Hierarchical visualization of materials space with graph convolutional neural networks

Cite as: J. Chem. Phys. **149**, 174111 (2018); https://doi.org/10.1063/1.5047803
Submitted: 09 July 2018 . Accepted: 05 October 2018 . Published Online: 06 November 2018

Tian Xie, and Jeffrey C. Grossman

Hierarchical visualization of materials space with graph convolutional neural networks

Tian Xie and Jeffrey C. Grossman
Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

(Received 9 July 2018; accepted 5 October 2018; published online 6 November 2018)

The combination of high throughput computation and machine learning has led to a new paradigm in materials design by allowing for the direct screening of vast portions of structural, chemical, and property spaces. The use of these powerful techniques leads to the generation of enormous amounts of data, which in turn calls for new techniques to efficiently explore and visualize the materials space to help identify underlying patterns. In this work, we develop a unified framework to hierarchically visualize the compositional and structural similarities between materials in an arbitrary material space with representations learned from different layers of graph convolutional neural networks. We demonstrate the potential for such a visualization approach by showing that patterns emerge automatically that reflect similarities at different scales in three representative classes of materials: perovskites, elemental boron, and general inorganic crystals, covering material spaces of different compositions, structures, and both. For perovskites, elemental similarities are learned that reflects multiple aspects of atom properties. For elemental boron, structural motifs emerge automatically showing characteristic boron local environments. For inorganic crystals, the similarity and stability of local coordination environments are shown combining different center and neighbor atoms. The method could help transition to a data-centered exploration of materials space in automated materials design. *Published by AIP Publishing.* https://doi.org/10.1063/1.5047803

I. INTRODUCTION

Efficient exploration of the materials space has been central to material discovery as a result of the limited experimental and computational resources compared with its vast size. Often compositional or structural patterns are sought from past experiences that might guide the design of new materials, improving the efficiency of material exploration.[1–5] Emerging high-throughput computation and machine learning techniques directly screen large amounts of candidate materials for specific applications,[6–13] which enables fast and direct exploration of the materials space. However, the large quantities of material data generated makes the discovery of patterns challenging with traditional, human-centered approaches. Instead, an automated, data-centered method to visualize and understand a given materials design phase space is needed in order to improve the efficiency of exploration.

The key in visualizing material space is to map materials with different compositions and structures into a lower dimensional manifold where the similarity between materials can be measured by their Euclidean distances. One major challenge in finding such manifolds is to develop a unified representation for different materials. A widely used method is representing materials with feature vectors, where a set of descriptors are selected to represent each material.[14–16] There are also methods that automatically select descriptors that are best for predicting a desired target property.[17] Recent work has also developed atomic-scale representations to map complex atom configurations into low dimensional manifolds, such as atom centered symmetry functions,[18] social permutation invariant (SPRINT) coordinates,[19] global minimum of root-mean-square distance,[20] smooth overlap of atomic positions (SOAP),[21] among many other methods.[22–24] These representations often have physically meaningful parameters that can highlight some structural or chemical features. Material descriptors and atomic representations are also used together to combine compositional and structural information.[23,25] They have been used to visualize the material and molecular similarities,[26–28] as well as explore the complex configurational space of biological systems[29–32] and water structures.[33,34] In addition to Euclidean distances, similarity kernels are also used to compare material similarities.[27,28] Combined with machine learning algorithms, these representations have been used to predict material properties[12,14–17,22,35,36] as well as construct force fields.[21,37,38]

In parallel to these efforts, the success of "deep learning" has inspired a group of representations purely based on neural networks. Instead of designing descriptors or atomic representations that are fixed or contain several physically meaningful parameters, these approaches use relatively general neural network architectures with a large number of trainable weights to learn a representation directly. This field started with building neural networks on molecular graphs[39–42] and was recently expanded to periodic material systems by us[43] and Schütt *et al.*[44] Deep neural networks have shown many advantages over conventional machine learning methods in computer vision and speech recognition with the large amounts of data available,[45] and they outperformed conventional

methods on 11/17 datasets for predicting molecular properties in a recent study.[46] However, the general neural network architecture may also limit performance when the data size is small since there is no material specific information built-in. It is worth noting that many machine learning force fields combine atomic representations and neural networks,[18,37,47] but they usually deal with different compositions separately and use a significantly smaller number of weights. It has been shown that the hidden layers of these neural networks can learn physically meaningful representations by proper design of the network architecture. For instance, several studies have investigated the ideas of learning atom energies[42,43,48] and elemental similarities.[49,50] In addition, recent work showed that elemental similarities can also be learned using a specially designed SOAP kernel.[51]

In this work, we aim to develop a unified framework to hierarchically visualize the compositional and structural similarities between materials in an arbitrary material space with representations learned from different layers of the neural networks. The network is based on a variant of our previously developed crystal graph convolutional neural network (CGCNN) framework,[43] but it is designed to focus on presenting the similarities between materials at different scales, including elemental similarities, local environment similarities, and local energies. We apply this approach to visualize three material spaces: perovskites, elemental boron, and general inorganic crystals, covering material spaces of different compositions, different structures, and both, respectively. We show that in all three cases, a pattern emerges automatically that might aid in the design of new materials.

II. METHODS

To visualize the crystal space at different scales, we design a variant of CGCNN[43] that has meaningful interpretation at different layers of the neural network. The learned CGCNN network provides a vector representation of the local environments in each crystal that only depends on its composition and structure without any human designed features, enabling us to explore the materials space hierarchically.

As shown in Fig. 1, we first represent the crystal structure with a multigraph $\mathcal{G}$ that encodes the connectivity of atoms in the crystal. Each atom is represented by a node i in $\mathcal{G}$ which stores a vector $\boldsymbol{v}_i$ corresponding to the element type of the atom. To avoid introducing any human bias, we initialize $\boldsymbol{v}_i$ to be a random 64 dimensional vector for each element and allow it to evolve during the training process. Then, we search for the 12 nearest neighbors for each atom and introduce an edge

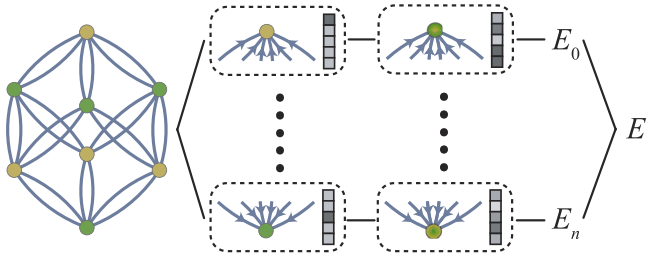

FIG. 1. The structure of the crystal graph convolutional neural networks.

