Journal of Applied Physics ARTICLE scitation.org/journal/jap

Practicing deep learning in materials science: An evaluation for predicting the formation energies

Cite as: J. Appl. Phys. **128**, 124901 (2020); doi: 10.1063/5.0012411
Submitted: 30 April 2020 · Accepted: 4 September 2020 ·
Published Online: 23 September 2020

Liyuan Huang and Chen Ling[a)]

AFFILIATIONS

Toyota Research Institute of North America, Ann Arbor, Michigan 48105, USA

Note: This paper is part of the special collection on Machine Learning for Materials Design and Discovery
[a)]**Author to whom correspondence should be addressed:** chen.ling@toyota.com

ABSTRACT

Deep learning in recent years has entered the chemistry and materials research arsenal with many successful accomplishments in tasks considered to be intractable using traditional means. However, the widespread application of this data-driven technology is still challenged by the requirement of large training data, poor model interpretability, and hard-to-detect errors that undermine the soundness of conclusion. Here, we performed a systematic study for the modeling of the formation energies of inorganic compounds using deep learning. Our results proved the advantage of deep learning methods over several non-deep learning methods in this specific task and demonstrated the abstraction of knowledge using deep learning, which was a unique ability compared to non-deep learning methods. Several aspects that critically affected the conclusion were also highlighted, including the importance to rigorously compare model performance with the same dataset, the design of input representation, and the careful selection of model architecture. Findings from the current study demonstrate the capabilities of deep learning solving complicated problems in materials research and serve as new guidelines for future practicing of deep learning in this field.

I. INTRODUCTION

In the modern term, deep learning (DL) is usually referred to as a neural network with more than three layers. Each layer contains parallel located neurons fully connected to those in the front and back adjacent layers, the combination of which gives a hierarchical data representation. More complex network structures, such as convolutional neural network, recurrent neural network (RNN), autoencoder, and others, are also considered in the category of deep learning. The concept of deep learning can be traced back to 1943 when the mathematical representation of neural network was for the first time proposed by McCulloch and Pitts.[1] Plentiful breakthroughs were made in the following decades, such as the first artificial neural network for pattern recognization,[2] learning from back-propagating errors,[3] the use of convolutional neural network in handwritten digits recognition,[4] the long short-term memory (LSTM) model,[5] and gradient-based learning.[6] At present, deep learning is becoming the most popular technique in the field of machine learning. Along with the emerging modern big data era and the great enhancement on computation power brought by more advanced graphic processing units, deep learning models have raised rapidly increased performance in a variety of problems ranging from computer vision, speech recognition, natural language processing to medical diagnosis, and the Go game playing.[7,8]

Extensive applications of deep learning have also been found in the chemistry and materials domain. For example, in the chemical toxicity challenge of TOX21, the highest performance was scored by a deep learning-powered application DeepTox.[9] Deep learning was used as the replacement of expensive quantum mechanical calculations to obtain fundamental properties such as wave functions,[10,11] formation energies,[12–17] and electronic ground and excited states,[18] and in tasks of chemical literature reading,[19–22] crystalline structure prediction,[23] and microstructure classification.[24–26] In a review paper, Goh *et al.* showed that the deep neural network (DNN)-based models performed equivalently or exceeding to the best non-deep learning methods in these tasks.[27] Although deep learning has accomplished many tasks considered to be intractable in traditional means, the widespread application of this data-driven technique is still challenged by the requirement of a relatively large amount of data and the black-box-like algorithm that prevents the model interpretability.[28,29] Deep learning is also prone to hard-to-detect errors

that undermine the soundness of conclusions, such as target leakage into training data, uncareful selection of modeling methods, and unnecessary complexification.[29–31] A thorough study to assess the performance of deep learning is therefore of importance to set up guidelines for future practicing in the chemistry and materials domain.

Here, we evaluated deep learning in a representative problem in materials science, predicting the formation energy of inorganic compounds. The modeling of formation energy is one of the most representative applications of machine learning in materials science, and the usage of deep learning to tackle this problem has been reported in a number of studies.[13,15–17] The current work studied several factors affecting the performance of deep learning models that have not been systematically analyzed in the past, including the training data, the algorithm variation, and the representation of input. Our results highlighted several important aspects of deep learning in modeling materials properties, and these results were thus served as new guidelines to future practicing of deep learning in this field.

II. METHODS

Deep neural network (DNN) and convolution neural network (CNN) were implemented through the high-level Keras API provided by TensorFlow. Recurrent neural network (RNN) was implemented through the low-level TensorFlow API. Non-deep learning models were implemented through the Scikit-learn python library.[32] A standalone iteratively grid search on the depth of the deep learning network, the number of training epochs, and the learning rate was carried out for every individual experiment. For the optimization of the DNN model, a total of 3, 6, 10, 17, or 21 layers with rectified linear unit as activation function were tied out to find the best structure. A set of learning rates ranging from $1e^{-1}$ to $1e^{-4}$ were used to get the best score. Two learning epochs, 100 and 500, were tested to find out the best training time. The rest of the hyperparameters were either set as default or by manually selected values. To implement the CNN architecture, the one-dimensional elemental dictionary that was used to define the compositional representation was transformed into a 10×10 two-dimensional one, in which each unit represented one unique element. Note that 86 units were utilized in this matrix, where 86 stood for the number of different elements in the dataset. With this two-dimensional dictionary, the composition was embedded as a 10×10 matrix. For the implementation of the RNN architecture, a three-dimensional training data matrix $R^{A \times B \times C}$ was created as input to the model, where the depth C of the matrix was set to be 100 to allow the inclusion of all the appeared elements. The width B of the matrix was set to be 8, which was the maximum number of elements appearing in a compound in the dataset. The height A was the total number of entries in the dataset. We used the TensorFlow implementation of a long short-term memory (LSTM) cell to generate hidden states and the output. An attention mechanism was used in conjunction with the LSTM cell in the block to learn and give weighted importance on each element in a compound. Afterward, the hidden state vector of the LSTM cell, calculated attention context vector, and two most important element vectors ordered by attention weights were concatenated to be used as a single input feature for the following abstraction layers, which were composed of nine highway network layers and one fully connected layer.

