# Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery

Cheol Woo Park and Chris Wolverton*
*Department of Materials Science and Engineering, Northwestern University, Evanston, Illinois 60208, USA*

(Received 4 February 2020; accepted 11 May 2020; published 1 June 2020)

The recently proposed crystal graph convolutional neural network (CGCNN) offers a highly versatile and accurate machine learning (ML) framework by learning material properties directly from graphlike representations of crystal structures ("crystal graphs"). Here, we develop an improved variant of the CGCNN model (iCGCNN) that outperforms the original by incorporating information of the Voronoi tessellated crystal structure, explicit three-body correlations of neighboring constituent atoms, and an optimized chemical representation of interatomic bonds in the crystal graphs. We demonstrate the accuracy of the improved framework in two distinct illustrations: First, when trained/validated on 180 000/20 000 density functional theory (DFT) calculated thermodynamic stability entries taken from the Open Quantum Materials Database (OQMD) and evaluated on a separate test set of 230 000 entries, iCGCNN achieves a predictive accuracy that is significantly improved, i.e., 20% higher than that of the original CGCNN. Second, when used to assist a high-throughput search for materials in the $ThCr_2Si_2$ structure-type, iCGCNN exhibited a success rate of 31% which is 155 times higher than an undirected high-throughput search and 2.4 times higher than that of the original CGCNN. Using both CGCNN and iCGCNN, we screened 132 600 compounds with elemental decorations of the $ThCr_2Si_2$ prototype crystal structure and identified a total of 97 unique stable compounds by performing 757 DFT calculations, accelerating the computational time of the high-throughput search by a factor of 65. Our results suggest that the iCGCNN can be used to accelerate high-throughput discoveries of new materials by quickly and accurately identifying crystalline compounds with properties of interest.

DOI: 10.1103/PhysRevMaterials.4.063801

## I. INTRODUCTION

Density functional theory (DFT) calculations have proven to be a valuable tool in characterizing materials properties and discovering new materials [1]. However, prediction of novel materials through DFT calculations remains a computationally challenging process due to the sheer size of the materials search space. Recently, with the availability of large material databases [2–7], data-driven materials design and discovery using machine learning (ML) has gained much attention for its potential to predict new materials with favorable properties much faster than DFT calculations with substantially less computational cost. ML models have been developed for various materials applications such as predicting formation energies [8–19], band-gap energies [20–24], melting temperatures [25–27], thermal conductivity [26,28], and mechanical properties of materials [29–31].

A working ML model requires three components: (1) training and testing data, (2) a ML algorithm, and (3) materials representation. Much of the creative efforts in materials informatics have been focused on developing representations that can uniquely define each material and best capture the chemistry that influences the property of interest. Recently, inspired by the breakthroughs made in other fields such as computer vision, there has been a rising effort to take advantage of neural networks to extract useful descriptors from inorganic compounds without having to construct them manually [32–39]. In particular, graph neural networks (GNNs), first used by the quantum chemistry community to extract descriptors from molecular graphs [32–36], have started being used on graph representations of crystal compounds to reach unprecedented accuracy in predicting materials properties and to gain chemical insight [36–38,40].

In this work, we show that frameworks utilizing GNNs can be further improved in predicting material properties. We build upon the recently proposed crystal graph convolution neural network (CGCNN) framework [37] which utilizes graph representations of crystals, referred to as crystal graphs, and present an improved framework (iCGCNN) that better represents the chemical nature of an inorganic compound. In this framework, descriptors extracted from the crystal graphs include the information of the Voronoi tessellated crystal structure, explicit three-body correlations of neighboring constituent atoms, and an optimized chemical representation of interatomic bonds, all of which are absent in the crystal graphs utilized by the original framework. The improvement of the model is illustrated through two distinct tests.

First, we compare the accuracy of CGCNN and iCGCNN in predicting the thermodynamic stability of inorganic materials, using a training/testing dataset of DFT-calculated stabilities from the Open Quantum Materials Database (OQMD) [2,3]. Thermodynamic stability in this work refers to the difference between the formation energy of a compound and the lowest-energy linear combination of phases corresponding to that composition, typically calculated from the so-called

*Corresponding author: c-wolverton@northwestern.edu

 

convex hull constructions (henceforth, "convex hull energy"). We use two different approaches to predict stability. In the first approach, we train ML models to predict the formation energy of phases, which is subsequently used to calculate stability relative to the convex hull energy derived from DFT calculated formation energies. In the second approach, we train and test the ML models directly on the DFT-calculated thermodynamic stability data, bypassing calculations of formation energy.

In the second illustration, we conduct separate ML-assisted high-throughput searches using both CGCNN and iCGCNN to discover stable compounds in the $ThCr_2Si_2$ structure type, one of the most commonly occurring ternary prototype structures. We compare the performances of CGCNN and iCGCNN based on the number compounds that were confirmed to be stable through DFT and the success rate, which we define as the ratio of the number of stable compounds identified to the number of DFT calculations that were performed to identify those stable compounds, in their respective high-throughput search.

In both studies, we find that iCGCNN significantly outperforms the original CGCNN model. In predicting thermodynamic stability, iCGCNN achieves an accuracy that is 20% higher than that of CGCNN. In predicting new stable $ThCr_2Si_2$-type materials, iCGCNN identifies nearly twice as many compounds with a success rate that is greater by a factor of 2.4 than the original CGCNN. Using both CGCNN and iCGCNN, we screened 132 600 $ThCr_2Si_2$-type compounds that were generated by substituting elements into the original $ThCr_2Si_2$ structure and identified 97 of them to be stable by only performing 757 DFT calculations, a success rate that is higher than that of an undirected high-throughput search by a factor of 65. Our findings show iCGCNN to be a highly efficient screening tool to predict potentially stable materials to accelerate the challenging task of materials design and discovery.

