Atomic-position independent descriptor for machine learning of material properties

Ankit Jain*
Department of Chemical Engineering, Stanford University, Stanford, California 94305, USA

Thomas Bligaard†
SUNCAT Center for Interface Science and Catalysis, SLAC, National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, California 94025, USA

(Received 17 September 2018; revised manuscript received 19 November 2018; published 20 December 2018)

The high-throughput screening of periodic inorganic solids using machine learning methods requires atomic positions to encode structural and compositional details into appropriate material descriptors. These atomic positions are not available *a priori* for new materials, which severely limits exploration of novel materials. We overcome this limitation by using only crystallographic symmetry information in the structural description of materials. We show that for materials with identical structural symmetry, machine learning is trivial, and accuracies similar to that of density functional theory calculations can be achieved by using only atomic numbers in the material description. For machine learning of formation energies of bulk crystalline solids, this simple material descriptor is able to achieve prediction mean absolute errors of only 0.07 eV/at on a test dataset consisting of more than 85 000 diverse materials. This atomic-position independent material descriptor presents a new route of materials discovery wherein millions of materials can be screened by training a machine learning model over a drastically reduced subspace of materials.

DOI: 10.1103/PhysRevB.98.214112

I. INTRODUCTION

Machine learning (ML) methods are now able to predict material properties with similar accuracy to those of density functional theory (DFT) calculations but with several orders of magnitude reduced computational cost [1–6]. ML methods require a computational representation of materials that is achieved by mapping structure and composition to descriptors [7,8]. The majority of descriptors used in ML of inorganic periodic solids, such as the extended Coulomb matrix [7] and the partial radial distribution function [9], require atomic positions. This explicit use of atomic coordinates depends on prior knowledge of accurate positions for all atoms in a material with limited applicability to new and/or unknown materials (where atomic positions are not known *a priori*, as in the case of high-throughput calculations). Recently, there have been attempts to reduce this dependence of structure description on atomic coordinates, for instance, with a lattice volume independent Voronoi-tessellation (VT) -based descriptor [10], volume-scaling-based normalized prototype structures [11], and a perturbation-tolerant crystal graph (CG) -based convolution neural network (CNN) [12], but a true atomic-position independent (apI) descriptor capable of describing structure details without explicit use of atomic positions is still nonexistent.

Here, we present a crystallographic method for the apI description of structural details for inorganic periodic solids. Using this method, materials are clustered into different structure types on the basis of symmetry. Within a given structure-type cluster, only material composition is needed to accurately describe the system. We first show that we can use an apI descriptor effectively within a structure type, and we then develop a universal apI descriptor (U-apI) that extends across structure types.

For the prediction of material formation energies (FE) relative to their standard elemental references, accuracies similar to that of DFT calculations are achieved within these structure-type clusters by using a representation learning feed-forward neural network (RLFNN) employing only atomic numbers of participating species. For FE predictions on the 20 most frequently appearing structure types from the open quantum materials database (OQMD) [6], an attention-based convolution neural network (ABCNN) employing a U-apI descriptor is able to match apD descriptor-based ML models by delivering a regularized performance with an average test mean absolute error (MAE) of 0.07 eV/at.

The U-apI descriptor-based ABCNN is the first-ever apI descriptor-based material ML model, which is capable of learning across different structure types while matching the performance of more involved apD descriptor-based ML models. The U-apI descriptor-based ABCNN presents a new paradigm of high-throughput material discovery where millions of compounds can be reliably screened using ML without relying on atomic coordinates.

II. STRUCTURE TYPES

Three-dimensional periodic solids can be classified into one of the 14 Bravais lattices [13]. These 14 Bravais lattices, when combined with 32 point groups, result in 230

*ankitj@alumni.cmu.edu
†bligaard@slac.stanford.edu

three-dimensional space groups (without considering spin) [14,15]. Each of these 230 space groups is divided into conjugate subgroups called Wyckoff sites [16]. Atoms in crystals can sit only on special positions, called Wyckoff positions, which are points belonging to the Wyckoff site of the space group of the crystal [17]. For some Wyckoff sites, all internal fractional coordinates are fixed; at the other sites, one or more fractional coordinate can be changed while maintaining the symmetry of the Wyckoff site.

The atomic structure of a periodic solid is uniquely identified by specifying (i) the crystal space group, (ii) chemical species, (iii) Wyckoff sites occupied by atoms, (iv) the unit-cell parameters, and (v) numerical values of the fractional coordinates for occupied Wyckoff sites (needed for sites with free fractional coordinates). Fixing the space group and Wyckoff site of a material uniquely set the symmetry environment of the atoms. Herein, this will be referred to as structure type. Specifying the chemical species that occupy specific Wyckoff sites results in a completely specified structure. Within this fixed chemical and symmetry environment, the numerical values of the unit-cell parameters and free fractional coordinates can be determined by performing DFT calculations or by experimental synthesis and characterization of material.