We used the same ratio of 81:9:10 for training:validation:testing in all the experiments. Before splitting, the whole dataset was randomly shuffled to remove artificial patterns. The mean squared error was selected as the cost function for the DNN and non-deep learning approaches. We used the mean absolute error (MAE) and Huber loss as the cost function for CNN and RNN, respectively. The stochastic gradient descent (SGD) optimizer,[33] Adadelta,[34] and Adam[35] were selected as the optimizer for the DNN, CNN, and RNN models, respectively. Mean absolute error (MAE) was selected as the evaluation measurement. An early stopping mechanism was adopted in the DNN models to prevent the learning process from overfitting. For the training of non-deep learning models, the optimization was carried out through the random search function provided by the Scikit-learn library.

III. RESULTS AND DISCUSSION

A. Influence of data sources

Despite the vast number of machine learning techniques in handling a variety type of problems, the basic principle is rooted in inferring the behavior from learning patterns in the dataset. Naturally, the degree of fitting is perceptibly determined by the dataset itself including the ways to scale the property and construct the features. We started our study to look into the influence of the dataset on predicting the formation energy of inorganic compounds. In particular, we considered two data sources: the materials project (MP) database[36] and the open quantum materials database (OQMD).[37,38] Both MP and OQMD use the first-principles electronic method at a similar level of theory to high-throughput calculate the thermodynamic, structural, and electronic properties of inorganic compounds. In our study, the MP dataset included the formation energies for 56 661 compounds with unique compositions. For OQMD, we considered two versions containing 256 622 (the same dataset used in Ref. 13, OQMD1) and 337 996 unique inorganic compounds (version 1.2, OQMD2). All three datasets excluded entries for single substances and outlines located out of the 5σ range, where σ is the standard deviation (SD) of the formation energy data.

Table I lists the mean absolute error (MAE) of the DNN models to predict the formation energies in these three datasets. The models used a total of 145 hand-crafted chemical descriptors as input features.[39] All the MAEs reported here and afterward were evaluated on the testing set. There are several observations that should be discussed. First, the standard deviation calculated from five independent divisions of training, validation, and testing sets

TABLE I. Mean absolute error (MAE, eV/atom) and standard deviation (SD, eV/atom) to predict the formation energy using different data sources.

	MP	OQMD1	OQMD2
Dataset size	56 661	256 622	337 996
MAE	0.130	0.055	0.060
SD	0.0044	0.0009	0.0006

were less than 5% of MAE, indicating for the prediction of formation energy a dataset with the size above 50 000 was adequate to alleviate the statistical variance caused by the random training–validation–testing splitting. Second, even for the same data source of OQMD, different versions affected the MAE by ∼10%. Interestingly, opposite to the common belief that more data benefits to reduce the prediction error, the OQMD2 dataset, which contained 32% more examples than OQMD1, gave larger MAE. This is probably due to the new examples added to the database included a higher portion of compounds whose existence is only theoretically hypothesized. Last but not least, the MAEs for the MP and OQMD datasets differed by more than twice, the reason of which will be further explored.

Both MP and OQMD databases were built following the similar procedure by starting with the Inorganic Crystal Structure Database (ICSD) of experimentally known compounds and expanding it with hypothetical compounds. At first glance, for the OQMD dataset with a denser population in the chemical space, the closer distance between a testing point and training examples should benefit to decrease the predicting error. To examine this, we randomly populated an OQMD2 subset with the same size as the MP dataset and performed the modeling on the sampled subset. The averaged MAE was 0.105 eV/atom from ten random samplings. While this experiment proved the apparent importance of data size in determining the testing error in our problem, a gap of ∼0.025 eV/atom was left unexplained after eliminating the size difference between training datasets.

A deeper analysis revealed more fundamental differences between MP and OQMD2 datasets. Figure 1(a) shows the population of elements in these datasets. The OQMD2 datasets contained a high population of metal elements, while the MP dataset included more elements from p-block, which indicated the sampling difference between these two databases: the MP had more ceramic compounds and the OQMD2 had more metals and intermetallics. Such sampling difference was also reflected in the crystalline structure space. Figure 1(b) shows the distributions of space groups in MP and OQMD2 datasets. The OQMD2 dataset 1 was dominated by cubic structures, especially those with the space group of $Fm\bar{3}m$ and $F\bar{4}3m$. On the other side, the MP dataset was more widely distributed in triclinic, monoclinic, orthorhombic, and cubic families of crystalline structures.

These sampling differences in the compositional and structure space led to the drastically different distributions of formation energies in the MP and OQMD2 datasets. As shown in Fig. 1(c), the data in the MP dataset were distributed at more negative values and, compared to OQMD2, MP had a significantly less population of samples with positive formation energies. Consequently, we found that the DNN model scored poorly for cases whose formation energies were positive due to the lack of training examples in this special range. Despite a small population in the entire dataset, the compounds with positive formation energies contributed to ∼0.031 eV/atom. Combining these analyses, we attributed the difference of 0.07 eV/atom of MAE for the models using MP and OQMD2 datasets to two major parts: the difference of data size,