## II. RESULTS

### A. Description of improvements in the iCGCNN

The original CGCNN framework [37] utilizes a graph representation of the crystals that is composed of two parts: (1) nodes that represent constituent atoms of the crystal, and (2) edges that represent the bonds between neighboring atoms. A node is embedded with a vector $\boldsymbol{v}_i$ to represent the properties of atom $i$, where we define embedding as the process of mapping a discrete object to a vector of real numbers. Each edge is also embedded with a vector $\boldsymbol{u}_{(i,\,j)_k}$ that contains the distance information between neighboring atoms $i$ and $j$ of the crystal unit cell. In order to account for the periodicity of the crystal, multiple edges between atoms $i$ and $j$, as indexed by $k$, can exist. Each node in the crystal graph is connected to its 12 nearest neighbors. During the training phase, the node vectors are updated iteratively according to a convolution function defined by

$$\boldsymbol{v}_i^{(t+1)} = \boldsymbol{v}_i^{(t)} + \sum_{j,k} \sigma\left(\boldsymbol{z}_{(i,j)_k}^{(t)} \boldsymbol{W}_1^{(t)} + \boldsymbol{b}_1^{(t)}\right) \odot g\left(\boldsymbol{z}_{(i,j)_k}^{(t)} \boldsymbol{W}_2^{(t)} + \boldsymbol{b}_2^{(t)}\right). \quad (1)$$

The terms in the sum operator represent the two-body correlation of an atom with its neighboring atoms. $\boldsymbol{z}_{(i,\,j)_k}^{(t)} = \boldsymbol{v}_i^{(t)} \oplus \boldsymbol{v}_j^{(t)} \oplus \boldsymbol{u}_{(i,\,j)_k}^{(t)}$ is the concatenation of the node and edge vectors. $\odot$ represents a matrix elementwise multiplication while $\sigma$ and $g$ respectively represent a sigmoid function and a nonlinear activation function. $\boldsymbol{W}^{(t)}$ and $\boldsymbol{b}^{(t)}$ represent the weight and bias matrices respectively for the $t$th convolution step. A more detailed explanation of the CGCNN framework can be found in Ref. [37].

CGCNN offers a highly flexible framework to represent different crystal structures and exhibits excellent performances in predicting a variety of material properties. However, the design of the crystal graphs utilized by CGCNN may not be optimal in representing the chemical environment of constituent atoms. We identify three possible drawbacks of the original CGCNN model that we attempt to address in the new iCGCNN model.

The first drawback of CGCNN is that regardless of the crystal structure being represented, each node in every crystal graph is connected to 12 of its nearest neighbors. While the local chemical environment of an atom is determined by all of its neighboring atoms, atoms in the first- and second-nearest-neighbor shell often have the largest impact on the local environment. Depending on the crystal structure, it is possible that the 12 nearest neighbors of an atom may include neighbors beyond the first- and second-nearest-neighbor shell. These neighbors may introduce information that could overshadow the more important information relayed from the nearest neighbors and deter the ML model from learning the most optimal local environment representation of an atom during training. While $\sigma$ in Eq. (1) functions as a learned weight matrix that allows CGCNN to differentiate the weaker bonds from the stronger bonds between an atom and its 12 neighboring atoms, these weights are still determined in a data-driven process which is more prone to error than explicitly dictating which of the neighboring atoms have significant interactions. In our improved model, we attempt to better represent the local environment of crystals by connecting each node to its Voronoi neighbors, as illustrated in Fig. 1. Furthermore, such a construction enables us to use information from the Voronoi tessellation of the crystal [19] in the edge vector embeddings of the crystal graph, in addition to interatomic distance information. Voronoi polyhedral information embedded in the edge vectors in the iCGCNN model includes attributes such as the solid angle, area, and volume of the Voronoi cell subtended for the facet as calculated by the open source library, Pymatgen [41].

The second drawback of the CGCNN is that only pairwise correlations are explicitly encoded into the convolution function. Higher-order, many-body correlations (e.g., three-body) are not explicitly encoded. We note that the original CGCNN implicitly encodes some information about the many-body correlations into the node vectors through multiple iterations of the convolution step. However, much of the information regarding the many-body correlations is inevitably lost as it is not explicitly encoded into convolutional function. To minimize the loss of information, we *explicitly* integrate information of three-body interactions between atoms $i$, $j$, and $l$

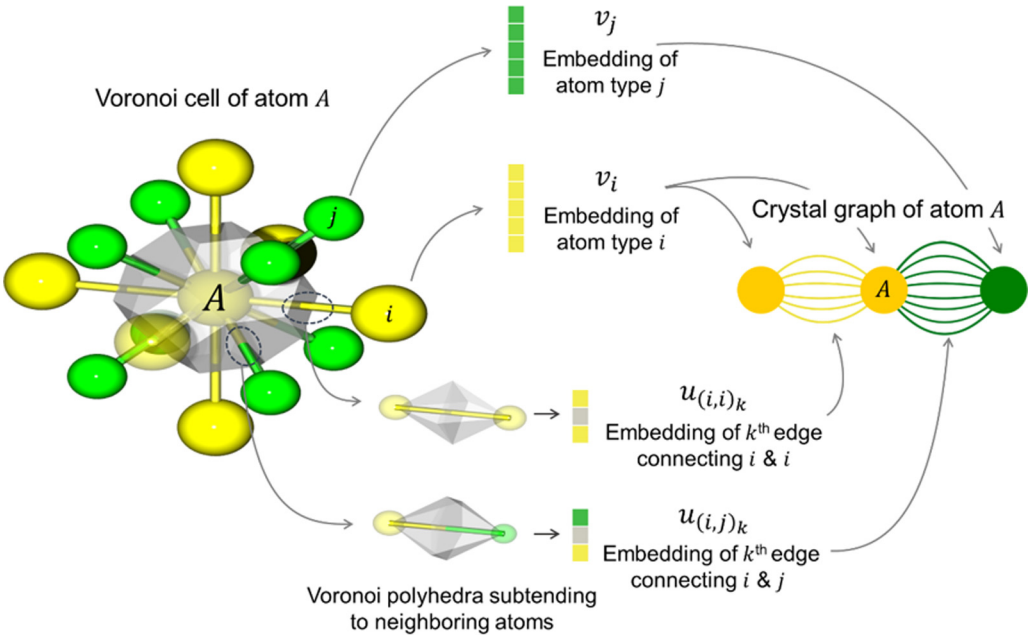


FIG. 1. Illustration of iCGCNN crystal graph. The crystal graph shown on the far right represents the local environment of atom $\boldsymbol{A}$. Multiple edges connect $\boldsymbol{A}$ to neighboring nodes to show the number of Voronoi neighbors. The nodes and edges are embedded with vectors that characterize the constituent atoms ($\boldsymbol{v}_i$, $\boldsymbol{v}_j$) and their correlations with neighboring atoms ($\boldsymbol{u}_{(i,\,i)_k}$, $\boldsymbol{u}_{(i,\,j)_k}$) respectively. Edge vectors include information about the Voronoi polyhedra such as solid angle, area, and volume.