For a given space group, many different Wyckoff site combinations, referred to as Wyckoff sets here, result in structurally identical compounds. For instance, for ABC_3 stoichiometry in space group 221, Wyckoff sets [(A, a), (B, b), (C, c)] and [(A, b), (B, a), (C, d)] represent the same perovskite structure type ([A, a) denotes specie A occupying Wyckoff site a]. The former can be transformed to the latter by a translation of $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$. This poses a difficulty in the unique representation of structures.

We overcome this difficulty by grouping all symmetrically equivalent Wyckoff sets to each other. For a Wyckoff set W consisting of N Wyckoff sites w_i, $i \in [1, N]$ and Wyckoff positions $s_{ij,\alpha}$ $j \in [1, n_i]$, where n_i and α represent the multiplicity of Wyckoff site w_i and lattice directions $(\vec{a}_1, \vec{a}_2, \vec{a}_3)$, we find a symmetric Wyckoff set W' such that (i) W' consists of N Wyckoff sites, (ii) the multiplicities of Wyckoff sites in set W' are the same as those of corresponding Wyckoff sites in set W, i.e., $n_i = n_i{}'$, for $i \in [1, N]$, and (iii) $\exists\, \vec{t}_i$ such that

$$f(\overline{\overline{R}}(\vec{s_{ij}} + \vec{T}), \vec{t}_i) = \vec{s_{ij}'}, \tag{1}$$

where $\overline{\overline{R}}$ and $\vec{T}$ represent a rotation matrix and a translation vector belonging to symmetry operations of a given lattice, and f transforms free parameters (x, y, z) in Wyckoff site positions $\vec{s_{ij}}$ to $(x - t_{i,x}, y - t_{i,y}, z - t_{i,z})$.

III. ML WITHIN STRUCTURE TYPE

We begin our apI-descriptor design with a simple composition-only descriptor consisting of 15 elemental properties of participating species [18]. The ML is performed using the random forest (RF) ML algorithm with 10 decision trees as implemented in the ML library SKlearn [19] (prediction MAE changes by less than 0.01 eV/at upon increasing the number of decision trees from 10 to 100 in the random forest). The ML is performed on 12 318 ABC_3 stoichiometry materials from the OQMD database for the prediction of

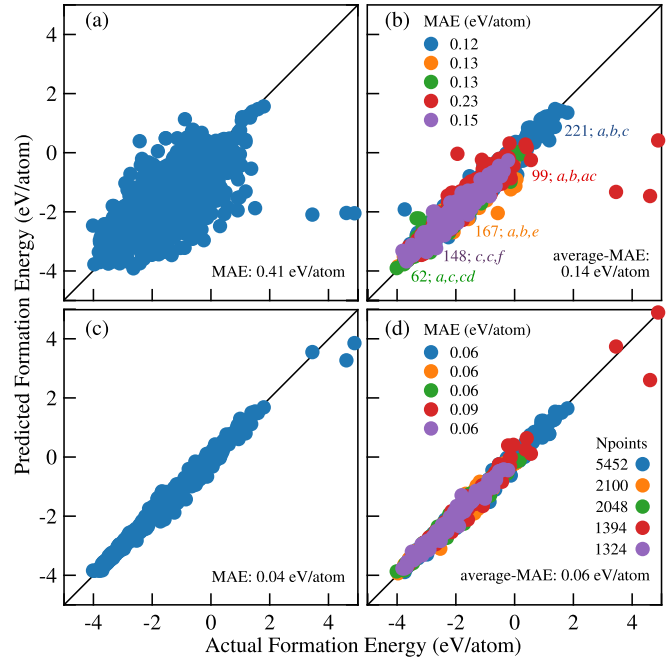

FIG. 1. Role of structural information in the ML of materials. The predicted FEs of 12 318 ABC_3 stoichiometry materials from OQMD. In (a) and (b), materials are described using a simple composition-only descriptor consisting of 15 elemental properties of species A, B, and C by training the ML model over all structure types simultaneously in (a) and separately in (b). In (c) and (d), predictions are made using a VT-based apD descriptor [10] by training the ML model over all structure types simultaneously in (c) and separately in (d). The labels next to the data points in (b) show the structure type (space group; Wyckoff sites). All predictions are made using a random forest ML algorithm. Only test data are shown (20% of data).

FEs of materials for decomposition into elemental standard states [10]. The materials are randomly assigned into training and test datasets consisting of 80% and 20% of the total materials, respectively. Figure 1 reports the predicted FEs of the materials in the test dataset materials. A MAE of 0.41 eV/at is achieved by using a simple composition-only descriptor; see Fig. 1(a). The large MAE is due to the presence of five different ABC_3 allotropes in the dataset, indistinguishable by elemental properties alone. To distinguish these allotropes, materials are preclustered on the basis of structure type before performing ML [Fig. 1(b)] reducing the average MAE to 0.14 eV/at. This threefold reduction in the MAE with structure-type clustering demonstrates the importance of retaining structural details. This is in contrast with [20], where the authors suggested a similar ML performance from structure-dependent and structure-independent descriptors for vibrational free energy and entropy predictions.