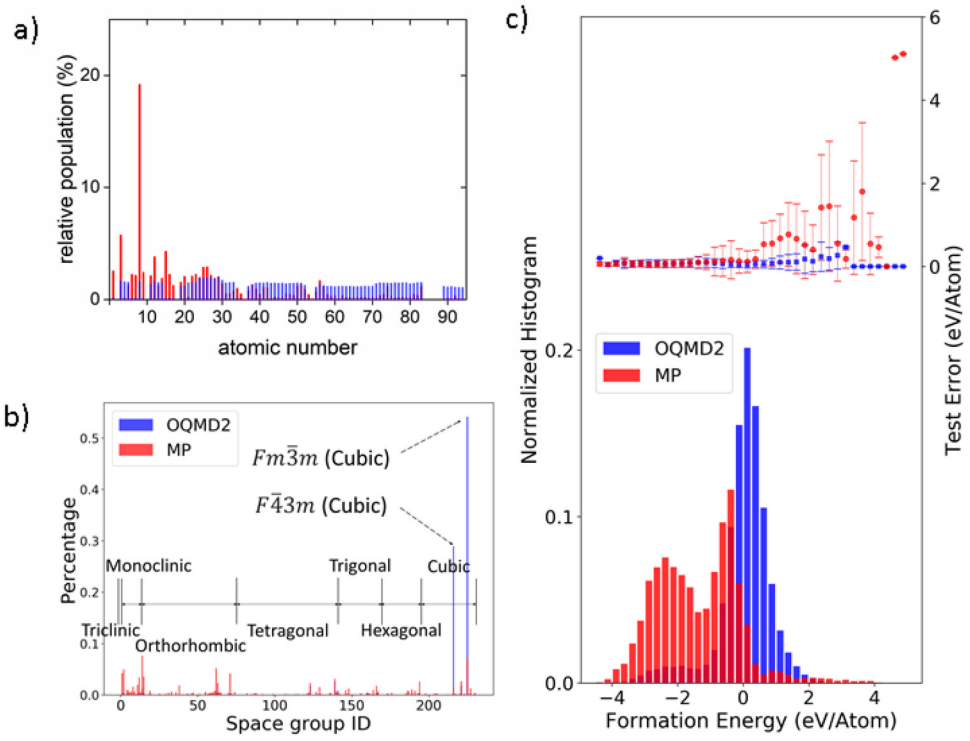

FIG. 1. Three analyses revealed more fundamental differences between MP and OQMD2 datasets. (a) Populations of elements in the MP and OQMD2 datasets. Red: MP dataset; blue: OQMD2 dataset. (b) Distribution of space groups in the MP and OQMD2 datasets. (c) The lower subplot shows the normalized truth formation energy distributions of MP and OQMD2, while the upper subplot shows the averaged test errors of the sampled bins that range from the lowest possible formation energy value to the largest possible formation energy value.

TABLE II. MAE (eV/atom) to predict the formation energy using non-deep learning algorithms.

	Linear algorithms				Non-linear algorithms			
	LR	LASSO	LRR	elastic net	RF	XGBoost	LightGBM	CatBoost
MAE	0.267	0.269	0.267	0.267	0.066	0.129	0.070	0.171

which contributed ~0.04 eV/atom of the difference, and the difference of data sampling, which contributed ~0.03 eV/atom.

B. Deep neural network vs non-deep learning algorithms

The next factor that we examined was the variations of learning algorithms. We evaluated the performance of several linear algorithms including linear regression (LR), least absolute shrinkage and selection operator (LASSO), linearized ridge regression (LRR), and elastic net; non-linear algorithms including random forest (RF); and more advanced tree-based ensemble learning algorithms such as XGBoost, LightGBM, and CatBoost on the same OQMD2 dataset. Two other non-linear algorithms of support vector regression (SVM) and kernel ridge regression (KRR) were tested but failed to yield an actual result due to the formidable computational cost for training such a large volume of data. We kept the same chemical descriptors for training and strictly used the same training–validation–testing splitting to avoid any variance caused in this procedure.

Table II lists the MAE for different non-deep learning algorithms in comparison with that for the DNN model as mentioned above. All four linear methods performed poorly compared to DNN on predicting the formation energy, suggesting that the relationship between this property of inorganic compound was far beyond linear with simple chemical descriptors. Among the four non-linear algorithms, the random forest scored the best MAE. However, the performance was still appreciably worse than the DNN model with the MAE ~10% higher than that for the DNN model.

Previously, Jha *et al.* showed a deep neural network model using the fractional amount of each element in a compound as the input outperformed the RF model using chemical descriptors in predicting the formation energy once the size of training dataset exceeded ~4000.[13] While the DNN model used the compositional representation, which will be discussed in Sec. III C, here we compared the size effect while keeping the same input of chemical descriptors. As shown in Fig. 2(a), the learning curves for the DNN and random forest models behaved remarkably similar as the predicting accuracy increased when more samples were added for the training. The DNN models performed consistently better than the random forest ones even for a small dataset of fewer than 1000 samples. However, for the data size below 30 000, the differences of MAEs were marginally small (2%–4%), suggesting that for the data size in the range of 1000–10 000, the two algorithms had the performance at the same level for this task.

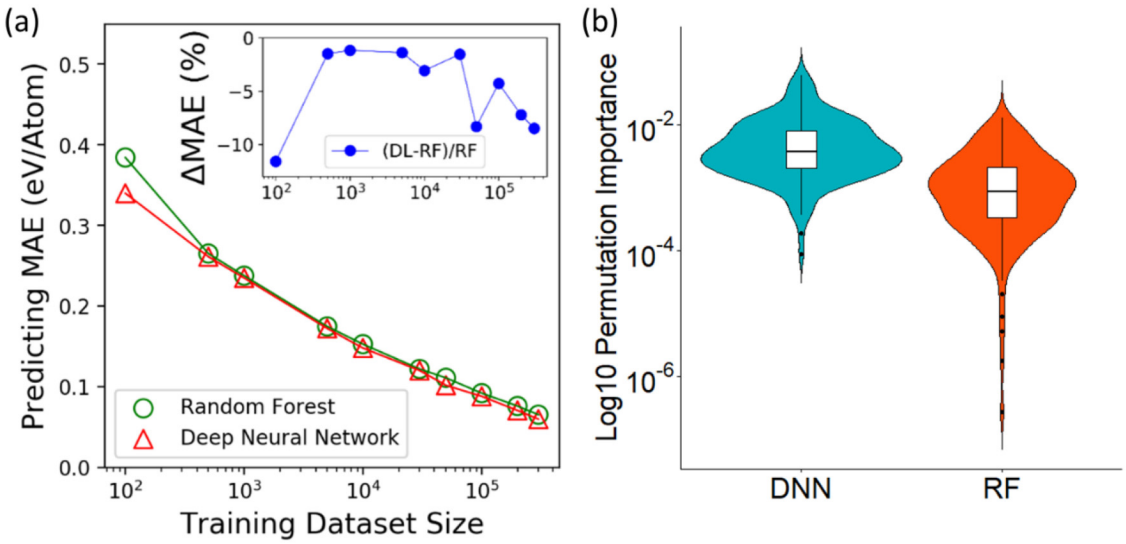

FIG. 2. Comparison of the deep neural network and random forest models for the prediction of formation energies. (a) Influence of the size of training data on the prediction of formation energies. The inset shows the relative improvement using a deep neural network. (b) Distribution of features with positive permutation importance of features for deep neural networks and random forest models while learning on the formation energy averaged by ten randomly sampling runs. In the deep neural network model, all 145 chemical descriptors had positive importance, while in the random forest model, 142 of them had positive importance.