$(i,j)_k$ between the center node i and neighbor j. The subscript k indicates that there can be multiple edges between the same end nodes as a result of the periodicity of the crystal. The edge $(i,j)_k$ stores a vector $\boldsymbol{u}_{(i,j)_k}$ whose tth element depends on the distance between i and j by

$$\boldsymbol{u}_{(i,j)_k}[t] = \exp(-(d_{(i,j)_k} - \mu_t)^2/\sigma^2), \tag{1}$$

where $\mu_t = t \cdot 0.2$ Å for $t = 0, 1, \ldots, 40$, $\sigma = 0.2$ Å, and $d_{(i,j)_k}$ denotes the distance between i and j at their kth edge.

In graph $\mathcal{G}$, each atom i is initialized by a vector $\boldsymbol{v}_i$ whose value solely depends on the element type of atom i. We call this iteration 0 where

$$\boldsymbol{v}_i^{(0)} = \boldsymbol{v}_i \tag{2}$$

Then, we perform convolution operations on the multigraph $\mathcal{G}$ with the convolution function designed in Ref. 43 which allows atom i to interact with its neighbors iteratively. In iteration t, we first concatenate neighbor vectors $\boldsymbol{z}_{(i,j)_k}^{(t-1)} = \boldsymbol{v}_i^{(t-1)} \oplus \boldsymbol{v}_j^{(t-1)} \oplus \boldsymbol{u}_{(i,j)_k}$ and then perform the convolution by

$$\boldsymbol{v}_i^{(t)} = \boldsymbol{v}_i^{(t-1)} + \sum_{j,k}\Big[\sigma(\boldsymbol{z}_{(i,j)_k}^{(t-1)}\boldsymbol{W}_f^{(t-1)} + \boldsymbol{b}_f^{(t-1)}) \odot g(\boldsymbol{z}_{(i,j)_k}^{(t-1)}\boldsymbol{W}_s^{(t-1)} + \boldsymbol{b}_s^{(t-1)})\Big], \tag{3}$$

where $\odot$ denotes element-wise multiplication, σ denotes a sigmoid function, and g denotes any non-linear activation function, and $\boldsymbol{W}$ and $\boldsymbol{b}$ denote weights and biases in the neural network, respectively. During these convolution operations, $\boldsymbol{v}_i^{(t)}$ forms a series of representations of the local environments of atom i at different scales.

After K iterations, we perform a linear transformation to map $\boldsymbol{v}_i^{(K)}$ to a scalar E_i,

$$E_i = \boldsymbol{v}_i^{(K)}\boldsymbol{W}_l + b_l \tag{4}$$

and then use a normalized sum pooling to predict the averaged total energy per atom of the crystal,

$$E = \frac{1}{n}\sum_i E_i, \tag{5}$$

where n is the number of atoms in the crystal. This introduces a physically meaningful term E_i to represent the energy of the local chemical environment.

The model is trained by minimizing the squared error between predicted properties relative to the density functional theory (DFT) calculated properties using backpropagation and stochastic gradient descent.

In this CGCNN model, each vector represents the local environment of each atom at different scales. Here, we focus three vectors that have the most representative physical interpretations.

1. *Element representation* $\boldsymbol{v}_i^{(0)}$ that depends completely on the type of element that atom i is composed of since no convolution operation has been performed, describing the similarities between elements.
2. *Local environment representation* $\boldsymbol{v}_i^{(K)}$ that depends on atom i and its Kth order neighbors after K convolution operations, describing the similarities between local environments that combines the compositional and structural information.

3. *Local energy representation* E_i that describes the energy of atom i.

III. RESULTS AND DISCUSSIONS

To illustrate how this method can help visualize the compositional and the structural aspects of the crystal space, we apply it to three datasets that represent different material spaces: (1) a group of perovskite crystals that share the same structure type but have different compositions; (2) different configurations of elemental boron that share the same composition but have different structures; and (3) inorganic crystals from the Materials Project[52] that have both different compositions and different structures.

For each material space, we train the CGCNN model with 60% of the data to predict the energy per atom of the materials. 20% of the data are used to select hyperparameters of the model and the last 20% are reserved for testing. In Fig. 2, we show the learning curves for the three representative material spaces where a subset of training data is used to show how the number of training data affects the model prediction performance. As we will show below, the representations learned by predicting the energies automatically gain physical meanings and can be used to explore the materials spaces.

A. Perovskite: Compositional space

First, we explore the compositional space of perovskites by visualizing the *element representations*. Perovskite is a crystal structure type with the form of ABC_3 as shown in Fig. 3(a). The dataset[53,54] that we used includes 18 928 different perovskites where the elements A and B can be any nonradioactive metals and the element C can be one or several from O, N, S, and F. We trained our model to predict the energy above the hull with 15 000 training data, and after hyperparameter optimization on 1890 validation data, we achieve a prediction mean absolute error (MAE) of 0.042 eV/atom on the 2000 test data. The prediction performance is excellent and lower than several recent ML models such as those of Schmidt *et al.* (0.121 eV/atom)[49] and Xie *et al.* (0.099 eV/atom).[43] The learning curve in Fig. 2 shows a straight line in the log-log

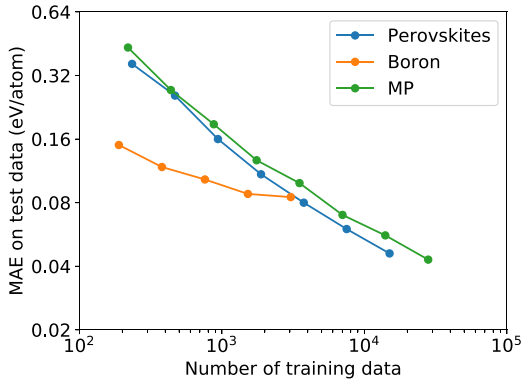

FIG. 2. Learning curves for the three representative material spaces. The mean absolute errors (MAEs) on test data is shown as a function of the number of training data for the perovskites,[53,54] elemental boron,[48] and Materials Project[52] datasets.