In Fig. 1(c), the VT-based apD descriptor [10] is used for the material description, and the prediction MAE is 0.04 eV/at. In comparison to the structure-type preclustering-based apI descriptor in Fig. 1(b), the superior performance of the VT-based apD descriptor in Fig. 1(c) can originate from (i) simultaneous training over all structure types, (ii) additional apD details about fractional volumes of atoms, and (iii) inclusion of 271 elemental/structural properties of

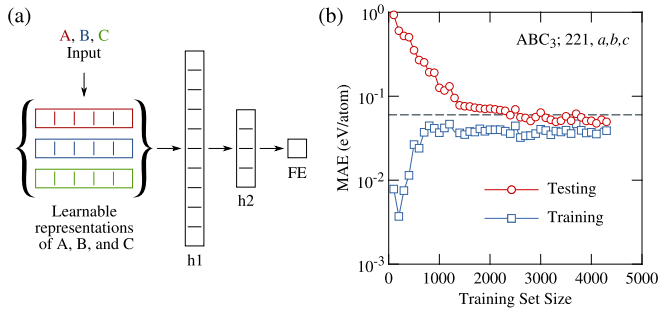

FIG. 2. ML for the same structure-type materials. (a) A representation-learning feedforward neural network employing atomic numbers of participating species for FE prediction of perovskite structure-type materials. (b) The effect of training set size on the prediction performance of the neural network. The network takes atomic numbers of species A, B, and C as input and learns their five-dimensional representation from the training data.

species [as opposed to 15 per species in Fig. 1(b)]. The effects of (i)–(iii) are investigated independently in the following paragraphs.

The effect of (i) is tested by separately training the VT-based apD descriptor ML model on the different structure types [Fig. 1(d)]. The average MAE increases only marginally to 0.06 eV/atom. The minimum MAE across different structure types is 0.06 eV/atom in Fig. 1(d) compared to an average MAE of 0.04 eV/at in Fig. 1(c) showing that all structure types gained accuracy improvements by simultaneous training of the ML model in Fig. 1(c).

The effect of including additional structural details [(ii) from above] was studied by focusing on the perovskite structure type with space group 221 and Wyckoff sites a, b, and c [blue data points in Figs. 1(b) and 1(d)]. For the perovskite structure type, the only free parameter is the lattice constant, i.e., atom connectivity and interatomic distances are uniquely identified upon specifying the lattice constant. Therefore, the use of fractional atomic volume (and fractional-volume weighted elemental properties) in the VT-based apD descriptor does not provide additional structural information beyond that already encoded in the Wyckoff sites. The prediction MAE with a structure-type preclustering-based apI descriptor of 0.12 eV/at in Fig. 1(b) compared to only 0.06 eV/at in Fig. 1(d) rules out the additional contribution from apD structure details as the source of the superior performance of a VT-based apD descriptor.

To confirm that the inferior performance of the structure-type preclustering-based apI descriptor in Fig. 1(b) originates from nonoptimized elemental properties of relevant species [factor (iii) from above], a RLFNN is employed for the ML of FEs of materials belonging to perovskite structure type in Fig. 2. Instead of using pretabulated elemental properties, the RLFNN was designed to learn a five-dimensional representation of species from the training data itself. The RLFNN architecture is reported in Fig. 2(a), where learnable representations of species A, B, and C are passed through two fully connected hidden layers consisting of 10 and 4 neutrons to make the FE prediction. With this network design and for 82 chemical species in the perovskite structure type from OQMD, the number of trainable weights is only 645, of which 410 are for elemental representations.

In Fig. 2(b), the effect of training dataset size on predictability is presented by training the network on varying sized datasets and evaluating the performance on ∼1100 test data points. For fewer than 1500 points in the training dataset, the network results in overfitting indicated by large differences in the MAEs between training and testing datasets (regularization is not used in this network). With more than 2500 points in the training dataset, the network is able to achieve a prediction MAE of less than 0.06 eV/at, the same as that from the VT apD descriptor-based RF model in Fig. 1(d).

Figure 3 reports the performance of the structure-type preclustering apI descriptor-based RLFNN compared to the VT apD descriptor-based RF model for the 20 most frequently appearing structure types in OQMD in Fig. 3. These structure types consist of more than 380 000 materials from 13 space groups involving 41 different Wyckoff sites and 84 elements. For the RLFNN, a different neural network [similar to that in Fig. 2(a)] is trained on a different structure type. For the VT apD descriptor-based RF model, a single RF was trained by randomly collecting 80% of the data from each structure type.

For structure types consisting of more than 3500 materials, RLFNN results in MAEs of less than 0.12 eV/at. For structure types with fewer materials, the MAEs from RLFNN increase up to a maximum of 0.25 eV/at. In comparison, the VT apD descriptor-based RF model delivers a more regularized performance with maximum MAEs of only 0.10 eV/at across all considered structure types.