To understand the reasoning that the DNN outperformed the random forest in predicting the formation energy, we interpreted both models by evaluating the importance of each individual descriptor in modeling this property. We used the permutation importance, a reliable and universally applicable measurement, to evaluate the importance of features in deep neural network and random forest models.[40,41] The permutation importance of a feature was calculated by taking the difference between a baseline metric score and the score after permuting that specific feature from the input. For both DNN and random forest models, the high ranked features included electronegativity, number of filled p valence orbitals, and maximum/mean ionic character between any two elements in the material. This indicates these models indeed identified similar correlations between the formation energy and chemical features. Most other features had low permutation importance. An important observation from the permutation importance analysis was the significant population of features with marginal importance in the random forest model, as shown in Fig. 2(b). Removing these features slightly improved the random forest model by 0.0004 eV/atom. It indicated that these features were indeed noises in the random forest modeling. The distribution of the permutation importance for the DNN model was obviously concentrated at higher values compared to that for the RF, suggesting better utilization of these features in DNN. We note, however, that a decent number of features still had low permutation importance of the DNN model, indicating that potential improvement from feature selection could be achieved. We will discuss the influence of using different input descriptors in Sec. III C.

C. Chemical descriptor vs compositional representation

Our study so far used hand-crafted descriptors to represent the chemical information of individual compounds. These descriptors embedded classical chemical terms such as electronegativity, atomic radius, and ionization energy.[39] Alternatively, a compound can be represented as simple as its chemical composition. For instance, let us define an arbitrary dictionary of elements such as [H, Li, Be, O, …]. The compound of Li_2O can be assigned with a vector of [0, 2/3, 0, 1/3, 0,…] where 2/3 and 1/3 are the fractional amount of Li and O in this compound, respectively, and the indexes correspond to the position of Li and O in the dictionary. Apparently, this "one-hot" vector only encodes the compositional information and, in sharp contrast with the chemical descriptor, does not contain any prior knowledge from human wisdom.

It is of interest to compare the performance of DNN models when using chemical descriptors and the one-hot compositional representation for the prediction of formation energy. Jha *et al.*[13] and Zeng *et al.*[17] showed that DL algorithms were capable to detect patterns underlying the formation energy data using only the compositional information. Jha *et al.* further showed that the DNN model using the compositional representation performed better than the RF one using the chemical descriptors.[13] However, the comparison of DNN models using different input features has not been thoroughly analyzed up to our knowledge. To shed light on this question, we trained two DNN models using chemical descriptor and compositional representation as the input. The examination was carried out in MP, OQMD1, and OQMD2 datasets. As showed in Fig. 3(a), the models with compositional representation consistently outperformed those with chemical descriptors on all the three datasets. The improvement was substantial, 5%, 14%, and 13% for MP, OQMD1, and OQMD2, respectively.

The performance of the DNN model using compositional representation was attributed to the capability to abstract knowledge from its "deep" feedforward architecture compared to some traditional machine learning method or shallow neural network.[42] The abstraction with a feedforward network with a backpropagation updating mechanism captures the most salient features out of all the input features and passes to the next layer. It enables the network to process the training data gradually and the abstracted layers have often been taken out as transfer learning knowledge to other studies. In Fig. 3(b), we visualized the self-taught

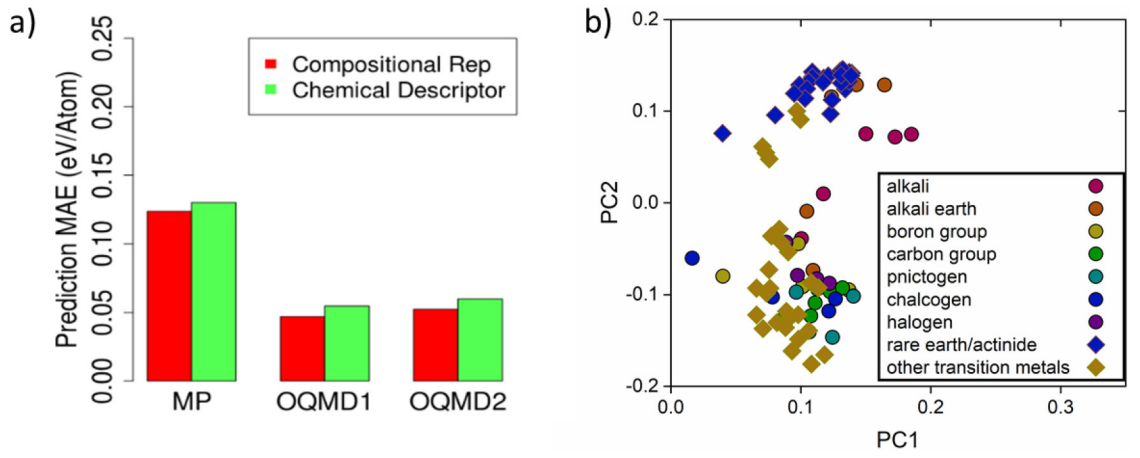

FIG. 3. Deep neural network in predicting the formation energy using compositional representation. (a) Performance benchmarked on different datasets. (b) Projecting the self-taught representation in the principal component space along the first (PC1) and second principal component directions (PC2).

representation abstracted from the first fully connected layer in the deep neural network model in the principal component space. The clustering of elements clearly reflected their group-wise similarities, indicating this neat representation alleviated the potential overfitting induced by the vast number of hand-crafted chemical descriptors that barely contributed to the model performance as revealed in Fig. 2(b).