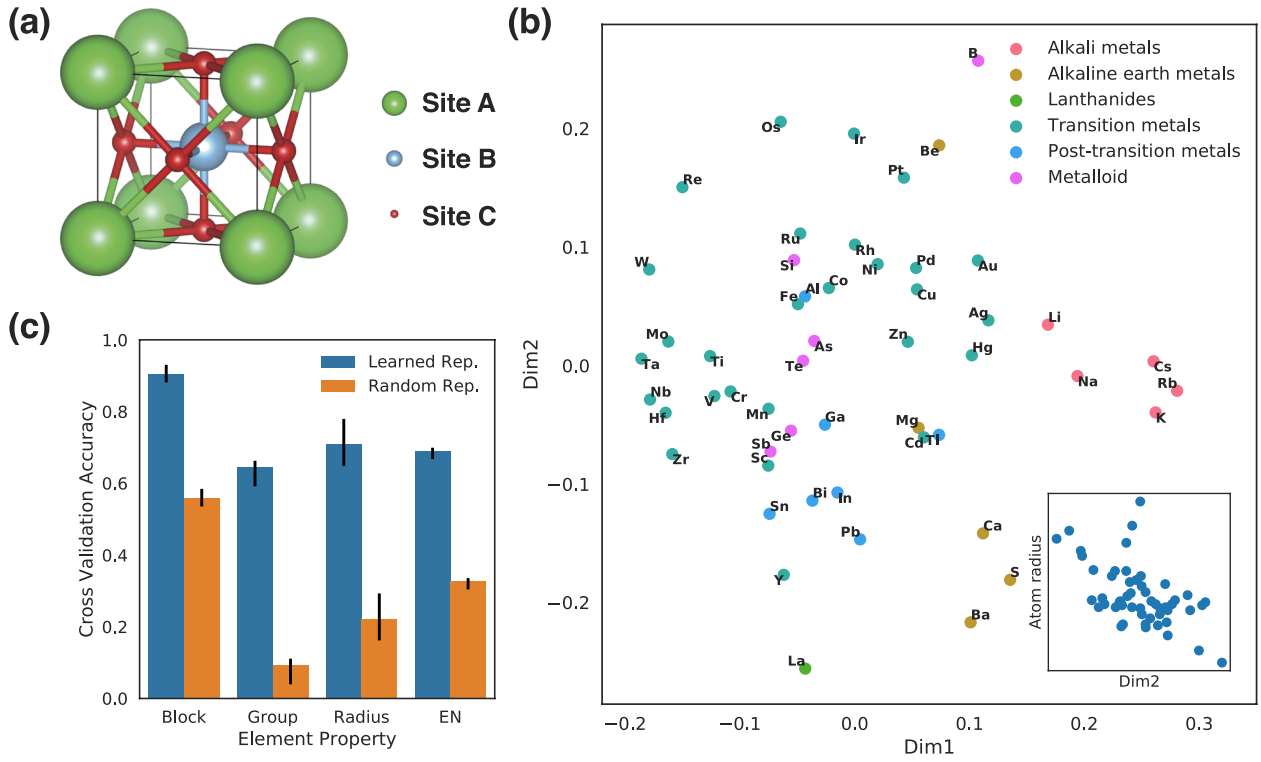

FIG. 3. Visualization of the element representations learned from the perovskite dataset. (a) The perovskite structure type. (b) Visualization of the two principal dimensions with principal component analysis. (c) Prediction performance of several atom properties, including the element block, group number, atom radius, and electronegativity, using a linear model on the element representations.

scale, indicating a steady increase of prediction performance as the number of training data increases.

In Figs. 3(b) and 3(c), the element representation $v_i^{(0)}$, a 64 dimensional vector, is visualized for every nonradioactive metal element after training with the perovskite dataset. Figure 3(b) shows the projection of these element representations on a 2D plane using principal component analysis, where elements are colored according to their elemental groups. We can clearly see that similar elements are grouped together based on their stability in perovskite structures. For instance, alkali metals are grouped on the right of the plot due to their similar properties. The large alkaline earth metals (Ba, Se, and Ca) are grouped on the bottom, distinct from Mg and Be, because their larger radii stabilize them in the perovskite structure. On the left side are elements such as W, Mo, and Ta that favor octahedral coordinations due to their configuration of *d* electrons, which might be related to their extra stability in the B site.[43] Interestingly, we can also observe a trend of decreasing atom radius from the bottom of the plot to the top, as shown in the insert of Fig. 3(b), except for the alkali metals as outliers. This indicates that CGCNN learns the atom radius as an important feature for perovskite stability. Recently, Schütt *et al.* also discovered a similar grouping of elements with data from the Materials Project.[44] In general, these visualizations can help discover similarities between elements for designing novel perovskite structures.

We also study how the element representations evolve as the number of training data changes. In Fig. S1, we show the 2D projections of the element representations when 234, 937, 3750, and 15 000 training data are used, respectively. The projection looks completely random with 234 training data, and some patterns start to emerge when 937 training data are used. In Fig. S1(b), transition metals are grouped on top of the figure, while large metals like La, Ca, Sr, Ba, and Cs are grouped at the bottom. With 3750 training data, the figure is already close to Fig. 3(b) and the relation between atom radius and the second dimension is clear. Figure 3(b) and Fig. S1(d) are almost identical after rotations because they both use 15 000 training data. Note that these representations start from different random initializations, but they result in similar patterns after training with the same perovskite data.