The RLFNN of Fig. 2(a) is designed to learn the elemental representations of species from the training dataset. As shown in Fig. 2(b), depending on the structure type, this requires an excess of 2500–3500 training data points. For densely populated structure types, while this trivial composition-only descriptor simplifies the material description and learning, for sparsely populated structure types, the prediction performance can be improved by sharing elemental information across structure types. To achieve this, we next focus our attention on the U-apI descriptor, which is capable of learning across different structure types.

IV. ML ACROSS STRUCTURE TYPES

As presented in Figs. 1(a) and 1(b), space group and Wyckoff sites are needed in the material description (along with the material composition) for learning across different structure types. While both space group and Wyckoff sites can be represented using simple scalars in the material description, this would suggest, for instance, that space groups 10 and 220 can be combined to form space group 230. To avoid this misrepresentation, space groups in the U-apI descriptor are represented by a one-hot vector of length 230. The occupied Wyckoff sites and corresponding atomic species are represented by a matrix of size 27×118 (referred to as a Wyckoff-species matrix in this work), where 27 is the maximum number of Wyckoff sites in any space group [21] and 118 is the number of elements from the Periodic Table. The element (i, j) of the Wyckoff-species matrix denotes the number of Wyckoff sites of type i occupied by an element

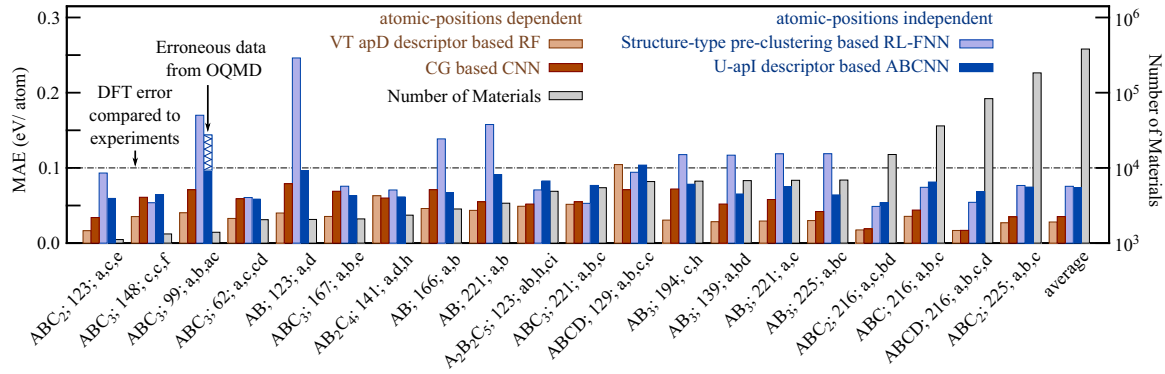

FIG. 3. ML across structure types. Comparison of atomic-position dependent and atomic-position independent descriptor-based ML models for the prediction of FE of the 20 most frequently appearing structure types from OQMD. For atomic-position dependent descriptors, VT apD descriptor-based RF [10] and CG-based CNN [12] are considered. For atomic-position independent descriptors, structure-type preclustering-based RLFNN and U-apI descriptor-based ABCNN are considered. The x-axis labels represent the material's stoichiometry, space-group number, and Wyckoff sites. For structure-type preclustering-based RLFNN, a different ML model is trained on each structure type.

of atomic number j. The U-apI descriptor for $CaTiO_3$ in the perovskite structure type is illustrated in Fig. 4(a).

The simple concatenation of the space group vector and Wyckoff-species matrix results in a tensor of size $230 \times 27 \times 118$. Training using simple ML models, such as linear regression and a feedforward neural network, on this gigantic tensor produces a minimum of 732 780 fitting parameters, which is more than the number of entries in modern high-throughput material databases, thereby potentially causing overfitting. To avoid this, the U-apI descriptor is used with weight-sharing ABCNN [22] as presented in Fig. 4(b). In this neural network, the columns of the Wyckoff-species matrix are passed through three weight-sharing layers followed by multiplication with species attention weights to get an embedded representation of the Wyckoff-species matrix. This embedded representation is next weighted by space group attention weights before finally passing through two consecutive fully connected hidden layers to make a prediction for the material property.