into the convolution function in the iCGCNN model by adding the following term to Eq. (1):

$$\sum_{j,l,k,k'} \sigma\left(\boldsymbol{z}'^{(t)}_{(i,\,j,l)_{k,k'}}\boldsymbol{W}'^{(t)}_1+\boldsymbol{b}'^{(t)}_1\right)\odot g\left(\boldsymbol{z}'^{(t)}_{(i,\,j,l)_{k,k'}}\boldsymbol{W}'^{(t)}_2+\boldsymbol{b}'^{(t)}_2\right). \tag{2}$$

The node vectors and edge vectors that connect atoms $i$, $j$, and $l$ are concatenated to form $\boldsymbol{z}'^{(t)}_{(i,\,j,l)_{k,k'}} = \boldsymbol{v}_i^{(t)} \oplus \boldsymbol{v}_j^{(t)} \oplus \boldsymbol{v}_l^{(t)} \oplus \boldsymbol{u}^{(t)}_{(i,\,j)_k} \oplus \boldsymbol{u}^{(t)}_{(i,l)_{k'}}$. The computational cost of the original CGCNN model scales linearly with the number of atoms in the crystal structure. The time complexity can be written out as $\Theta(mN)$ where $N$ is the number of atoms in a crystal structure and $m$ is the number of edges connected to each node in the crystal graph representation. The time complexity of iCGCNN is $\Theta(m^2N)$ which indicates that the computational cost of iCGCNN is greater than that of CGCNN, but it still scales linearly with respect to the number of atoms even after explicitly calculating the three-body correlations

The third drawback of the CGCNN is that the chemical representations of interatomic bonds, as defined by the edge vectors, are not optimized. In the original CGCNN, node vectors are iteratively optimized to better represent the local chemical environment of an atom, but the edge vectors remain unchanged during training. Thus, in CGCNN, interatomic bonds are not as well represented by the edge vectors as the local chemical environments are represented by the node vectors. To address this issue, we implemented the following convolutional function to update the edge vectors in iCGCNN:

$$\begin{aligned}\boldsymbol{u}^{(t+1)}_{(i,\,j)_k} = \boldsymbol{u}^{(t)}_{(i,j)_k} &+ \sigma\left(\boldsymbol{z}^{(t)}_{(i,j)_k}\boldsymbol{W}^{(t)}_1+\boldsymbol{b}^{(t)}_1\right)\odot g\left(\boldsymbol{z}^{(t)}_{(i,j)_k}\boldsymbol{W}^{(t)}_2+\boldsymbol{b}^{(t)}_2\right)\\ &+\sum_{l,k'}\sigma\left(\boldsymbol{z}'^{(t)}_{(i,\,j,l)_{k,k'}}\boldsymbol{W}'^{(t)}_1+\boldsymbol{b}'^{(t)}_1\right)\\ &\odot g\left(\boldsymbol{z}'^{(t)}_{(i,\,j,l)_{k,k'}}\boldsymbol{W}'^{(t)}_2+\boldsymbol{b}'^{(t)}_2\right).\end{aligned} \tag{3}$$

The terms in this convolutional function closely resemble those in Eqs. (1) and (2). The terms in $\sigma(\boldsymbol{z}^{(t)}_{(i,j)_k}\boldsymbol{W}^{(t)}_1+\boldsymbol{b}^{(t)}_1)\odot g(\boldsymbol{z}^{(t)}_{(i,j)_k}\boldsymbol{W}^{(t)}_2+\boldsymbol{b}^{(t)}_2)$ represent how the chemical properties of the atoms $i$ and $j$ affect their chemical bond. The sum operator term represents how the interatomic bonds are affected by the presence of other nearby atoms.

### B. Predicting thermodynamic stability using CGCNN and iCGCNN

In this section, we compare the predictive accuracies of the original and improved CGCNN model in predicting the thermodynamic stability of crystal compounds. The thermodynamic stability of a compound is determined by what is often referred to as "distance to convex hull" or simply "hull distance." The *convex hull* is defined as the envelope connecting the lowest energy compounds in the chemical space (e.g., binary Li-O space, ternary Cu-Mn-Al space). For example, in the binary Li-O chemical space, the convex hull is simply the envelope that connects the stable Li, $Li_2O$, $Li_2O_2$, $LiO_2$, $LiO_3$, and $O_2$ phases. The hull distance of a compound $i$, $\Delta H^i_{\text{stab}}$ is given by $\Delta H^i_{\text{stab}} = \Delta H^i_f - E^i_{\text{hull}}$, where $\Delta H^i_f$ is the formation energy of compound $i$ and $E^i_{\text{hull}}$ is the convex hull (*constructed without including* $i$) energy at the composition

of $i$. Using the previous example, $\Delta H^{\mathrm{Li_2O}}_{\mathrm{stab}}$ is calculated by constructing a convex hull in the Li-O space after explicitly excluding $Li_2O$. It thus follows, as defined here, that all stable compounds have a hull distance $\Delta H^{i}_{\mathrm{stab}} \leqslant 0$ meV/atom. Note that our definition of hull distance is different from one of the common ways of defining it where $E^{i}_{\mathrm{hull}}$ is the energy of the convex hull constructed including $i$. Under such a definition, stable compounds have a hull distance of zero.

In this study, we considered two strategies to determine the hull distances, using a combination of ML and DFT energies. In strategy 1, we train the ML models on DFT-calculated formation energies of the compounds in the training dataset. Hull distances of the compounds in the test dataset are evaluated by taking the differences between the ML-predicted formation energies and DFT-calculated convex hull energies as in the OQMD. This approach requires us to calculate the convex hull energy every time we are predicting the hull distance of a new compound. In strategy 2, we train the ML models directly on the DFT-calculated hull distances. This approach requires us to construct the convex hulls for compounds in the training data, but it allows us to bypass the convex hull construction altogether when predicting the stability of a new compound.