These 2D plots only account for part of the 64-dimensional element representation vectors. To fully understand how element properties are learned by CGCNN, we use linear logistic regression (LR) models to predict the block type, group number, radius, and electronegativity of each element from their learned representation vectors. In Fig. 3(c), we show the 3-fold cross validation accuracy of the LR models and compare them with LR models learned from random representations, which helps to rule out the possibility that the predictions are caused by coincidences. We discover a significantly higher prediction accuracy of the learned representations for all four properties, demonstrating that the element representations can reflect multiple aspects of element properties. For instance, the model predicts the block of the element with over 90% accuracy, and the same representation also predicts the group number, radius, and electronegativity with over 60% accuracy. This is surprising considering that these elements are from 16 different elemental

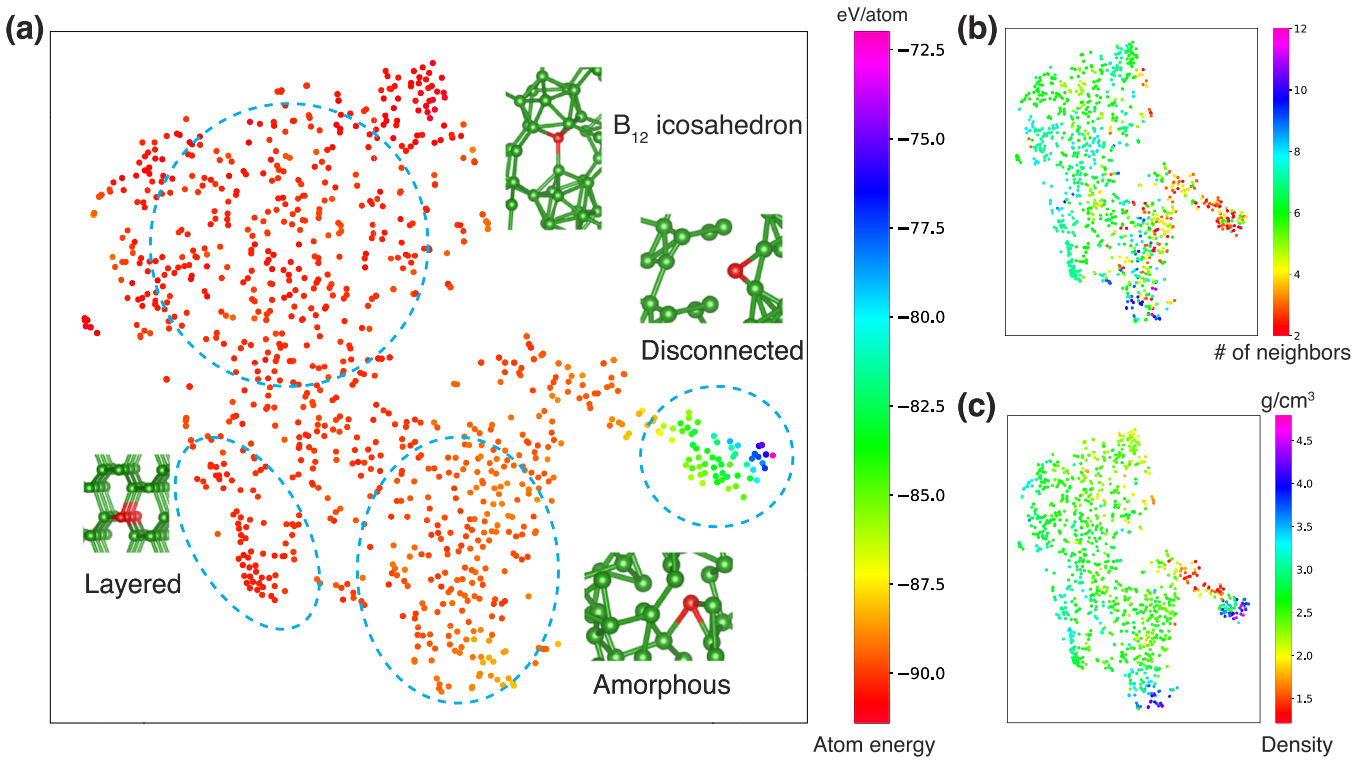

FIG. 4. Visualization of the local environment representations learned from the elemental boron dataset. The original 64D vectors are reduced to 2D with the t-distributed stochastic neighbor embedding algorithm. The color of each plot is coded with learned local energy (a), number of neighbors calculated by the Pymatgen package[55] (b), and density (c). Representative boron local environments are shown with the center atom colored in red.

groups. It is worth noting that these representations are learned only from the perovskite structures and the total energy above hull, but they are in agreement with these empirical element properties reflecting decades of human chemical intuition.

B. Elemental boron: Structural space

As a second example, we explore the structural space of elemental boron by visualizing the *local environment representations* and the corresponding *local energies*. Elemental boron has a number of complex crystal structures due to its unique, electron-deficient bonding nature.[48,56] We use a dataset that includes 5038 distinct elemental boron structures and their total energies calculated using density functional theory.[48] We train our CGCNN model with 3038 structures, and perform hyperparameter optimization with 1000 validation structures. The MAE of the predicted energy relative to DFT results from the remaining 1000 test structures is 0.085 eV/atom. The learning curve in Fig. 2 shows a much smaller slope compared with the other material spaces. One explanation is that there exist many highly unstable boron structures in the dataset, whose energies might be hard to predict given the limited structures covered by the training data.