In spite of improved accuracy through leveraging simple composition representations as features input, there are still some requirements that should be discussed further. For example, one can imagine the scenario that the dataset contains few compounds with one element. In such a case, the DNN may not be able to decipher the chemical knowledge of that specific element accurately. We used the following experiment to examine this scenario. First, we divided all compounds in the OQMD2 dataset into a group of Mg-containing compounds that included 16 252 samples and a non-Mg-containing group containing 321 744 compounds. Next, we randomly sampled 10 000 compounds from the non-Mg-containing group and kept it as the testing set. The rest of this group was used to train a baseline model using DNN and compositional representation. After that, we added a randomly selected set of samples from Mg-containing groups to the training set and trained the new DNN model. The testing was carried out on the pre-isolated testing set as well as on Mg-containing compounds that were not used for training.

Figure 4 shows the MAEs of predicting the formation energies of non-Mg-containing and Mg-containing compounds when the training data were varied by the number of Mg-containing compounds. The accuracy of the non-Mg-containing testing set was barely affected, indicating that adding a new dimension in the elemental space did not affect the interpretation of other dimensions. On the other side, a training set containing too few Mg-compounds resulted in significantly worse performance to predict the formation energies of Mg-containing compounds, in agreement with the expectation that the DNN model was not able to abstract the chemical knowledge of this particular element from deficient examples. However, this behavior was quickly alleviated by adding Mg-compounds in the training set. After around one-fifth of the total magnesium-containing compounds were added for training, the model behaved similarly as the times it saw the majority of the Mg-containing data. It indicated at least a decent performance when the number of Mg-containing compounds in the training set exceeded ∼3000. The experiments were performed using calcium- and potassium-containing compounds as examples, and the same trends were consistently observed.

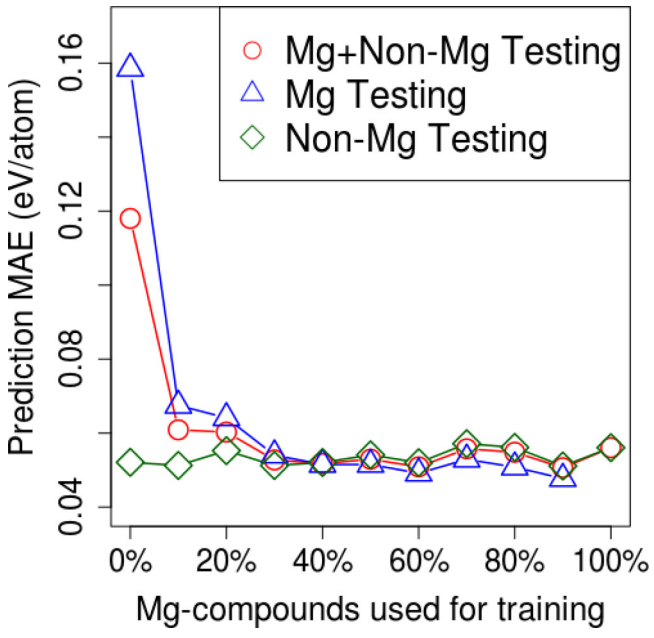

FIG. 4. Performance of deep neural network using the input of compositional representation varied with different percentages of magnesium-containing compound samples used for training. 0% means all magnesium-containing compounds were used for testing and 100% means they were all used for training. Different lines represent different combinations of compound samples used for testing.

In Fig. 5(a), the MAE evaluated on the testing set was elementally resolved and plotted against the number of compounds containing that specific element in the training set. The results displayed two distinct groups of elements. For elements appearing in low frequencies (less than ∼5000), the performance was significantly worse than those appearing than larger frequencies. For some elements such as carbon and nitrogen, the element resolved MAEs were about 2–3 times to the average MAE. This result proved the importance of adequate sampling to correctly interpret the complex pattern using DNN. We note that the critical boundary of the frequency of appearance that separated elements with high and low MAE coincided with the above-mentioned results in Fig. 4, ∼3000 to 5000.

One plausible argument is that the presence of a critical frequency of appearance is attributed to the compositional representation that projects the information of a compound in the discrete compositional space without any explicit relation between different elemental dimensions. Assuming the training data have zero compounds containing a specific element, for example, magnesium, it would be impossible for the model to learn anything about magnesium from other elements presented in the training data. This issue may be alleviated with the chemical descriptors, which project the compound in a space with pre-defined dimensions. In this way, the model is pre-educated with the chemical knowledge and the inference in a domain with less populated examples can still be made by interpolating from other domains. To examine this hypothesis, we analyzed the DNN model using chemical descriptors as input and the result is presented in Fig. 5(b). The likeness between Figs. 5(a) and 5(b) is immediately noticed. Figure 5(c) compared the MAEs from these two charts in detail. For the elements with less appearance frequency, both models agreed with better accuracies for predicting the formation energies of selenium and sulfur compounds and scored the worst for compounds with carbon or nitrogen. One exception was xenon compounds, for which the model using chemical descriptors performed notably worse, which was attributed to the less accurate chemical knowledge about this noble gas element to form a compound. In general, these results overthrew the expectation that chemical descriptors bring advantages to infer the behavior of a compound by leveraging pre-educated chemical