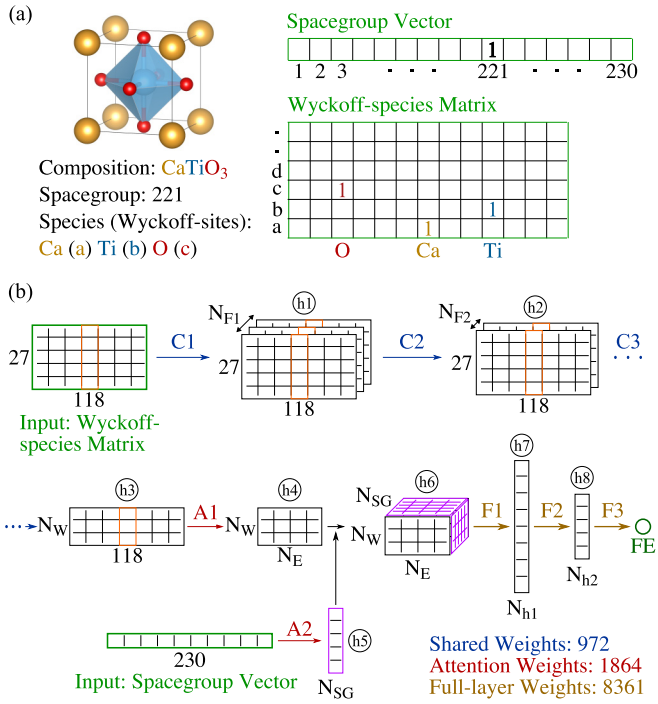

FIG. 4. U-apI descriptor and attention-based convolution neural network. (a) Illustration of U-apI descriptor (consisting of space group vector and Wyckoff-species matrix) for $CaTiO_3$ in the perovskite structure type. Only nonzero entries are explicitly listed for simplicity. (b) An attention-based convolution neural network for the prediction of material properties using the U-apI descriptor. The shared, attention, and fully connected layers are represented by blue, red, and brown. F1, F2, and F3 represent $(W1, b1)$, $(W2, b2)$, and $(W3, b3)$, respectively. The total number of trainable weights in the network are 11 197, which are obtained by setting the values of N_{F1}, N_{F2}, N_W, N_E, N_{SG}, N_{h1}, and N_{h2} at 4, 4, 4, 8, 4, 60, and 10, respectively.

Mathematically, the input Wyckoff-species matrix, $W_{i,j}$, is passed through two convolution layers consisting of N_{F1} and N_{F2} filters followed by a weight-sharing layer to obtain a N_W length embedded representation of columns of the Wyckoff-species matrix as

$$h1^k_{i,j} = f\left(C1^k_i \times W_{i,j}\right), \tag{2}$$

$$h2^l_{i,j} = f\left(\sum_k C2^{k,l}_i \times h1^k_{i,j}\right), \tag{3}$$

$$h3_{m,j} = f\left(\sum_{i,l} C3^m_{i,l} \times h2^l_{i,j}\right), \tag{4}$$

where k, l, and m loop from 1 to N_{F1}, N_{F2}, and N_W, $f()$ is the activation function introducing nonlinearity in the network, and $C1$, $C2$, and $C3$ are trainable convolution filter weights. The obtained embedded representation of the Wyckoff-species matrix is next weighted by elements and space group attention weights to obtain the hidden material representation $h6_{m,n,q}$ as

$$h4_{m,n} = \sum_j A1_{j,n} \times h3_{m,j}, \tag{5}$$

$$h5_q = \sum_p A2_{p,q} \times S_p, \tag{6}$$

$$h6_{m,n,q} = h4_{m,n} \times h5_q, \tag{7}$$

TABLE I. The effect of different hyperparameters on the training error of U-apI descriptor based ABCNN.

N_{F1}	N_{F2}	N_W	N_E	N_{SG}	N_{h1}	N_{h2}	Number of trainable parameters	Activation function	Training error (eV/at)
4	4	4	4	4	20	10	3885	Sigmoid	0.10
8	4	4	4	4	20	10	4425	Sigmoid	0.11
4	8	4	4	4	20	10	4749	Sigmoid	0.12
4	4	8	4	4	20	10	5597	Sigmoid	0.11
4	4	4	8	4	20	10	5637	Sigmoid	0.08
4	4	4	12	4	20	10	7389	Sigmoid	0.08
4	4	4	4	8	20	10	6085	Sigmoid	0.11
4	4	4	4	4	40	10	5385	Sigmoid	0.08
4	4	4	4	4	60	10	6885	Sigmoid	0.09
4	4	4	4	4	80	10	8385	Sigmoid	0.09
4	4	4	4	4	20	20	4105	Sigmoid	0.11
4	4	4	4	4	20	10	3885	Elu	0.11
4	4	4	4	4	20	10	3885	Softplus	0.12
4	4	4	4	4	20	10	3885	Softsign	0.10
4	4	4	4	4	20	10	3885	Tanh	0.10

where S_p is the input space group vector of length 230, $A1$ and $A2$ are trainable elemental and space group attention weights, and n and q loop from 1 to N_E and N_{SG}. The hidden material representation $h6_{m,n,q}$ is finally passed through two fully connected layers of N_{h1} and N_{h2} neurons to obtain the prediction for the material property as

$$h7_r = f\left(b1_r + \sum_{m,n,q} W1^r_{m,n,q} \times h6_{m,n,q}\right), \tag{8}$$

$$h8_s = f\left(b2_s + \sum_{r} W2_{r,s} \times h7_r\right), \tag{9}$$

$$FE = b3 + \sum_{s} W3_s \times h8_s, \tag{10}$$

where $(W1, b1)$, $(W2, b2)$, and $(W3, b3)$ are trainable fully connected layer weights and r and s vary from 1 to N_{h1} and N_{h2}.