In Fig. 4, 1000 randomly sampled boron local environment representations are visualized in 2 dimensions using the t-distributed stochastic neighbor embedding (t-SNE) algorithm.[57] We observe primarily four different regions of different boron local environments, and we discover a smooth transition of local energy, number of neighbor atoms, and the density between different regions. The disconnected region consists of boron atoms at the edges of boron clusters [Fig. 4 and Figs. S1(a)–S1(c)]. These atoms have very high local energies and lower number of neighbors, as to be expected, and their density varies depending on the distances between clusters. The amorphous region includes boron atoms in a relatively disordered local configuration, and their local energies are lower than the disconnected counterparts but higher than other configurations [Fig. 4 and Figs. S1(d)–S1(f)]. We can see that the number of neighbors fluctuates drastically in these two regions due to the relatively disordered local structures. The layered region is composed of boron atoms in layered boron planes, where neighbors on one side are closely bonded and the neighbors on the other side are further away [Fig. 4 and Figs. S1(g)–S1(i)]. The B_{12} icosahedron region includes boron local environments with the lowest local energy, which have a characteristic icosahedron structure [Fig. 4 and Figs. S1(j)–S1(l)]. The local environments in each region share common characteristics but are slightly different in detail. For instance, most boron atoms in the B_{12} icosahedron region are in a slightly distorted icosahedron, and the local environments in Fig. S1(l) only have certain features of an icosahedron. Note that these representations are rather localized. The global structure of Fig. S1(c) is layered, but the representation of the highlighted atom at the edge is closer to the disconnected region locally. Some experimentally observed boron structures, like boron fullerenes, are not present in the dataset. We calculate the local environment representations of every distinct boron atom of two boron fullerenes[58] using the trained CGCNN, and plot them into the original 2D visualization in Fig. S3. They form a small cluster close to the B_{12} icosahedron region, which can be explained by the fact that they share many common characteristics to the B_{12} icosahedron structure. In addition, the representations of the less symmetric $B_{40}(C_s)$ are more spread out than the more symmetric $B_{40}(D_{2d})$. Note that the pattern in Fig. S3 is slightly different from that in Fig. 4 due to the

TABLE I. Comparison of the prediction performance of formation energy per atom. The mean absolute errors (MAEs) on test data reported in several recent studies are summarized. Data are from several different but similar inorganic crystal material datasets. MP represents Materials Project,[52] OQMD represents the open quantum materials database,[61] and the ternary compounds are $A_xB_yC_z$ compounds calculated by Ref. 15.

Method	MAE (eV/atom)	Data source	Training size
This work	0.042	MP	28 046
CGCNN[43]	0.039	MP	28 046
SchNet[44]	0.035	MP	60 000
Generalized coulomb matrix[62]	0.37	MP	3 000
Decision trees + heuristic[15]	0.12	Ternary compounds	15 000
Voronoi + composition[25]	0.08	OQMD	30 000
QML[23]	~0.11	OQMD	2 000
Random subspace + REPTree[16]	0.088	OQMD	228 676

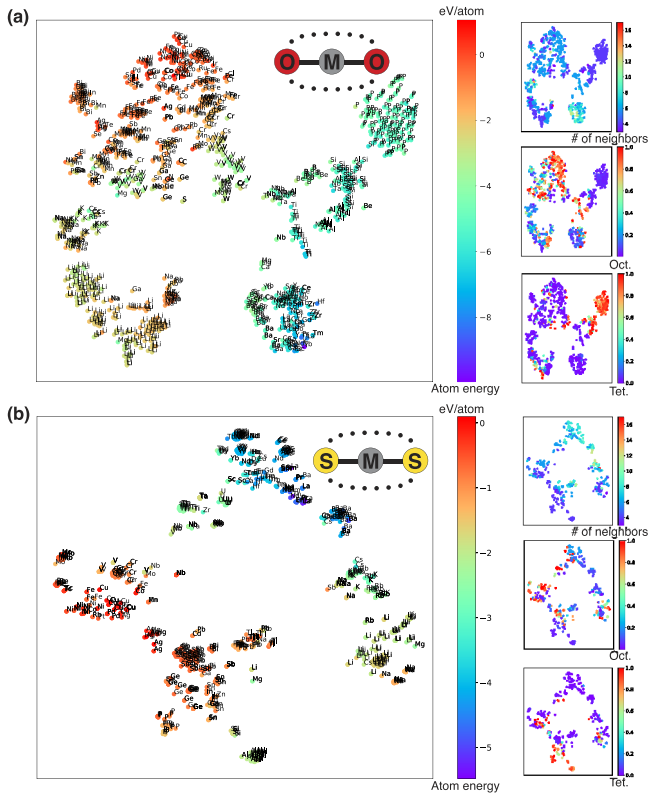

FIG. 5. Visualization of the local oxygen (a) and sulfur (b) coordination environments. The points are labeled according to the type of the center atoms in the coordination environments. The colors of the left parts are coded with learned local energies, and the color of the right parts are coded with number of neighbors,[55] octahedron order parameter, and tetrahedron order parameter.[59]

random nature of the t-SNE algorithm, but the overall structure of the patterns is preserved.

Taken together, such a visualization approach provides a convenient way to explore complex boron configurations, enabling the identification of characteristic structures and systematic exploration of structural space.

C. Materials Project: Compositional and structural space

As a third example of applying this approach, we explore the material space of crystals in the Materials Project dataset,[52] which includes both compositional and structural differences, by visualizing the *element representation*, *local environment representation*, and the *local energy representation*. The dataset includes 46 744 materials that cover the majority of crystals from the Inorganic Crystal Structure Database,[60] providing a good representation of known inorganic materials. After training with 28 046 crystals and performing hyperparameter optimization with 9348 crystals, our model achieves a MAE of predicted energy relative to DFT calculations on the 9348 test crystals of 0.042 eV/atom, slightly higher than the MAE of our previous work, 0.039 eV/atom, with a CGCNN structure focusing on prediction performance.[43] The learning curve in Fig. 2 is similar to that of the perovskites dataset, which might indicate a similar prediction performance to the datasets that are composed of stable inorganic compounds. In Table I, we compare the prediction performance of this method with several recently published studies.

In Fig. S2, the element representation of 89 elements learned from the dataset is shown using the same method as that used to generate Fig. 3(b). We observe a similar grouping of elements from the same elemental groups, but the overall pattern differs since it reflects the stability of each element in general inorganic crystals rather than perovskites. For instance, the non-metal and halogen elements stand out because their properties deviate from other metallic elements.