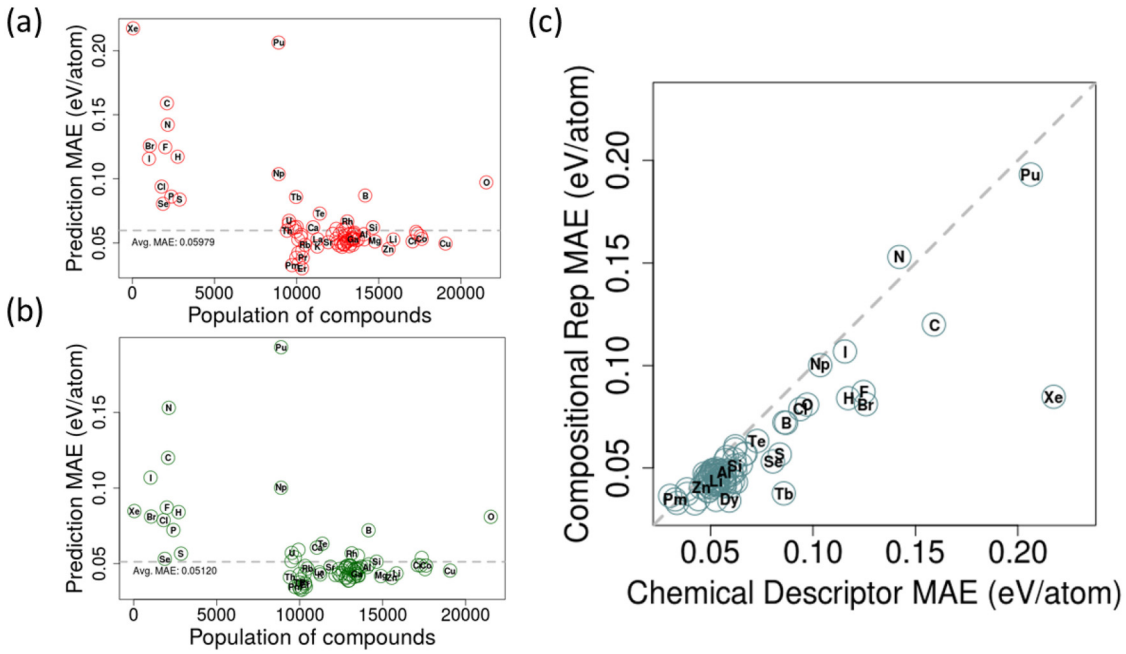

FIG. 5. Effect of training size on the model performance. (a) Accuracy to predict the formation energy for compounds containing a specific element using the hand-crafted chemical descriptors. (b) Same as (a) but using the input of compositional representation. (c) Direct comparison of the MAEs from (a) and (b). In these, overlapped elements are not labeled for a better reading experience.

knowledge. Combining these results, we may safely draw the conclusion that deep learning is able to abstract the compositional information of a compound and utilizes it for the task of predicting the formation energy at least at the same level as, if not significantly exceeding the manually crafted descriptors.

D. Other deep learning algorithms

As stated in the Introduction, the broad concept of deep learning includes many different architectures designed for various types of problems. The analysis presented in Secs. III A–III C relied on one of the simplest architectures, the deep neural network. In our final analysis, we extended the study to other DL architectures of the convolutional neural network and recurrent neural network. Convolutional neural network (CNN) is one of the most extensively used architectures in deep learning, especially for problems with data in a grid-like format. In the RNN architecture, the information is sequentially processed, and connections are built up between historical and currently processing data. This method has been extensively used in machine translation, speech recognition, and time series prediction area.[43,44] The illustrations of the DNN, CNN, and RNN architectures are shown in Fig. 6. We chose to compare DNN, CNN, and RNN due to their similar input structure as well as means to transform the input of a one-hot vector to a new representation of chemical composition. We note that recent studies using other advanced deep learning algorithms are utilizing different representation and model architectures such as graphic neural network have achieved the state-of-art level of accuracy for the prediction of formation energies.[12,15,45]

The performance of DNN, CNN, and RNN in predicting the formation energies were evaluated and compared on OQMD datasets, and the results are presented in Table III. In both tasks, the RNN model improved the capability of the DNN model by more than 10%, while the CNN model had the worst performance among the three models. The same phenomena were also observed in the OQMD1 dataset, where substantial improvement was achieved with the RNN model and the CNN decreased the performance by ~20%. In both DNN and CNN architectures, each individual layer only interacts with its adjacent layers to abstract the input to lower dimensions until the final prediction is made. In the RNN architecture, the model leveraged the weights sharing characteristic that enables the model to capture important information of elements in a sequence regardless of their actual locations. The attention weight further added a mechanism to abstract the information from the full sequence. These results hence suggested the positive role of these two aspects in the RNN architecture in capturing the trend in the formation energy data.

IV. DISCUSSION

Our current study primarily focused on the predicting performance of deep learning models in predicting the formation energies of inorganic compounds. Particularly, we discussed the influence of data source, descriptors, and algorithms on this

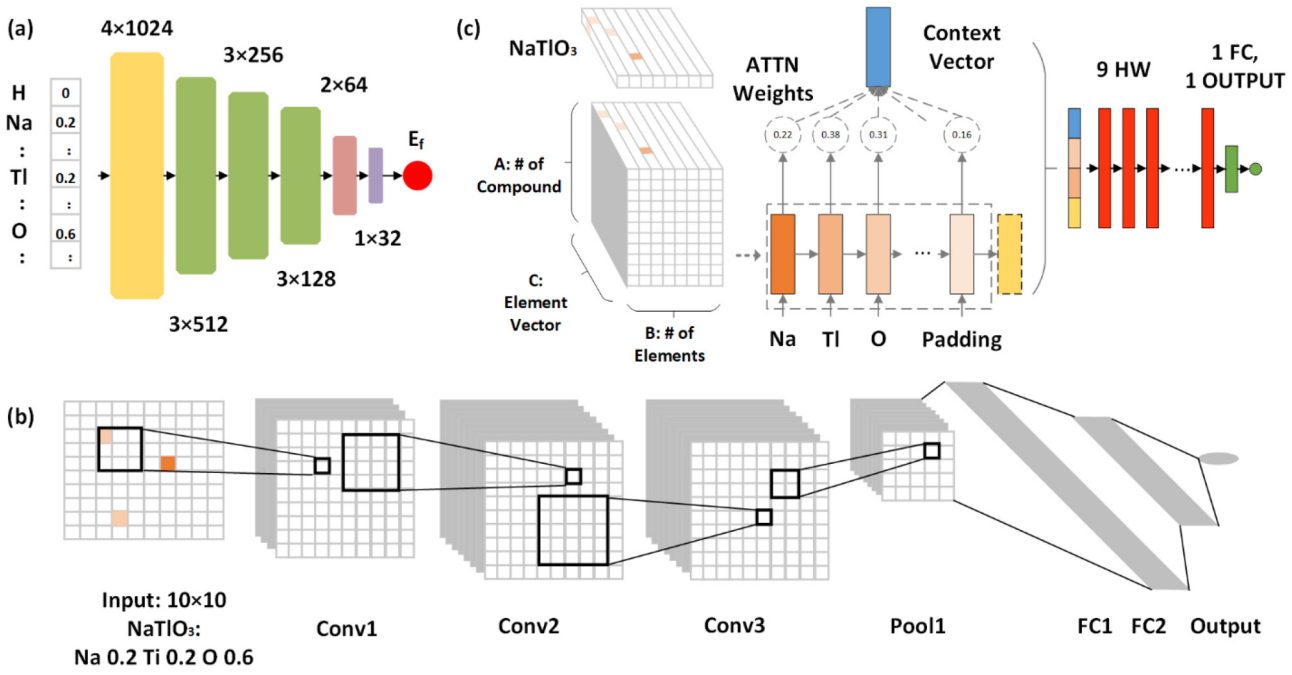

FIG. 6. Architecture illustrations of (a) deep neural network, (b) convolutional neural network, and (c) recurrent neural network.

