go far beyond oxides. This indicates that the proposed motif-centric learning model can be readily expanded to all the inorganic crystalline materials. In addition, the general applicability of motif-centric models can be further enhanced by using the technologies that are under development in the field of GNNs. There has been a recent exciting development to train GNNs in a self-supervised manner to automatically extract graph motifs from large graph datasets of molecules (*48*). The self-supervised learning-guided graph analysis can be applied to crystal graphs to enable a general learning architecture for automatic motif identification in crystal structures and its consequent use in graph-based neural networks for various downstream tasks. Therefore, our work provides the important initial step toward a general motif-centric GNN learning model that can be applied to arbitrary crystal systems.

AMDNet is a general learning framework for solid-state atomistic systems that can be used to predict other materials properties, such as mechanical and excited state properties, and applied to other motif-based systems including two-dimensional materials and metal-organic frameworks. Several directions related to the motif-centric learning methods here are worthy to explore in the future. Although we perform the test on perfect crystalline systems, through the addition of extra types of local motif information, the motif-enhanced graph network framework can be expanded for the learning and prediction of surface and defective material systems. Besides the use of a dual graph network architecture, motif information and the physical principles behind it can be incorporated into a learning framework in other manners, such as through a motif-enhanced convolutional process in an atom-based GCN or other novel algorithms that are actively developing in the graph theory including graph attention.

MATERIALS AND METHODS

Training process for atom motif dual GNN

In the AMDNet with L layers, the module generates a sequence of atomic graph representation ($G_1^{\text{atom}}, G_2^{\text{atom}}, \ldots, G_L^{\text{atom}}$) and motif graph representations ($G_1^{\text{motif}}, G_2^{\text{motif}}, \ldots, G_L^{\text{motif}}$), where each graph has the same number of nodes and edges as in the input graphs G_0^{atom} and G_0^{motif}, respectively. Through a graph convolutional process called AtomNet block for atom graphs and MotifNet block for motif graphs, information of each edge and its respective connecting nodes are passed through a dense neural network with a nonlinear activation function (we use the shifted softplus function), which creates a new edge representation. To generate the new node representation, the node information, together with the information of the new incident edges, is passed through a separate dense neural network with the same nonlinear activation function.

Each graph convolutional block has a hidden dimension of 64 for both node and edge convolution. In our work, we use three graph convolutional blocks to apply the graph convolution, which creates an output graph representation. The graph representation is transformed into vector form by averaging over all nodes and edges, respectively, which is denoted as set2set(E) and set2set(V) in Fig. 4A. These set2set vectors are concatenated before going through two densely connected layers as shown in Fig. 4A. This results in a low-dimensional vector representation of the original atom and motif graph representation of the crystal. These representations are concatenated again and passed through two densely connected layers to make a single real-valued prediction.

For the training and test process, we choose a 60-20-20 train validation test splits. We initialize the hyperparameters based on the best values from MEGNet to train our neural network. All deep models are trained with Adam optimizer (*49*) with initial learning rate $\alpha = 0.001$. Training formation energy prediction was slower to converge to the best solutions than for the bandgap prediction; therefore, we adjusted some parameters to adapt to each prediction task. We stop training when the validation error does not improve for 20 and 100 epochs to train bandgap prediction and formation energy prediction, respectively. We save the model with the lowest observed validation error and use it to evaluate the models on the test data. We use 64 compounds per minibatch for bandgap prediction and 32 compounds per minibatch for formation energy prediction. Note that in the network setup of AMDNet, it is possible that the model predicts negative bandgaps. We truncate the negative bandgap values to 0 for evaluation purposes. The procedure we apply here is the same as what was used by the atom-based model MEGNet. This treatment allows the use of the same model for both formation energy and bandgap predictions.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/7/17/eabf1754/DC1

REFERENCES AND NOTES

1. N. S. A. Technology, *Materials Genome Initiative for Global Competitiveness* (Gen. Books 2011).
2. C. J. Long, J. Hattrick-Simpers, M. Murakami, R. C. Srivastava, I. Takeuchi, V. L. Karen, X. Li, Rapid structural mapping of ternary metallic alloy systems using the combinatorial approach and cluster analysis. *Rev. Sci. Instrum.* **78**, 072217 (2007).
3. R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. Von Lilienfeld, Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 140022 (2014).