To illustrate how the compositional and structural spaces can be explored simultaneously, we visualize the oxygen and sulfur coordination environments in the Materials Project dataset using the local environment representation

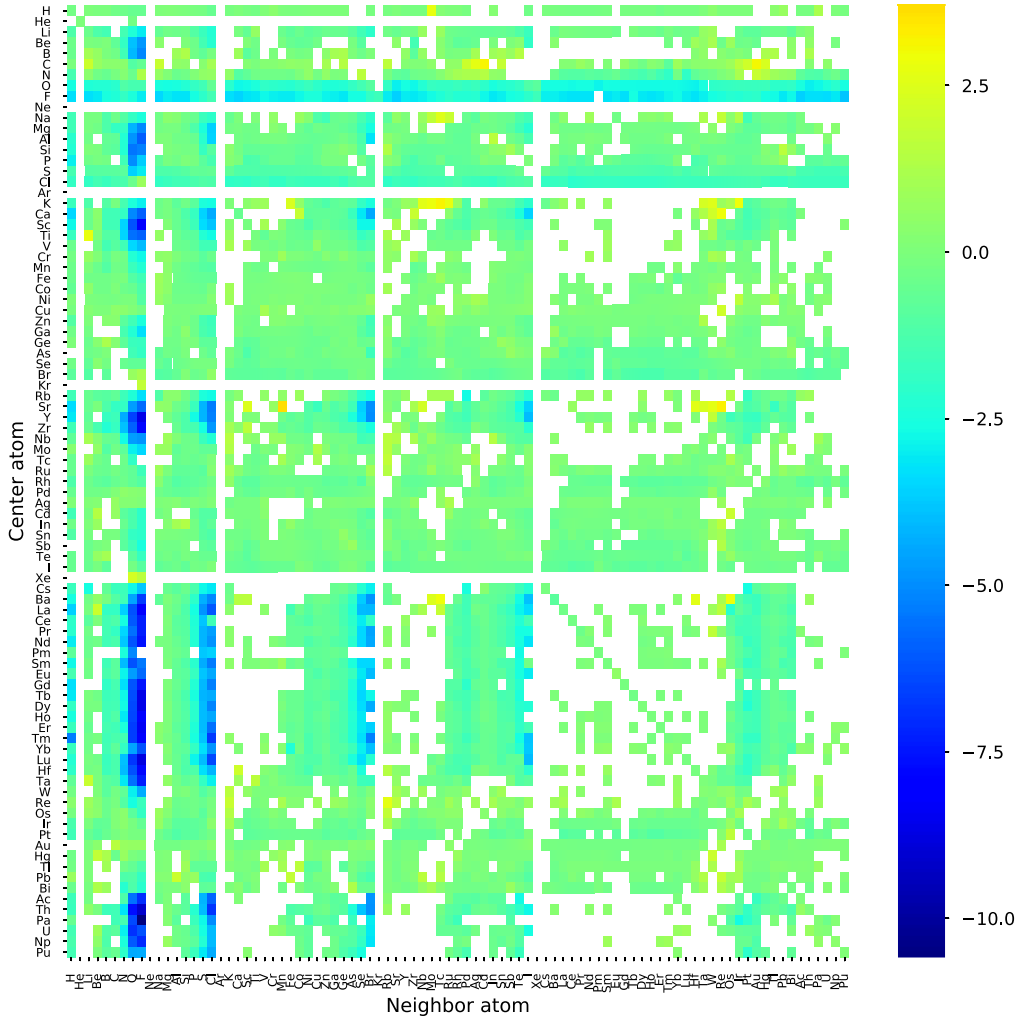

FIG. 6. The averaged local energy of 734 077 distinct coordination environments in the Materials Project dataset. The color is coded with the average of learned local energies while having the corresponding elements as the center atom and the first neighbor atom. White is used when no such coordination environment exists in the dataset.

and local energy. 1000 oxygen and 803 sulfur coordination environments are randomly selected and visualized using the t-SNE algorithm. As shown in Fig. 5(a), the oxygen coordination environments are clustered into 4 major groups. The upper right group has the center atom of non-metal elements like P, Al, Si, forming tetrahedron coordinations. The center atoms of the upper left environments are mostly transition metals, and they mostly form octahedron coordinations. The lower left group has center atoms of alkali metals, and the lower right group has those of alkaline earth metals and lanthanides which have larger radii and therefore higher coordination numbers. The sulfur coordination environment visualization [Fig. 5(b)] shares similar patterns due to the similarities between oxygen and sulfur, and a similar four-cluster structure can be observed. However, instead of non-metal elements, the lower center group has center atoms of metalloids like Ge, Sn, and Sb since these elements will be more stable in a sulfur with respect to an oxygen coordination environment.

The local energies of oxygen and sulfur coordination environments are determined by their relative stability to the pure elemental states since the model is trained using formation energy data, which treats the pure elemental states as the reference energy states. In Fig. S3, we show the change of oxygen and sulfur local energies as a function of atomic number. We can clearly see that it follows a similar trend as the electronegativity of the elements: elements with lower electronegativity tend to have lower local energy and vice versa. This is because elements with lower electronegativity tend to give the oxygen and sulfur more electrons and thus form stronger bonds. The local energies of alkali metals are slightly higher since they form weaker ionic bonds due to lower charges. Interestingly, the strong covalent bonds between oxygen and Al, Si, P, S result in a V-shaped curve in the figure, with Si–O environments having the lowest energy, which contrasts the trend of electronegativity and sulfur coordination environments, whose local energies are dominated by the strength of ionic bonds. We also observe a larger span of local energies in oxygen coordination environments than their sulfur counterparts due to the stronger ionic interactions.

Inspired by these results, we visualize the averaged local energy of 734 077 distinct coordination environments in the Materials Project by combining different center and neighbor atoms in Fig. 6. This figure illustrates the stability of the local coordination environment while combining the corresponding center and neighbor elements. The diagonal line represents coordination environments made up with the same elements with local energies close to zero, which corresponds to elemental substances with zero formation energy. The coordination environments with lowest local energy consist of high valence metals and high electronegativity non-metals, which can be explained by the large cohesive energies due to strong ionic bonds. One abnormality is the stable Al–O, Si–O, P–O, S–O coordination environments, although this can be attributed to their strong covalent bonds. We can also see that Tm-H coordination stands out as a stable hydrogen solid solution.[63] It is worth noting that each local energy in Fig. 6 is the average of many coordination environments with different shapes and outer layer chemistries, and we can obtain more information by using additional visualizations similar to Fig. 5.