The neural network is implemented using the openly available tensor library Tensorflow [23]. The trainable weights are all initialized using a truncated Normal distribution of zero mean and unity standard deviation. The training is performed using the Adam optimizer [24] with an initial learning rate of 1×10^{-3}. The training is performed for 10 000 epochs with a batch size of 1024. The MAE is used as the loss function for the training of the network. For training on the top 20 most frequently appearing structure types from OQMD, the loss on individual materials is scaled inversely by the number of materials in the structure type. The values of the hyperparameters N_{F1}, N_{F2} N_W, N_E, N_{SG}, N_{h1}, N_{h2}, and the activation function are tested on the training error, and the results are summarized in Table I.

The final network has a total of 11 197 trainable weights that are obtained by setting the values of N_{F1}, N_{F2}, N_W, N_E, N_{SG}, N_{h1}, and N_{h2} at 4, 4, 4, 8, 4, 60, and 10, respectively. The sigmoid function is used to introduce nonlinearity in the network. The python code for generating the U-apI descriptor from structure files and for performing ML using U-apI descriptor-based ABCNN can be obtained through GITHUB [25].

The prediction performance of U-apI descriptor-based ABCNN is reported in Fig. 3 for the 20 most frequently appearing structure types from the OQMD. The reported MAEs in Fig. 3 are on test datasets that are obtained by randomly assigning total materials from each structure type into training and test datasets consisting of 80% and 20% of the total materials, respectively. For these structure types, the VT apD descriptor-based RF and the structure-type preclustering apI descriptor-based RLFNN result in maximum MAEs of 0.10 and 0.25 eV/at and a factor of 5 difference in the performance on sparsely and densely populated structure types. For the U-apI descriptor-based ABCNN, with an exception of structure-type (ABC_3; 99; a,b,ac), the maximum MAE is 0.10 eV/at with only a factor of 2 difference in the performance across structure types.

As can be seen in Fig. 1(b), for structure-type (ABC_3; 99; a,b,ac), several materials have FE as high as 5 eV/at. These materials have compositions $HfTlO_3$, $TmTlO_3$, $LuTlO_3$, $SmTlO_3$, $HoTlO_3$, and $OsTlO_3$. All of these materials are erroneously reported to have identical lattice constants of 3.398 and 3.639AA in the a and c directions in the relaxed configurations in the OQMD, thereby suggesting failed/nonrelaxed geometries. For these materials, the prediction errors are as high as 6 eV/at. Removing these erroneous entries from the dataset brings down the MAE to below 0.10 eV/at for (ABC_3; 99; a,b,ac) structure-type using the U-apI descriptor-based ABCNN (not shown in Fig. 3).

The ability of the U-apI descriptor-based ABCNN to handle structurally diverse materials is tested by performing the learning on more than 428 000 materials from OQMD. These materials belong to 90 different space groups with 289 participating Wyckoff sites and 84 elements and are obtained by eliminating materials for which participating Wyckoff sites appeared in fewer than 100 materials in the OQMD [26]. On these materials, the U-apI descriptor-based ABCNN results in a test MAE of only 0.07 eV/at.

The CNN has also been employed recently by Xie *et al.* [12] with a CG descriptor to predict the material properties of inorganic periodic solids. When used in the prediction of FE of materials from the OQMD [27], this CG-based CNN results

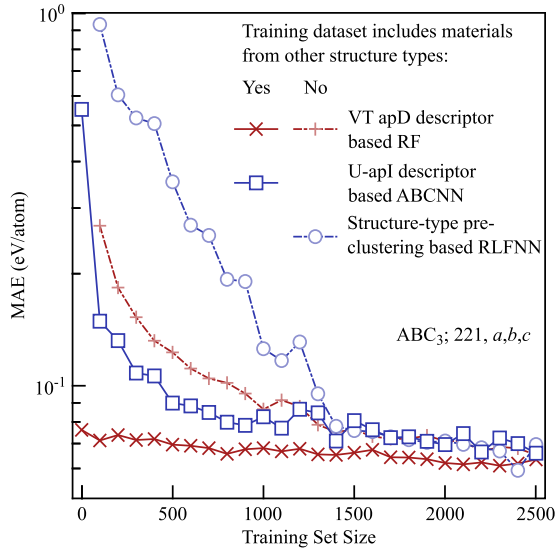

FIG. 5. ML on unseen structure types. The variation of test MAEs with the training set size for perovskite structure-type materials. For models trained on all data, all materials from Fig. 3 (except those belonging to the perovskite structure type) are used in training. For materials belonging to the perovskite structure type, the total materials are randomly split into train and test datasets consisting of 80% and 20% of the total materials. The varying number of materials from the 80% set are added in the training dataset, and the prediction performances are evaluated on a test dataset consisting of only perovskite structure-type materials.