CGCNN and iCGCNN models using strategy 1 were trained and validated on the formation energies of the ∼200 000 compounds in the training and validation data taken from the OQMD. The models were then used to evaluate both the formation energies and hull distances of the ∼230 000 compounds in the test data, where the hull distances are computed by taking the differences between the ML-predicted formation energies and the convex hull energies as calculated in the OQMD at that composition. Model performances utilizing strategy 1 are summarized in Fig. 2. Mean absolute errors (MAEs) for formation energy predictions are 41.3 and 30.5 meV/atom for the CGCNN and iCGCNN respectively [Figs. 2(a) and 2(b)]. In comparison, Ward *et al.* report a MAE of 80 meV/atom in cross validation for the Voronoi tessellation model [19] when trained and tested on 435 000 formation energies taken from the OQMD. This shows that both CGCNN and iCGCNN offer highly accurate estimations of DFT-calculated formation energies compared to previously developed ML models, consistent to the results reported in Ref. [37]. In another comparison, DFT is widely considered to be a reliable method in estimating various material properties where, for many cases, the differences between DFT and experimental results are trivial. For measuring formation energies, the difference between the DFT and experiments is around 100 meV/atom [3]. This implies that both CGCNN and iCGCNN can be used as a reliable method to estimate DFT-calculated material properties. MAE for hull distances are 41 and 30 meV/atom for CGCNN and iCGCNN respectively, identical to the formation energy prediction errors as expected since the hull distances are calculated directly from the formation energies [Figs. 2(c) and 2(d)]. Overall, iCGCNN hull distance prediction errors are lower than CGCNN by 25% and 15% in terms of MAE and root-mean-square error (RMSE) respectively.

A common trend observed in Figs. 2(a) and 2(b) is the position of the outliers indicated by the red arrows. Both CGCNN and iCGCNN have prediction errors greater than 3 eV/atom for the formation energies of the same 11 compounds. These outliers in the ML predictions were found to consist of compounds with failed DFT calculations that occasionally occur in automated, high-throughput calculations due to the "hands-off" nature of the automation. All 11 outliers with their corresponding DFT errors are listed in Table S3 of the Supplemental Material [42]. For 9 out of the 11 outliers, the indication of DFT error shows up as a large energy difference (several eV/atom) between the relaxation calculation and the final, static calculation, further implying that the DFT-calculated formation energies of these outliers may significantly deviate from the ground truth. Thus, it is possible that when we use highly accurate ML models such as CGCNN or iCGCNN to predict the formation energies of these compounds, the ML predictions are closer to the true formation energy values than the DFT calculations, resulting in large differences between the ML-predicted and DFT-calculated formation energies. While it is concerning that such errors can exist in a high-throughput database, we perceive that an analysis of ML-predicted vs DFT-calculated energies can provide us with the means to quickly locate and ultimately correct these errors.

While it is important that a ML model does not predict a highly unstable compound to be stable, the ability to correctly predict new stable materials is often more closely related to how accurately a ML model can predict the hull distances of compounds that are stable or nearly stable, i.e., compounds that have hull distances less than ∼50 meV/atom. To evaluate the ML models in this aspect, hull distance predictions of stable/nearly stable compounds are shown separately in Figs. 2(e) and 2(f). For this latter set of compounds, iCGCNN hull distance prediction errors are lower by 33% and 28% in terms of MAE and RMSE respectively compared to those of CGCNN. For CGCNN and iCGCNN respectively, MAE measured on the nearly stable compounds are 28% and 16% higher compared to when they are measured on the entire test dataset. The worse predictive accuracy for the stable/nearly stable compounds is likely because there are far fewer stable compounds than there are unstable compounds in the training data (e.g., out of the training set of ∼200 000 compounds only 5.1% are stable), making it more difficult to learn the formation energies/hull distances of the stable/nearly stable compounds.

CGCNN and iCGCNN models using strategy 2 were trained directly on the hull distances of the training entries as queried from the OQMD, and then used to predict the hull distances of the testing entries without having to construct the convex hulls. Model performances for strategy 2 are summarized in Fig. 3. The overall MAEs for the original and improved models are 48.9 and 38.7 meV/atom respectively [Figs. 3(a) and 3(b)]. We find that the improved model outperforms the original model by close to 20%. For nearly stable compounds, MAEs for the original and improved models are 61.6 and 50.1 meV/atom respectively [Figs. 3(c) and 3(d)]. The difference between the MAE values for stable/nearly stable compounds is close to 20%, consistent with the gap between the overall MAE values. Again, as in strategy 1, we observe that MAE measured on the stable/nearly stable compounds are higher compared to when they are measured on the entire test dataset by 26% and 29% for CGCNN and iCGCNN respectively.