specific task. We should emphasize that other aspects are also of great importance for the machine learning models. For instance, we have not discussed the time and computational resources for model training, which are important especially for handling a large volume of data. Another aspect is the model interpretability, especially for the deep learning model, which is often criticized as a black box lacking the transparency of interpretability. Nonetheless, the results from the current study provide instructive information about learning a fundamental materials' property using deep learning methods, which we hope could serve as a guideline for future research on similar topics.

To summarize our key findings, we compared the performance of different models in Fig. 7. We used the DNN model equipped with chemical descriptors and tested on the OQMD 2 dataset as the baseline for comparison. Several conclusions can be drawn from this comparison:

(1) The current work showed the advantage of deep learning over several non-deep learning algorithms in predicting the formation energies. Linearized algorithms performed extremely poor on this task, suggesting the difficulty to handle the complex, perhaps also highly non-linear, patterns embodied in the properties of the material. Non-linear algorithms such as random forest performed significantly better, although compared to the baseline DNN model, the MAE was still ∼10% worse.

TABLE III. MAEs (eV/atom) of DNN, CNN, and RNN on predicting the formation energies.

	DNN	CNN	RNN
OQMD1	0.047	0.057	0.042
OQMD2	0.051	0.061	0.045

(2) The comparison between models trained with MP, OQMD1, and OQMD2 datasets highlighted the importance of data sources in evaluating the performance of any model. Even for the same property, the model performance was strongly affected by the different sizes, sampling, and distribution of data used for training. The difference between the models using OQMD1 and OQMD2 datasets was at the same level as that between the DNN and RF models. It is, therefore, of great importance not to use the models reported in the literature for the conclusive performance comparison unless the comparison is carefully performed using rigorously the same dataset. It is, therefore, our opinion that future publications should be encouraged to openly share the data or deposit it into public domains to allow the full transparency that benefits the advances in this field.

(3) The importance of data amount for deep learning has been highlighted in several experiments in the current work. In comparison of MP and OQMD2 models, the MP model predicted much poorer for positive formation energy compounds due to the lack of examples in this special range. In comparison of DNN and RF, we showed that the DNN was only

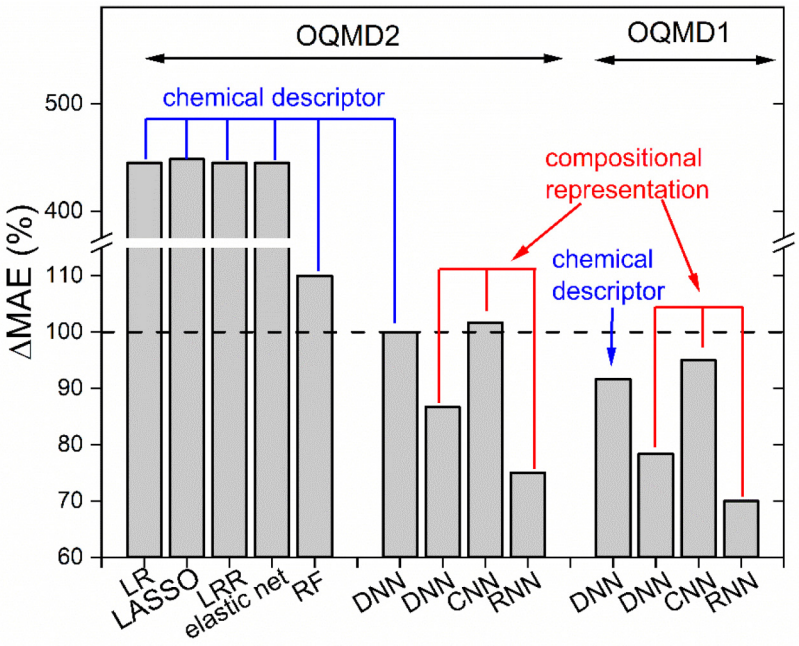

FIG. 7. Mean absolute errors for different models of formation energy.

noticeably better when the data exceeded a certain amount. In the training of DNN with compositional descriptors, the abstracting of elemental information was achieved when the data for a specific element reached a certain level. Finally, we showed that the predicting error for compounds containing a specific element was tightly associated with the frequency of the appearance of that element in the training set. While the data amount is certainly critical for any machine learning problem,[46] the complex architecture of deep learning model is likely to be more prone to the lack of training data. This aspect of deep learning should be explored for future studies in materials science.

(4) Our results demonstrated the capability of deep learning to abstract chemical knowledge from materials properties. Note that this capability is unique to deep learning methods. Non-deep learning algorithms such as random forest performed poorly if only the compositional representation were served. Compared to the chemical descriptor that was constructed based on the empirical chemical abstraction of elements, the DNN was better to utilize the compositional representation and improved the prediction by 5%–14%.

(5) In the three deep learning algorithms considered in the current work, CNN had a MAE ~20% higher than that of DNN, while RNN showed a MAE ~10% less. The worse performance of CNN may be attributed to the lack of spatial features used in training the model, which made it difficult to take the advantage of CNN to deal with interactions within each spatial local environment and long-term interactions transfer toward global environment. On the other side, the better performance of RNN indicated the importance of including inter-component interactions to correctly infer materials' properties. Materials' properties often display complex interplay among various degrees of freedom. We thus recommend algorithms that better model these types of interactions such as RNN and graphic neural network for the future practice of machine learning in materials research.[15,45]

Formation energy is one of the most fundamental properties of a chemical compound and, perhaps, the simplest metric to assess its thermodynamic stability. A more stringent measurement of thermodynamic stability is the energy cost or gain against the decomposition into stable ground states in a canonical ensemble, which is usually termed as the convex hull energy. A recent work by Bartel *et al.* showed that the prediction of the convex hull energies is a more challenging task than the prediction of formation energies.[47] While the challenge of predicting the stability of a chemical substance will certainly be addressed in future studies, the same guidelines should be followed to practice deep learning for this problem as well as other problems in materials science.