IV. CONCLUSION

In summary, we developed a unified approach to visualize the compositional and structural space of materials. The method provides hierarchical representations of the local environments at different scales, which enables a general framework to explore different material systems and measure material similarities. The insights gained from the visualizations could help to discover patterns from a large pool of candidate materials that may be impossible by human analysis, and provide guidance to the design of new materials. In addition to energies, this method can potentially be applied to other material properties for the exploration of novel functional materials.

SUPPLEMENTARY MATERIAL

See supplementary material for the details of the hyperparameters for each model, the results of the effects of the number of training data on element representations, additional figures showing the structures of boron local environments and the location of boron fullerene local environment representations with respect to the representations of other boron structures, the results of the element representations learned from the Materials Project dataset, and the results of the change of local energy as a function of the atomic number.

ACKNOWLEDGMENTS

This work was supported by the Toyota Research Institute. Computational support was provided through the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, and the Extreme Science and Engineering Discovery Environment, supported by the National Science Foundation Grant No. ACI-1053575.

[1]G. Niu, X. Guo, and L. Wang, "Review of recent progress in chemical stability of perovskite solar cells," J. Mater. Chem. A **3**, 8970–8980 (2015).
[2]H. J. Snaith, "Perovskites: The emergence of a new era for low-cost, high-efficiency solar cells," J. Phys. Chem. Lett. **4**, 3623–3630 (2013).
[3]M. Xu, T. Liang, M. Shi, and H. Chen, "Graphene-like two-dimensional materials," Chem. Rev. **113**, 3766–3798 (2013).
[4]S. Z. Butler, S. M. Hollen, L. Cao, Y. Cui, J. A. Gupta, H. R. Gutiérrez, T. F. Heinz, S. S. Hong, J. Huang, A. F. Ismach *et al.*, "Progress, challenges, and opportunities in two-dimensional materials beyond graphene," ACS Nano **7**, 2898–2926 (2013).
[5]O. Madelung, *Physics of III-V Compounds* (J. Wiley, 1964).
[6]J. Greeley, T. F. Jaramillo, J. Bonde, I. Chorkendorff, and J. K. Nørskov, "Computational high-throughput screening of electrocatalytic materials for hydrogen evolution," Nat. Mater. **5**, 909 (2006).
[7]S. M. Senkan, "High-throughput screening of solid-state catalyst libraries," Nature **394**, 350 (1998).
[8]R. Potyrailo, K. Rajan, K. Stoewe, I. Takeuchi, B. Chisholm, and H. Lam, "Combinatorial and high-throughput screening of materials libraries: Review of state of the art," ACS Comb. Sci. **13**, 579–633 (2011).