4. R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. Von Lilienfeld, Big data meets quantum chemistry approximations: The Δ-machine learning approach. *J. Chem. Theory Comput.* **11**, 2087–2096 (2015).
5. S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, G. Ceder, Predicting Crystal Structures with Data Mining of Quantum Calculations. *Phys. Rev. Lett.* **91**, 135503 (2003).
6. L. Zhang, J. Han, H. Wang, R. Car, E. Weinan, Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Phys. Rev. Lett.* **120**, 143001 (2018).
7. J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, K. Burke, Finding density functionals with machine learning. *Phys. Rev. Lett.* **108**, 253002 (2012).
8. Y. Dong, C. Wu, C. Zhang, Y. Liu, J. Cheng, J. Lin, Bandgap prediction by deep learning in configurationally hybridized graphene and boron nitride. *npj Comput. Mater.* **5**, 26 (2019).
9. S. Gong, T. Xie, T. Zhu, S. Wang, E. R. Fadel, Y. Li, J. C. Grossman, Predicting charge density distribution of materials using a local-environment-based graph convolutional network. *Phys. Rev. B* **100**, 184103 (2019).
10. H. Bai, P. Chu, J.-Y. Tsai, N. Wilson, X. Qian, Q. Yan, H. Ling, Graph neural network for hamiltonian-based material property prediction. arXiv:2005.13352 (2020).
11. T. Xie, J. C. Grossman, Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
12. C. Chen, W. Ye, Y. Zuo, C. Zheng, S. P. Ong, Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* **31**, 3564–3572 (2019).
13. Z. Zhang, P. Cui, W. Zhu, Deep Learning on Graphs: A Survey. *IEEE Trans. Knowl. Data Eng.* **14**, 1–1 (2020).
14. C. Park, C. Wolverton, Developing an improved Crystal Graph Convolutional Neural Network framework for accelerated materials discovery. *Phys. Rev. Mater.* **4**, 63801 (2020).
15. P. Zhang, A. C. To, Point group symmetry and deformation-induced symmetry breaking of superlattice materials. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **471**, 20150125 (2015).
16. L. Pauling, The Nature of the chemical bond. III. The transition from one extreme bond type to another. *J. Am. Chem. Soc.* **54**, 988–1003 (1932).
17. Q. Yan, J. Yu, S. K. Suram, L. Zhou, A. Shinde, P. F. Newhouse, W. Chen, G. Li, K. A. Persson, J. M. Gregoire, J. B. Neaton, Solar fuels photoanode materials discovery by integrating high-throughput theory and experiment. *Proc. Natl. Acad. Sci.* **114**, 3040–3043 (2017).
18. H. Peng, P. F. Ndione, D. S. Ginley, A. Zakutayev, S. Lany, Design of semiconducting tetrahedral Mn1-xZnxO alloys and their application to solar water splitting. *Phys. Rev. X* **5**, 021016 (2015).
19. M. Liu, Z. Rong, R. Malik, P. Canepa, A. Jain, G. Ceder, K. A. Persson, Spinel compounds as multivalent battery cathodes: A systematic evaluation based on ab initio calculations. *Energy Environ. Sci.* **8**, 964–974 (2015).
20. A. Karpov, C. K. Dobner, R. Glaum, S. A. Schunk, F. Rosowski, Catalytic properties of silver vanadium phosphates in n-butane oxidation-Considerations on the impact of the [VxOy] substructure. *Chem. Ing. Tech.* **83**, 1697–1704 (2011).
21. R. Schlögl, Active Sites for Propane Oxidation: Some Generic Considerations. *Top. Catal.* **54**, 627–638 (2011).
22. N. Mizuno, *Modern Heterogeneous Oxidation Catalysis : Design, Reactions, and Characterization* (Wiley-VCH, 2009).
23. Y. Li, J. Yu, R. Xu, Criteria for zeolite frameworks realizable for target synthesis. *Angew. Chem. Int. Ed.* **52**, 1673–1677 (2013).
24. Z. Rong, R. Malik, P. Canepa, G. Sai Gautam, M. Liu, A. Jain, K. Persson, G. Ceder, Materials Design Rules for Multivalent Ion Mobility in Intercalation Structures. *Chem. Mater.* **27**, 6016–6021 (2015).
25. D. Di Stefano, A. Miglio, K. Robeyns, Y. Filinchuk, M. Lechartier, A. Senyshyn, H. Ishida, S. Spannenberger, D. Prutsch, S. Lunghammer, D. Rettenwander, M. Wilkening, B. Roling, Y. Kato, G. Hautier, Superionic Diffusion through Frustrated Energy Landscape. *Chem* **5**, 2450–2460 (2019).
26. J. Dshemuchadse, W. Steurer, Some statistics on intermetallic compounds. *Inorg. Chem.* **54**, 1120–1128 (2015).
27. P. Villars, K. Cenzual, J. Daams, Y. Chen, S. Iwata, Data-driven atomic environment prediction for binaries using the Mendeleev number: Part 1. Composition AB. *J. Alloys Compd.* **367**, 167–175 (2004).
28. V. A. Blatov, Voronoi-Dirichlet polyhedra in crystal chemistry: Theory and applications. *Crystallogr. Rev.* **10**, 249–318 (2004).