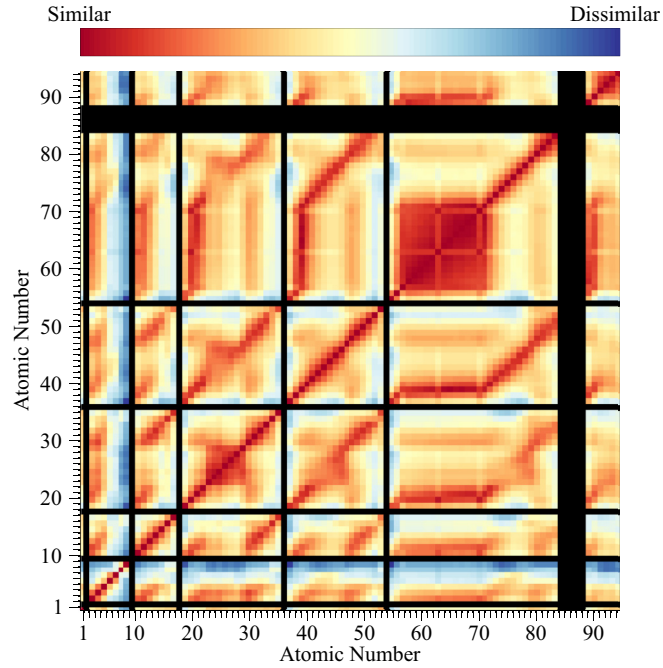

FIG. 6. Physical significance of attention weights. The heatmap shows the similarity of elements to form chemical compounds. The similarities between elements are characterized using the Euclidean distance of elemental attention weights learned by the neural network. The nonparticipating elements are represented using black.

in a similar prediction performance to that of the VT apD descriptor-based RF (Fig. 3). The CG-based CNN, however, employs additional structural information about relaxed bond lengths of neighboring atoms, which, similar to VT, requires relaxation of atomic coordinates and therefore inhibits prediction for new materials. In contrast, the U-apI descriptor-based ABCNN employs only space group and Wyckoff sites of atoms and therefore allows for screening of millions of compositionally different new materials within the same structure type.

The prediction performance of a U-apI descriptor-based ABCNN on new, unseen structure types is presented in Fig. 5. For this, all materials other than those belonging to the perovskite structure type from Fig. 3 are used in the training. For materials belonging to the perovskite structure type, 20% of the total materials are randomly picked and are used to evaluate the model performance. From the remaining 80% of the materials with perovskite structure type, a varying number of materials are added in the training set, and the dependence of model performance on training set size is evaluated.

Without any perovskite structure-type material in the training dataset, the test MAE from the U-apI descriptor-based ABCNN is 0.55 eV/at. With an inclusion of 500 perovskite structure-type materials in training, the test MAE drops to less than 0.09 eV/at. This is more than a factor of 4 improvement in the prediction performance compared to that from RLFNN (trained only on materials from the perovskite structure type) for which the test MAE is 0.35 eV/at for 500 materials in the training dataset and the test MAE drops to less than 0.10 eV/at when 1300 materials are included in the training dataset. This demonstrates that the representation weights learned by ABCNN are applicable over different structure types.

When trained on only perovskite structure-type materials, the prediction performance of the VT apD descriptor-based RF model is inferior to that from U-apI descriptor-based ABCNN, and the test MAE reduces to 0.09 eV/at only when 1000 materials are included in the training dataset. On training the VT apD descriptor-based RF on all materials, however, the test MAE is 0.08 eV/at even when no perovskite structure-type materials are included in the training dataset.

The physical significance of attention weights of elements in the U-apI descriptor-based ABCNN is tested by identifying the similarities of elements to form chemical compounds. For this, the Euclidean distance is evaluated between elemental representation attention weights (A1 in Fig. 4) as learned by the ABCNN. The resulting heatmap of distances is presented in Fig. 6. As can be seen from Fig. 6, the ABCNN learned correct elemental trends include (i) chemical similarities of the same group elements of the Periodic Table to form chemical compounds, (ii) correctly identifying 8, 8, 18, and 18 elements (corresponding to the number of elements in the Periodic Table periods 2, 3, 4, and 5) after which chemical bonding properties are similar, and (iii) chemical similarities between lanthanides except for elements Eu and Yb, which have half- and full-filled f-orbitals, respectively. We note that these properties are learned solely using the material's formation energies. We further note that this heatmap is similar to that obtained by Glawe *et al.* [28] using the statistical analysis

of experimentally known materials, thereby establishing the physical relevance of attention weights in the ABCNN.