[9]S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, "The high-throughput highway to computational materials design," Nat. Mater. **12**, 191 (2013).
[10]G. Hautier, C. C. Fischer, A. Jain, T. Mueller, and G. Ceder, "Finding natures missing ternary oxide compounds using machine learning and density functional theory," Chem. Mater. **22**, 3762–3767 (2010).
[11]R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha, T. Wu *et al.*, "Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach," Nat. Mater. **15**, 1120 (2016).
[12]F. A. Faber, A. Lindmaa, O. A. Von Lilienfeld, and R. Armiento, "Machine learning energies of 2 million elpasolite (ABC_2D_6) crystals," Phys. Rev. Lett. **117**, 135502 (2016).
[13]M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," Phys. Rev. Lett. **108**, 058301 (2012).
[14]G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad, "Accelerating materials property predictions using machine learning," Sci. Rep. **3**, 2810 (2013).
[15]B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, "Combinatorial screening for new materials in unconstrained composition space with machine learning," Phys. Rev. B **89**, 094104 (2014).
[16]L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, "A general-purpose machine learning framework for predicting properties of inorganic materials," npj Comput. Mater. **2**, 16028 (2016).
[17]L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, "Big data of materials science: Critical role of the descriptor," Phys. Rev. Lett. **114**, 105503 (2015).
[18]J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials," J. Chem. Phys. **134**, 074106 (2011).
[19]F. Pietrucci and W. Andreoni, "Graph theory meets *ab initio* molecular dynamics: Atomic structures and transformations at the nanoscale," Phys. Rev. Lett. **107**, 085504 (2011).
[20]A. Sadeghi, S. A. Ghasemi, B. Schaefer, S. Mohr, M. A. Lill, and S. Goedecker, "Metrics for measuring distances in configuration spaces," J. Chem. Phys. **139**, 184118 (2013).
[21]A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," Phys. Rev. B **87**, 184115 (2013).
[22]K. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. Müller, and E. Gross, "How to represent crystal structures for machine learning: Towards fast prediction of electronic properties," Phys. Rev. B **89**, 205118 (2014).
[23]F. A. Faber, A. S. Christensen, B. Huang, and O. A. von Lilienfeld, "Alchemical and structural distribution based representation for universal quantum machine learning," J. Chem. Phys. **148**, 241717 (2018).
[24]A. Glielmo, C. Zeni, and A. De Vita, "Efficient nonparametric n-body force fields from machine learning," Phys. Rev. B **97**, 184307 (2018).
[25]L. Ward, R. Liu, A. Krishna, V. I. Hegde, A. Agrawal, A. Choudhary, and C. Wolverton, "Including crystal structure attributes in machine learning models of formation energies via voronoi tessellations," Phys. Rev. B **96**, 024104 (2017).
[26]O. Isayev, D. Fourches, E. N. Muratov, C. Oses, K. Rasch, A. Tropsha, and S. Curtarolo, "Materials cartography: Representing and mining materials space using structural and electronic fingerprints," Chem. Mater. **27**, 735–743 (2015).
[27]S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, "Comparing molecules and solids across structural and alchemical space," Phys. Chem. Chem. Phys. **18**, 13754–13769 (2016).
[28]F. Musil, S. De, J. Yang, J. E. Campbell, G. M. Day, and M. Ceriotti, "Machine learning for the structure–energy–property landscapes of molecular crystals," Chem. Sci. **9**, 1289–1300 (2018).
[29]P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, "Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction," Proc. Natl. Acad. Sci. U. S. A. **103**, 9885–9890 (2006).
[30]M. Ceriotti, G. A. Tribello, and M. Parrinello, "Simplifying the representation of complex free-energy landscapes using sketch-map," Proc. Natl. Acad. Sci. U. S. A. **108**, 13023–13028 (2011).
[31]V. Spiwok and B. Králová, "Metadynamics in the conformational space nonlinearly dimensionally reduced by isomap," J. Chem. Phys. **135**, 224504 (2011).
[32]M. A. Rohrdanz, W. Zheng, and C. Clementi, "Discovering mountain passes via torchlight: Methods for the definition of reaction coordinates and pathways in complex macromolecular reactions," Annu. Rev. Phys. Chem. **64**, 295–316 (2013).
[33]F. Pietrucci and R. Martoňák, "Systematic comparison of crystalline and amorphous phases: Charting the landscape of water structures and transformations," J. Chem. Phys. **142**, 104704 (2015).
[34]E. A. Engel, A. Anelli, M. Ceriotti, C. J. Pickard, and R. J. Needs, "Mapping uncharted territory in ice from zeolite networks to ice structures," Nat. Commun. **9**, 2173 (2018).
[35]A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput, and I. Tanaka, "Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and bayesian optimization," Phys. Rev. Lett. **115**, 205901 (2015).
[36]O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, and A. Tropsha, "Universal fragment descriptors for predicting properties of inorganic crystals," Nat. Commun. **8**, 15679 (2017).
[37]J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," Phys. Rev. Lett. **98**, 146401 (2007).
[38]V. Botu, R. Batra, J. Chapman, and R. Ramprasad, "Machine learning force fields: Construction, validation, and outlook," J. Phys. Chem. C **121**, 511–522 (2016).
[39]D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in Neural Information Processing Systems* (Neural Information Processing Systems Foundation, 2015), pp. 2224–2232.
[40]S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley, "Molecular graph convolutions: Moving beyond fingerprints," J. Comput.-Aided Mol. Des. **30**, 595–608 (2016).
[41]J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," Proc. Mach. Learn. Res. **70**, 1263–1272 (2017), available at http://proceedings.mlr.press/v70/gilmer17a.html.
[42]K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks," Nat. Commun. **8**, 13890 (2017).
[43]T. Xie and J. C. Grossman, "Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties," Phys. Rev. Lett. **120**, 145301 (2018).
[44]K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "SchNet—A deep learning architecture for molecules and materials," J. Chem. Phys. **148**, 241722 (2018).
[45]I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, 2016), http://www.deeplearningbook.org.
[46]Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "MoleculeNet: A benchmark for molecular machine learning," Chem. Sci. **9**, 513–530 (2018).
[47]T. B. Blank, S. D. Brown, A. W. Calhoun, and D. J. Doren, "Neural network models of potential energy surfaces," J. Chem. Phys. **103**, 4129–4137 (1995).
[48]V. L. Deringer, C. J. Pickard, and G. Csányi, "Data-driven learning of total and local energies in elemental boron," Phys. Rev. Lett. **120**, 156001 (2018).
[49]J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, and M. A. Marques, "Predicting the thermodynamic stability of solids combining density functional theory and machine learning," Chem. Mater. **29**, 5090–5103 (2017).
[50]Q. Zhou, P. Tang, S. Liu, J. Pan, Q. Yan, and S.-C. Zhang, "Learning atoms for materials discovery," Proc. Natl. Acad. Sci. U. S. A. **115**, E6411 (2018).
[51]M. J. Willatt, F. Musil, and M. Ceriotti, "A data-driven construction of the periodic table of the elements," preprint arXiv:1807.00236 (2018).
[52]A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. a. Persson, "The materials project: A materials genome approach to accelerating materials innovation," APL Mater. **1**, 011002 (2013).
[53]I. E. Castelli, D. D. Landis, K. S. Thygesen, S. Dahl, I. Chorkendorff, T. F. Jaramillo, and K. W. Jacobsen, "New cubic perovskites for one-and two-photon water splitting using the computational materials repository," Energy Environ. Sci. **5**, 9034–9043 (2012).
[54]I. E. Castelli, T. Olsen, S. Datta, D. D. Landis, S. Dahl, K. S. Thygesen, and K. W. Jacobsen, "Computational screening of perovskite metal oxides for optimal solar light capture," Energy Environ. Sci. **5**, 5814–5819 (2012).
[55]S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, "Python materials

genomics (pymatgen): A robust, open-source python library for materials analysis," Comput. Mater. Sci. **68**, 314–319 (2013).
[56]T. Ogitsu, E. Schwegler, and G. Galli, "β-Rhombohedral boron: At the crossroads of the chemistry of boron and the physics of frustration," Chem. Rev. **113**, 3425–3449 (2013).
[57]L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," J. Mach. Learn. Res. **9**, 2579–2605 (2008).
[58]H.-J. Zhai, Y.-F. Zhao, W.-L. Li, Q. Chen, H. Bai, H.-S. Hu, Z. A. Piazza, W.-J. Tian, H.-G. Lu, Y.-B. Wu *et al.*, "Observation of an all-boron fullerene," Nat. Chem. **6**, 727 (2014).
[59]N. E. R. Zimmermann, M. K. Horton, A. Jain, and M. Haranczyk, "Assessing local structure motifs using order parameters for motif recognition, interstitial identification, and diffusion path characterization," Front. Mater. **4**, 34 (2017).
[60]M. Hellenbrandt, "The inorganic crystal structure database (ICSD): Present and future," Crystallogr. Rev. **10**, 17–22 (2004).
[61]J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, "Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD)," Jom **65**, 1501–1509 (2013).
[62]F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, "Crystal structure representations for machine learning models of formation energies," Int. J. Quantum Chem. **115**, 1094–1101 (2015).
[63]J. Bonnet and J. Daou, "Study of the hydrogen solid solution in thulium," J. Phys. Chem. Solids **40**, 421–425 (1979).