V. CONCLUSIONS

In this paper, we took the modeling of formation energies of inorganic compounds and systematically studied the performance of deep learning in this representative problem. Our results showed the advantage of deep learning methods over several non-deep learning methods in this task and demonstrated the abstraction of knowledge from patterns in the dataset using deep learning, a unique capability compared to non-deep learning methods. In addition, we highlighted several aspects that critically affected the conclusion, including the rigorous comparison using the same dataset, the design of input representation, and the careful selection of deep learning architectures. While the results proved the capability of deep learning methods in solving complicated materials problems, these results serve as new guidelines for future practicing of deep learning in materials research.

ACKNOWLEDGMENTS

The authors thank D. Banerjee from the Toyota Research Institute of North America and Y. Kawamura from the Toyota Motor Corporation for their support.

DATA AVAILABILITY

The raw data of formation energies should be acquired directly from the open quantum database licensed under CC-BY 4.0 and materials project database licensed under a Creative Commons Attribution 4.0 International License. The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1]W. S. McCulloch and W. Pitts, Bull. Math. Biophys. **5**, 115 (1943).
[2]K. Fukushima, Neural Networks **1**, 119 (1988).
[3]D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Nature **323**, 533 (1986).
[4]Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, Neural Comput. **1**, 541 (1989).
[5]F. A. Gers, J. Schmidhuber, and F. Cummins, Neural Comput. **12**, 2451 (2000).
[6]Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Proc. IEEE **86**, 2278 (1998).
[7]I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, 2016).
[8]D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, and M. Lanctot, Nature **529**, 484 (2016).
[9]A. Mayr, G. Klambauer, T. Unterthiner, and S. Hochreiter, Front. Environ. Sci. **3**, 80 (2016).
[10]K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müler, and R. J. Maurer, Nat. Commun. **10**, 5024 (2019).
[11]K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, Nat. Commun. **8**, 13890 (2017).
[12]C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, Chem. Mater. **31**, 3564 (2019).
[13]D. Jha, L. Ward, A. Paul, W. K. K. Liao, A. Choudhary, C. Wolverton, and A. Agrawal, Sci. Rep. **8**, 17593 (2018).
[14]K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, J. Chem. Phys. **148**, 241722 (2018).
[15]T. Xie and J. C. Grossman, Phys. Rev. Lett. **120**, 145301 (2018).
[16]W. Ye, C. Chen, Z. Wang, I.-H. Chu, and S. P. Ong, Nat. Commun. **9**, 3800 (2018).
[17]S. Zeng, Y. Zhao, G. Li, R. Wang, X. Wang, and J. Ni, NPJ Comput. Mater. **5**, 84 (2019).
[18]W.-K. Chen, X.-Y. Liu, W.-H. Fang, P. O. Dral, and G. Cui, J. Phys. Chem. Lett. **9**, 6702 (2018).
[19]L. Huang and C. Ling, ACS Omega **4**, 18510 (2019).
[20]E. Kim, K. Huang, S. Jegelka, and E. Olivetti, NPJ Comput. Mater. **3**, 53 (2017).
[21]E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, and E. Olivetti, Chem. Mater. **29**, 9436 (2017).
[22]V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, and A. Jain, Nature **571**, 95 (2019).
[23]K. Ryan, J. Legyel, and M. Shatruk, J. Am. Chem. Soc. **140**, 10158 (2018).
[24]A. Chowdhury, E. Kautz, B. Yener, and D. Lewis, Comput. Mater. Sci. **123**, 176 (2016).
[25]B. L. DeCost, T. Francis, and E. A. Holm, Acta Mater. **133**, 30 (2017).
[26]H. Xu, R. Liu, A. Choudhary, and W. Chen, J. Mech. Des. **137**, 051403 (2015).
[27]G. B. Goh, N. O. Hodas, and A. Vishnu, J. Comput. Chem. **38**, 1291 (2017).
[28]J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, NPJ Comput. Mater. **5**, 83 (2019).
[29]T. F. G. G. Cova and A. A. C. C. Pais, Front. Chem. **7**, 809 (2019).
[30]A. Mignan and M. Broccardo, Nature **574**, E1 (2019).
[31]A. Mignan and M. Broccardo, in *Advances in Computational Intelligence. IWANN 2019* (Springer, Cham, 2019), Vol. 11506.
[32]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, J. Machine Learning Res. **12**, 2825 (2011).
[33]H. Robbins and S. Monro, Ann. Math. Statist. **22**(3), 400 (1951).
[34]M. D. Zeiler, arXiv:1212.5701 (2012).
[35]D. P. Kingma and J. Ba, arXiv:1412.6980 (2014).
[36]A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, and G. Ceder, APL Mater. **1**, 011002 (2013).
[37]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).
[38]J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, JOM **65**, 1501 (2013).
[39]L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, NPJ Comput. Mater. **2**, 16028 (2016).
[40]L. Breiman, Mach. Learn. **45**, 5 (2001).
[41]C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, BMC Bioinf. **9**, 307 (2008).
[42]G. Larsson, M. Maire, and G. Shakhnarovich, arXiv:1605.07648 (2016).
[43]G. Petneházi, arXiv:1901.00069 (2019).
[44]I. Sutskever, O. Vinyals, and Q. Le, arXiv:1409.3215 (2014).
[45]R. E. A. Goodall and A. A. Lee, arXiv:1910.00617 (2020).
[46]Y. Zhang and C. Ling, NPJ Comput. Mater. Sci. **4**, 25 (2018).
[47]C. J. Bartel, A. Trewartha, Q. Wang, A. Dunn, A. Jain, and G. Ceder, NPJ Comput. Mater. **6**, 1098 (2020).