29. R. Hoppe, The Coordination Number. *Angew. Chem. Int.* **9**, 25–34 (1970).
30. D. Waroquiers, X. Gonze, G. M. Rignanese, C. Welker-Nieuwoudt, F. Rosowski, M. Göbel, S. Schenk, P. Degelmann, R. André, R. Glaum, G. Hautier, Statistical analysis of coordination environments in Oxides. *Chem. Mater.* **29**, 8346–8360 (2017).
31. N. E. R. Zimmermann, M. K. Horton, A. Jain, M. Haranczyk, Assessing Local Structure Motifs Using Order Parameters for Motif Recognition, Interstitial Identification, and Diffusion Path Characterization. *Front. Mater.* **4**, 34 (2017).
32. A. Zimmermann, Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity. *RSC Adv.* **10**, 6063–6081 (2020).
33. A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 11002 (2013).
34. J. George, D. Waroquiers, D. Di Stefano, G. Petretto, G.-M. Rignanese, G. Hautier, The Limited Predictive Power of the Pauling Rules. *Angew. Chem. Int. Ed.* **59**, 7569–7575 (2020).
35. Q. Zhou, P. Tang, S. Liu, J. Pan, Q. Yan, S.-C. Zhang, Learning atoms for materials discovery. *Proc. Natl. Acad. Sci.* **115**, E6411–E6417 (2018).
36. S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, G. Ceder, Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
37. J. Lima-de-Faria, E. Hellner, F. Liebau, E. Makovicky, E. Parthé, Nomenclature of inorganic structure types. Report of the International Union of Crystallography Commission on Crystallographic Nomenclature Subcommittee on the Nomenclature of Inorganic Structure Types. *Acta Crystallogr. Sect. A* **46**, 1–11 (1990).
38. R. M. Hartshorn, E. Hey-Hawkins, R. Kalio, G. J. Leigh, Representation of configuration in coordination polyhedra and the extension of current methodology to coordination numbers greater than six (IUPAC technical report). *Pure Appl. Chem.* **79**, 1779–1799 (2007).
39. L. van der Maaten, G. Hinton, Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
40. C. Rincon Castaneda, S. Tapia Rodriguez, L. I. Gonzalez Luna, M. Ortiz Ramirez, Comparacion De Inhibicion De La Prueba Cutanea De La Histamina Con Astemizole, Loratadina Y Terfenadina. *Rev. Alerg. Mex.* **40**, 86–90 (1993).
41. P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. F. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, Ç. Gülçehre, H. F. Song, A. J. Ballard, J. Gilmer, G. E. Dahl, A. Vaswani, K. R. Allen, C. Nash, V. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, R. Pascanu, Relational inductive biases, deep learning, and graph networks. arXiv:1806.01261 (2018).
42. K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, K.-R. Müller, SchNet – a deep learning architecture for molecules and materials. *J. Chem. Phys.* **148**, 241722 (2018).
43. B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, C. Wolverton, Combinatorial screening for new materials in unconstrained composition space with machine learning Prediction. *Phys. Rev. B* **89**, 94104 (2014).
44. L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2**, 16028 (2016).
45. E. E. Santiso, B. L. Trout, A general set of order parameters for molecular crystals. *J. Chem. Phys.* **134**, 064109 (2011).
46. A. M. Ganose, A. Jain, grapher: automated crystal structure text descriptions and analysisRobocrystallo. *MRS Commun.* **9**, 874–881 (2019).
47. K. M. Jablonka, D. Ongari, S. M. Moosavi, B. Smit, Using Collective Knowledge to Assign Oxidation States. *chemrXiv*, 11604129 (2020).
48. S. Zhang, Z. Hu, A. Subramonian, Y. Sun, Motif-driven contrastive learning of graph representations. arXiv:2012.12533 (2020).
49. D. Kingma, J. Ba, Adam: A Method for Stochastic Optimization. *Int. Conf. Learn. Represent*, (2014).
50. C. Dryzun, Continuous symmetry measures for complex symmetry group. *J. Comput. Chem.* **35**, 748–755 (2014).
51. M. Pinsky, D. Avnir, Continuous Symmetry Measures. 5. The Classical Polyhedra. *Inorg. Chem.* **37**, 5575–5582 (1998).
52. M. Pinsky, C. Dryzun, D. Casanova, P. Alemany, D. Avnir, Analytical methods for calculating Continuous Symmetry Measures and the Chirality Measure. *J. Comput. Chem.* **29**, 2712–2721 (2008).
53. H. Abdi, L. J. Williams, Principal Component Analysis. *WIREs Comput. Stat.* **2**, 433–459 (2010).
54. G. H. Golub, C. Reinsch, Singular value decomposition and least squares solutions. *Numer. Math.* **14**, 403–420 (1970).
55. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python. arXiv:1201.0490 (2018).