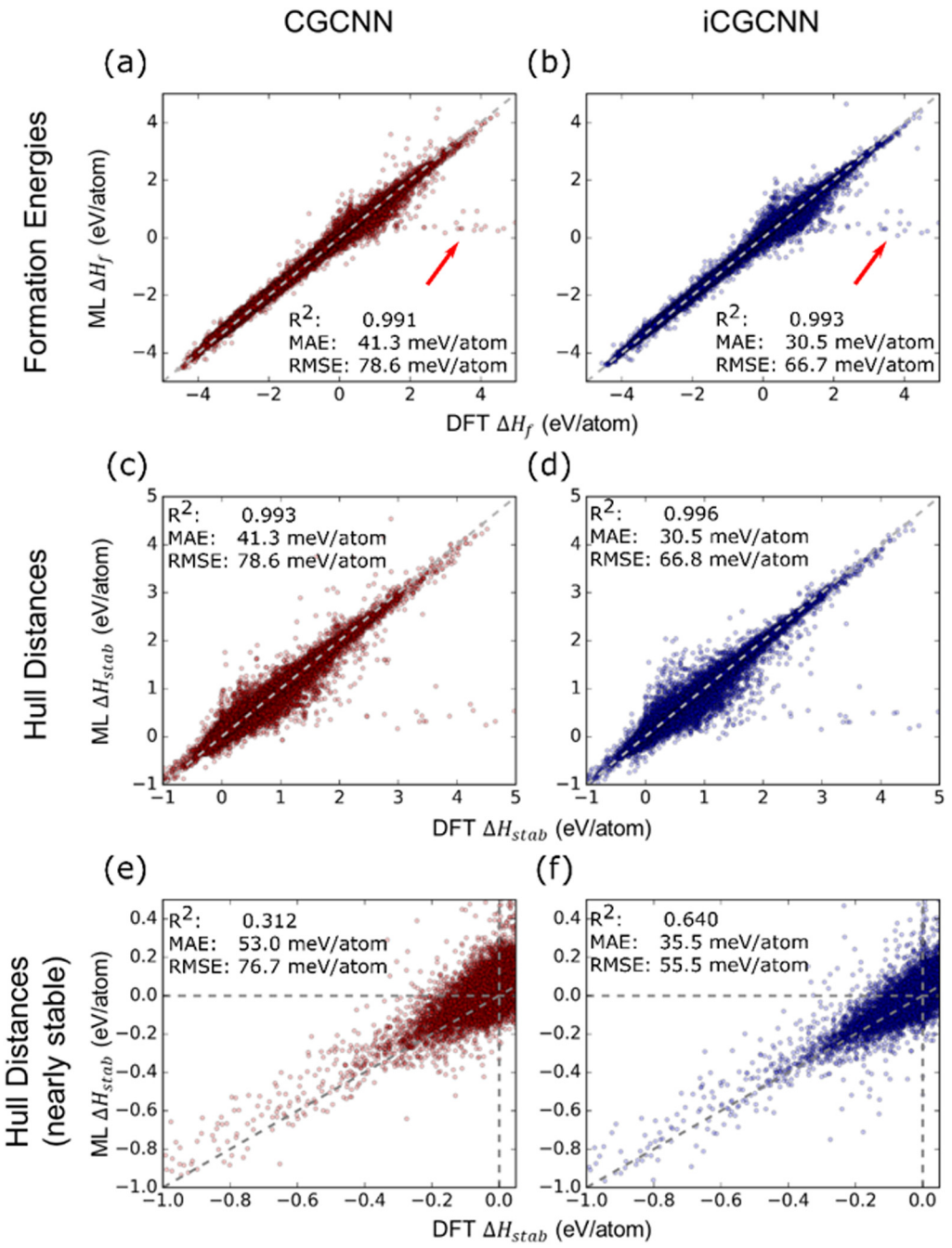


FIG. 2. Predictive accuracies of CGCNN and iCGCNN using strategy 1. Hull distances are evaluated by taking the differences between ML-predicted formation energies and DFT-calculated convex hull energies. (a), (b) DFT vs ML formation energies for (a) CGCNN and (b) iCGCNN. Red arrows indicate the common outliers in CGCNN and iCGCNN predictions. (c), (d) DFT vs ML predicted hull distances for (c) CGCNN and (d) iCGCNN. Close up on compounds with hull distances smaller than 50 meV/atom for (c) and (d) are shown in (e) and (f) respectively.

Comparing strategies 1 and 2, we find that hull distance prediction errors in general are higher for strategy 2. For stable/nearly stable compounds, the MAE resulting from strategy 1 was almost 40% and 30% lower for the original and improved CGCNN models respectively. Furthermore, for strategy 2, the DFT-calculated and ML-predicted hull distances have a negative coefficient of determination ($R^2$) as shown in Figs. 3(c) and 3(d), implying the lack of a linear correlation between the calculated and predicted stabilities. We speculate that there are two causes for this result: First, the hull distance is inherently a more complicated property to learn compared to formation energy. As opposed to formation energy which is calculated with respect to the elemental reference states, hull distance is calculated with respect to the ground-state phases. The number of ground-state phases that are required to be learned for strategy 2 is much larger

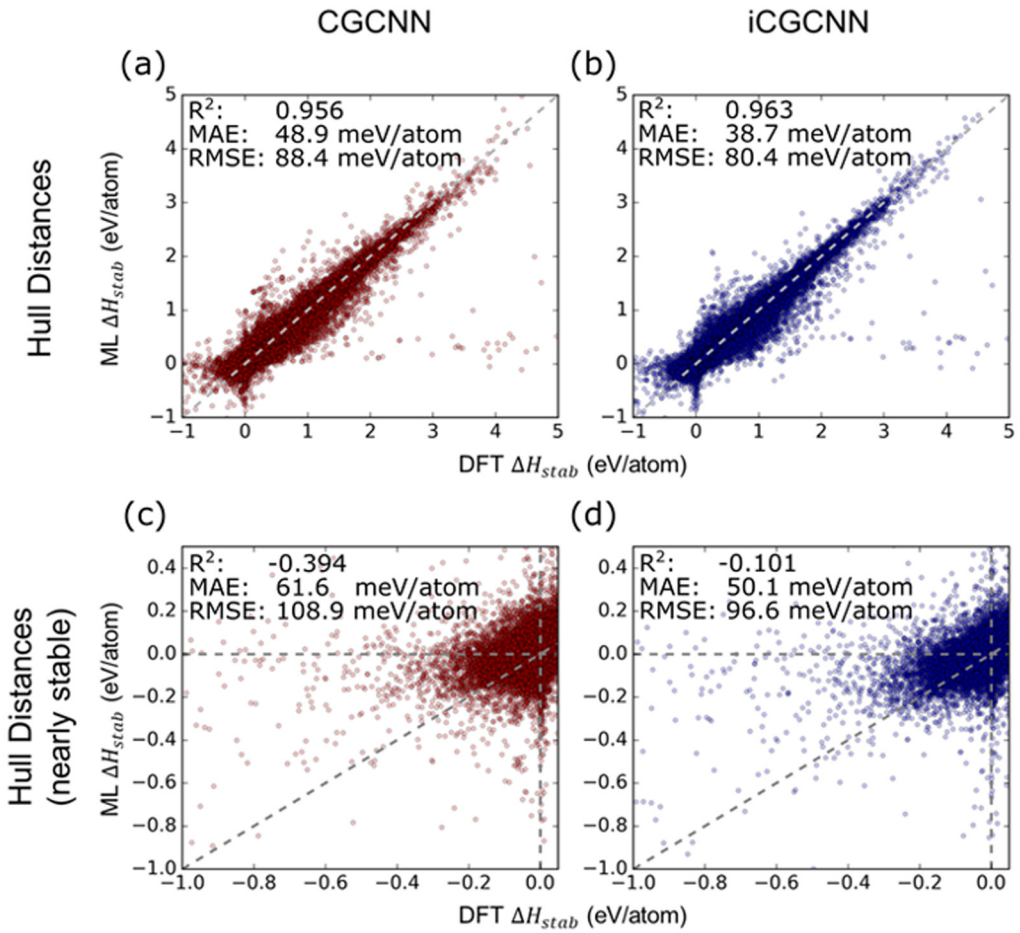


FIG. 3. Predictive accuracies of CGCNN and iCGCNN using strategy 2. ML models are trained to directly predict the hull distances. (a), (b) DFT vs ML predicted hull distances for (a) CGCNN and (b) iCGCNN. Closeup on compounds with hull distances smaller than 50 meV/atom for (a) and (b) are shown in (c) and (d) respectively.

than the number of elemental reference states that needs to be learned for strategy 1. For example, as opposed to only having to learn the elemental reference states of Li and O to predict the formation energy of $Li_2O$, a ML model must learn about the ground state Li, $Li_2O_2$, $LiO_2$, $LiO_3$, and $O_2$ phases to directly predict the hull distance. Second, convex hulls in the OQMD, or any high-throughput materials database that is currently available, are very much incomplete as there are still potentially thousands of compounds that have not been discovered yet. Thus, for strategy 2, it is very likely that the training data include incorrect hull distance values derived from incomplete convex hulls. For example, suppose there is a compound with the composition of $Li_4O$ in the binary Li-O space that is stable but is yet to be discovered. If such hypothetical compound does exist, it means that our knowledge of the convex hull in the Li-O space so far has been incomplete, and the convex hull distances that we calculate for compounds with compositions such as $Li_3O$ in the OQMD are incorrect because they are based on a convex hull construction that excludes the stable $Li_4O$ phase. Such incorrect information in the training data may hinder the learning process of the ML models.