In summary, we used a crystallographic space group and Wyckoff sites to describe the structure details of materials in ML models. We find that space group and Wyckoff sites fix the symmetry environment of materials where composition-only material descriptors are sufficient to distinguish different materials. We show that with sufficient materials in this fixed space group and Wyckoff-site space and with a representation learning feedforward neural network, the minimal material descriptor employing only atomic numbers is able to match the performance of the apD descriptor. Lastly, we present a U-apI descriptor that, when used with an attention-weight sharing convolution neural network, is able to learn across diverse structure types and delivers a prediction MAE of 0.07 eV/at. This accurate and regularized performance of a U-apI descriptor across diverse structure types without explicit use of atomic positions (as in the case of other material descriptors) establishes the broader applicability of the U-apI descriptor in the ML of material properties.

ACKNOWLEDGMENTS

We thank Raul Abram Flores, Jr. and Christopher Paolucci from Stanford University for their valuable feedback. We acknowledge financial support from the Toyota Research Institute.

[1] Q.-J. Hong and A. van de Walle, J. Chem. Phys. **139**, 094114 (2013).
[2] K. Fujimura, A. Seko, Y. Koyama, A. Kuwabara, I. Kishida, K. Shitara, C. A. Fisher, H. Moriwake, and I. Tanaka, Adv. Energy Mater. **3**, 980 (2013).
[3] A. M. Deml, V. Stevanović, C. L. Muhich, C. B. Musgrave, and R. O'Hayre, Energy Environ. Sci. **7**, 1996 (2014).
[4] J. Carrete, W. Li, N. Mingo, S. Wang, and S. Curtarolo, Phys. Rev. X **4**, 011019 (2014).
[5] A. Seko, T. Maekawa, K. Tsuda, and I. Tanaka, Phys. Rev. B **89**, 054303 (2014).
[6] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, npj Comput. Mater. **1**, 15010 (2015).
[7] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, Int. J. Quantum Chem. **115**, 1094 (2015).
[8] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, Phys. Rev. Lett. **114**, 105503 (2015).
[9] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, and E. K. U. Gross, Phys. Rev. B **89**, 205118 (2014).
[10] L. Ward, R. Liu, A. Krishna, V. I. Hegde, A. Agrawal, A. Choudhary, and C. Wolverton, Phys. Rev. B **96**, 024104 (2017).
[11] A. Seko, H. Hayashi, K. Nakayama, A. Takahashi, and I. Tanaka, Phys. Rev. B **95**, 144110 (2017).
[12] T. Xie and J. C. Grossman, Phys. Rev. Lett. **120**, 145301 (2018).
[13] T. Hahn, U. Shmueli, and A. J. C. Wilson, *International Tables for Crystallography* (Reidel, Dordrecht, 1983), Vol. 1.
[14] E. S. Fedorov, *Symmetry of Crystals (American Crystallographic Association Monograph No. 7)* (American Crystallographic Association, Buffalo, NY, 1971), pp. 50–131.
[15] A. Schoenflies, *Theorie der Kristallstruktur* (Borntraeger, Berlin, 1923).
[16] R. W. G. Wyckoff, *The Analytical Expression of the Results of the Theory of Space-groups* (Carnegie Institution of Washington, Washington, 1922), p. 318.
[17] R. T. Downs and M. Hall-Wallace, Am. Mineral. **88**, 247 (2003).
[18] The elemental properties used in the description of atoms are as follows: (i) atomic number, (ii) number of electrons, (iii) number of protons, (iv) atomic weight, (v) mass number, (vi) dipole polarizability, (vii) vdw radius uff, (viii) melting point, (ix) boiling point, (x) number of neutrons, (xi) covalent radius pyyko, (xii) covalent radius cordero, (xiii) vdw radius, (xiv) atomic radius rahm, and (xv) vdw radius mm3. These atomic properties are taken from the MENDELEEV database [29].
[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, J. Machine Learning Res. **12**, 2825 (2011).
[20] F. Legrain, J. Carrete, A. van Roekeghem, S. Curtarolo, and N. Mingo, Chem. Mater. **29**, 6220 (2017).
[21] M. I. Aroyo, J. M. Perez-Mato, C. Capillas, E. Kroumova, S. Ivantchev, G. Madariaga, A. Kirov, and H. Wondratschek, Z. Krist.-Cryst. Mater. **221**, 15 (2006).
[22] W. Yin, H. Schütze, B. Xiang, and B. Zhou, arXiv:1512.05193.
[23] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, in *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16 (USENIX Association, Berkeley, CA, 2016), pp. 265–283.
[24] D. P. Kingma and J. Ba, arXiv:1412.6980.
[25] https://gitlab.com/ankitjainmeiitk/abcnn.
[26] This screening eliminated only 7275 of 435 583 total materials from OQMD.
[27] We used default (optimized) values of hyperparameters as provided by authors in their code.
[28] H. Glawe, A. Sanna, E. Gross, and M. A. Marques, New J. Phys. **18**, 093011 (2016).
[29] Ł. Mentel, Mendeleev—A python resource for properties of chemical elements, ions and isotopes (2014), https://bitbucket.org/lukaszmentel/mendeleev.