Acknowledgments

Funding: H.R.B., W.G., and Q.Y. acknowledge support from the U.S. Department of Energy under award number DE-SC0020310. F.R. acknowledges financial support from F.R.S-FNRS. This project benefitted from the supercomputing resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science user facility operated under contract no. DE-AC02-05CH11231. **Author contributions:** Q.Y. and S.V. conceived and coordinated the research project. H.R.B. and S.H. contributed equally to this

work. H.R.B. and Q.Y. collected the motif data and designed the motif-learning framework. S.H., S.Z., and S.V. performed the training and test of motif-based GNN models. F.R. and G.H. were involved in the automatic generation of structure motif information. W.G. performed the training and test of the atom-based GNN model. All authors participated in discussing the data and editing the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 8 October 2020
Accepted 2 March 2021
Published 21 April 2021
10.1126/sciadv.abf1754

Citation: H. R. Banjade, S. Hauri, S. Zhang, F. Ricci, W. Gong, G. Hautier, S. Vucetic, Q. Yan, Structure motif–centric learning framework for inorganic crystalline systems. *Sci. Adv.* **7**, eabf1754 (2021).

Science Advances

Structure motif–centric learning framework for inorganic crystalline systems

Huta R. Banjade, Sandro Hauri, Shanshan Zhang, Francesco Ricci, Weiyi Gong, Geoffroy Hautier, Slobodan Vucetic and Qimin Yan

Sci Adv **7** (17), eabf1754.
DOI: 10.1126/sciadv.abf1754