### C. Using CGCNN and iCGCNN to accelerate high-throughput DFT searches: Discovering new stable $ThCr_2Si_2$-type materials

The $ThCr_2Si_2$ structure-type, illustrated in Fig. 4(a), is one of the most commonly observed crystal structures in nature. It is the fifth most common crystal structure among ternary intermetallics [43], accounting for 289 of the 13 026 ternary compounds in Pearson's Crystal Data [44]. The crystal structure has often been identified with materials that exhibit interesting properties such as superconductivity and valence fluctuation [45]. The number of possible compositions for the $ThCr_2Si_2$ structure is about 500 000, and at the time of this study, there were 538 stable $ThCr_2Si_2$-type compounds in the OQMD. This implies that if we conducted an undirected high-throughput search for these compounds, we would identify approximately one or two new stable compounds for every 1000 DFT calculations. For a high-throughput DFT search, we could define a success rate as the ratio of number of stable compounds identified to the number of DFT calculations that were performed to identify those stable compounds. The success rate of an undirected high-throughput search for $ThCr_2Si_2$-type materials would then be around 0.2%. In this section, we conduct a ML-assisted high-throughput search for

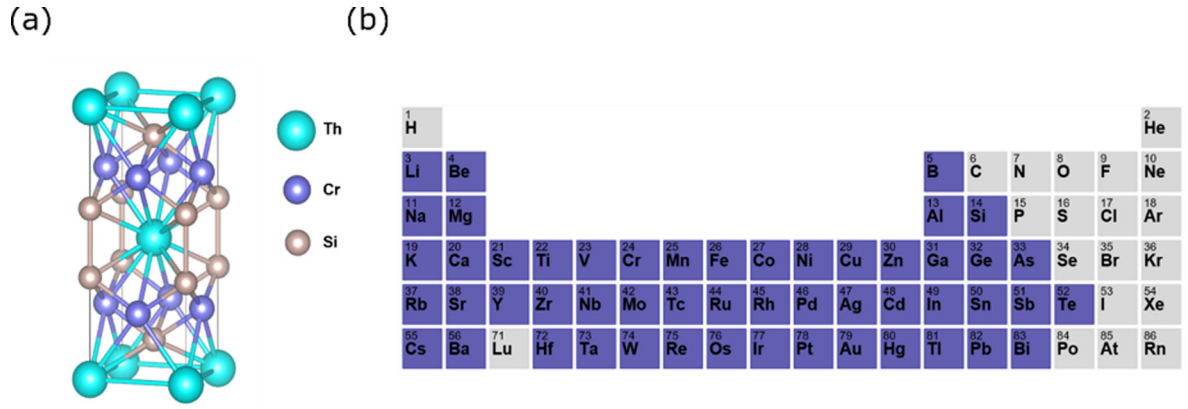


FIG. 4. (a) Structure of $ThCr_2Si_2$. (b) Periodic table where colored elements were substituted into $ThCr_2Si_2$ structure to generate new compounds.

materials in the $ThCr_2Si_2$ structure-type by using the original and improved CGCNN models in parallel with strategies 1 and 2 to improve the success rate.

First, new prototype compounds were generated by substituting elements into the original $ThCr_2Si_2$ structure. Only metallic elements that are not rare earth, totaling 52 elements, were considered for substitution [Fig. 4(b)], resulting in 132 600 ($52\times51\times50$) variations. The 538 stable $ThCr_2Si_2$-type compounds that already exist in the OQMD were then removed to avoid discovering duplicates. CGCNN and iCGCNN models that have been trained for the comparative study in the previous section were then used to predict and screen for potentially stable compounds among the newly generated prototype compounds. For the compounds that were predicted to be stable by the ML models, DFT was used to calculate the formation energies and subsequently the hull distances to validate their thermodynamic stability. The convex hulls that were used to assess the hull distances and confirm the DFT stability of the crystal structures identified by the ML models were constructed from the DFT-calculated phase data available through the OQMD. All DFT calculations were performed within the OQMD framework. Finally, we evaluate the performance of each model based on the number of newly discovered compounds and success rate of their respective high-throughput search.

The results of the search are summarized in Table I. Combining the compounds predicted to be stable from both strategies 1 and 2 without overlap, the original CGCNN predicted a total of 556 unique compounds to be stable, out of which 72 were confirmed to stable through DFT. Out of the 423 unique compounds predicted to be stable by iCGCNN, 133 were confirmed to be stable through DFT, nearly twice the number of stable compounds identified by CGCNN. The 31% success rate of iCGCNN is a factor of 2.4 greater than the 13% success rate of the original CGCNN and a factor of 155 greater than the 0.2% success rate of an undirected high-throughput search. Further investigation of the compounds that were not stable as calculated by DFT revealed that the number of nearly stable compounds found by iCGCNN was 82% and 41% higher than that of CGCNN for strategies 1 and 2 respectively. The average hull distances of the compounds predicted by iCGCNN that were DFT unstable were also lower by 3.9% and 57.7% for strategies 1 and 2 respectively. Both results show that iCGCNN predicted compounds are consistently closer to the convex hull than the compounds predicted by CGCNN. Overall, iCGCNN is far more accurate and efficient in discovering new stable compounds than CGCNN.

We note that for iCGCNN, strategy 2 has identified more stable compounds with a marginally higher success rate than strategy 1, which seemingly contradicts the results of the previous section where strategy 1 had a higher performance than strategy 2. We speculate that there are various factors that led to this contradiction. The first factor is bias in the ML training data. The data that were used to train CGCNN and iCGCNN contain 227 $ThCr_2Si_2$-type compounds, all of which have hull distance values below 0. Thus, when the ML models are trained to directly predict the stability as done in strategy 2, the models are more likely to predict the hull

TABLE I. Performance breakdown of CGCNN and iCGCNN using strategies 1 and 2 in predicting stable compounds out of the 132 600 new prototype $ThCr_2Si_2$-type materials.

| | | No. of compounds predicted to be stable by ML | No. of compounds validated to be stable by DFT[a] | Success rate % | Avg. hull distance of DFT unstable compounds (meV/atom) |
|---|---|---|---|---|---|
| CGCNN | strategy 1 | 134 | 35 (29) | 26 | 103.3 |
| | strategy 2 | 443 | 52 (57) | 12 | 179.6 |
| iCGCNN | strategy 1 | 197 | 69 (55) | 35 | 98.3 |
| | strategy 2 | 260 | 93 (82) | 36 | 82.3 |

[a]Numbers in parentheses indicate the number of compounds that are not stable but are nearly stable as calculated by DFT.

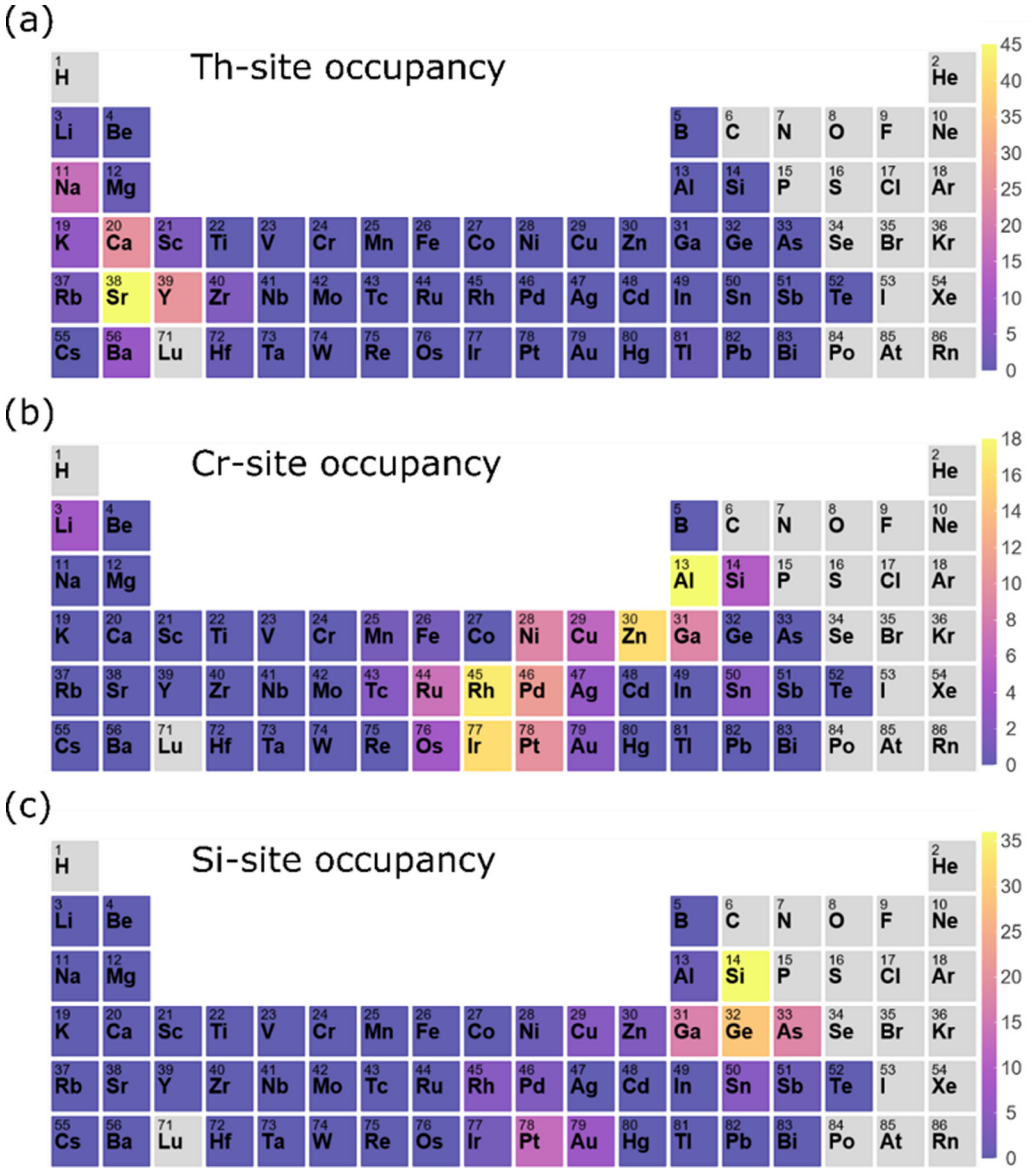


FIG. 5. Periodic tables where elements are color-coded based on the number of stable occurrences on the (a) Th site, (b) Cr site, and (c) Si site.

distances of the prototype $ThCr_2Si_2$-type compounds to be below 0 as well. This bias is reflected in the fact that for both CGCNN and iCGCNN, the number of compounds predicted to be stable by strategy 2 is significantly higher than that of strategy 1 (more than threefold difference for CGCNN). With more compounds predicted to be stable, it is more likely that strategy 2 will identify more stable compounds than strategy 1, even with a lower hull distance predictive accuracy. The slightly higher success rate of strategy 2 over strategy 1 for iCGCNN seems to be the combined result of iCGCNN's improved performance over CGCNN and statistical noise.

Through this survey, we performed a total of 757 DFT calculations and identified 143 stable unique compounds, of which 97 compounds have not yet been reported in the literature to the best of our knowledge. Among the 97 compounds, 42 are significantly stable with hull distances less than −50 meV/atom. While all 97 compounds identified in this survey are computational predictions that await experimental validation, compounds with the significantly negative convex hull distances are most likely to be synthesizable, and hence should be prioritized in any experimental synthesis effort. All 97 compounds are listed by the convex hull distance in Table S4 of the Supplemental Material [42]. The number of stable occurrences for each element on the Th, Cr, and Si site are shown in Fig. 5. The Th site, which has the highest coordination among the three sites, is mostly occupied by alkaline elements that generally have large radii. The Cr site and the Si site are mostly occupied by transition metals and metalloids respectively. Finally, we emphasize that we have discovered 97 potentially new stable compounds by only performing 757 DFT calculations, a success rate of 13% that implies that we have accelerated the high-throughput search for $ThCr_2Si_2$-type materials by a factor 65 using both CGCNN and iCGCNN. Compared to some of the previous works that utilize ML to accelerate materials discovery through high-throughput DFT studies, CGCNN and iCGCNN boast

state-of-the-art performances. For example, Faber *et al.* predicted 2133 elpasolites to be stable and confirmed 128 of them were DFT stable and Kim *et al.* performed 909 DFT calculations and identified 55 stable quaternary Heusler compounds, indicating that their ML models both had success rates of 6% [16,46]. In another benchmark, Schmidt *et al.* showed that their ML model can accelerate high-throughput DFT calculations by at least a factor of 5 [47]. While it is difficult to make an apples-to-apples comparison, our results imply that CGCNN and iCGCNN are more effective in identifying new stable compounds than the previous ML models. However, we note that CGCNN and iCGCNN can be limited by their dependency on the atomic coordinates which we further discuss in Sec. III of the Supplemental Material [42].

## III. METHODS

### A. Data

We use DFT-calculated thermodynamic data from the OQMD for training, validating, and testing ML models throughout this work. OQMD v1.1 contains about ∼450 000 DFT calculations of unique ordered inorganic compounds, including ∼40 000 experimentally known ones from the Inorganic Crystal Structure Database (ICSD) [48,49], and the rest hypothetical ones generated from commonly occurring structural prototypes. All DFT calculations in the OQMD are performed with the Vienna *Ab Initio* Simulation Package (VASP) [50,51]. The details of the methodology and settings used for the high-throughput calculations are explained in Ref. [3]. All ML models in this work are trained on a set of ∼180 000 compounds and validated on another ∼20 000 compounds, all randomly selected from the OQMD with no overlap. Models are tested on a separate set of ∼230 000 compounds that are not included in the training or validation data [2,3].

*Code implementation*. The iCGCNN was implemented based on the CGCNN code available through github [52].

## IV. CONCLUSIONS

The CGCNN model provides a highly accurate and flexible ML framework in which material descriptors are adaptively extracted according to the task at hand and thus allows us to bypass the painstaking process of handcrafting the material descriptors ourselves. We have presented an improved CGCNN model (iCGCNN) to demonstrate that this framework can be further improved. This was done by integrating the Voronoi tessellation information of the crystal structure, explicitly encoding the three-body correlations of neighboring constituent atoms, and optimizing the chemical representation of interatomic bonds in the crystal graphs. We trained and tested the original and improved CGCNN models on OQMD data to compare their hull distance predictive accuracy using two approaches: (1) predic the formation energy and subsequently calculating the hull distance relative to OQMD constructed convex hull and (2) bypass the need to construct the convex hull by directly predicting the hull distance. In both approaches, there were significant gaps between the predictive accuracies, where the iCGCNN performed 25% and 20% better than CGCNN for the former and latter approach respectively in terms of the MAE measured on the entire testing data. Finally, when used to predict new stable compounds with $ThCr_2Si_2$-structure, iCGCNN not only identified twice as many more stable compounds than the original CGCNN, it exhibited a success rate that was greater by a factor 2.4. Beyond comparing the two ML models, we discovered 97 stable compounds in a high-throughput search that was accelerated by a factor of 65 using both CGCNN and iCGCNN. Its excellent performances in screening stable compounds suggests that iCGCNN can be used to greatly accelerate materials discovery.

## ACKNOWLEDGMENTS

This work was performed under the following financial assistance: Award No. 70NANB14H012 from US Department of Commerce, National Institute of Standards and Technology as part of the Center for Hierarchical Materials Design (CHiMaD). The authors also acknowledge the financial support of Toyota Research Institute (TRI). This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation Grant No. ACI-1548562. This work was also supported in part through the computational resources and staff contributions provided for the Quest high performance computing facility at Northwestern University which is jointly supported by the Office of the Provost, the Office for Research, and Northwestern University Information Technology. The authors thank V. Hegde, E. Isaacs, M. Liu, S. Griesemer, A. Gopakumar, and K. Pal for helpful discussion.

[1] 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).

[2] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, The open quantum materials database (OQMD): Assessing the accuracy of DFT formation energies, npj Computat. Mater. **1**, 15010 (2015).

[3] 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 (2013).

[4] 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, Commentary: The materials project: A materials genome approach to accelerating materials innovation, APL Mater. **1**, 011002 (2013).

[5] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. Hart, S. Sanvito, M. Buongiorno-Nardelli, and N. Mingo, AFLOWLIB. ORG: A distributed materials properties repository from high-throughput ab initio calculations, Comput. Mater. Sci. **58**, 227 (2012).

[6] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway, and A. Aspuru-Guzik, The Harvard clean energy project: Large-scale computational screening and design of organic photovoltaics on the world community grid, J. Phys. Chem. Lett. **2**, 2241 (2011).

[7] NOMAD, https://nomad-coe.eu.

[8] S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, and G. Ceder, Predicting Crystal Structures with Data Mining of Quantum Calculations, Phys. Rev. Lett. **91**, 135503 (2003).

[9] E. O. Pyzer-Knapp, G. N. Simm, and A. A. Guzik, A Bayesian approach to calibrating high-throughput virtual screening results and application to organic photovoltaic materials, Mater. Horiz. **3**, 226 (2016).

[10] T. Moot, O. Isayev, R. W. Call, S. M. McCullough, M. Zemaitis, R. Lopez, J. F. Cahoon, and A. Tropsha, Material informatics driven design and experimental validation of lead titanate as an aqueous solar photocathode, Mater. Discov. **6**, 9 (2016).

[11] H. Wu, A. Lorenson, B. Anderson, L. Witteman, H. Wu, B. Meredig, and D. Morgan, Robust FCC solute diffusion predictions from ab-initio machine learning methods, Comput. Mater. Sci. **134**, 160 (2017).

[12] A. O. Oliynyk, E. Antono, T. D. Sparks, L. Ghadbeigi, M. W. Gaultois, B. Meredig, and A. Mar, High-throughput machine-learning-driven synthesis of full-Heusler compounds, Chem. Mater. **28**, 7324 (2016).

[13] G. Pilania, P. V. Balachandran, C. Kim, and T. Lookman, Finding new perovskite halides via machine learning, Front. Mater. **3**, 19 (2016).

[14] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. 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).

[15] 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 (2015).

[16] F. A. Faber, A. Lindmaa, O. A. Von Lilienfeld, and R. Armiento, Machine Learning Energies of 2 Million Elpasolite (A B C 2 D 6) Crystals, Phys. Rev. Lett. **117**, 135502 (2016).

[17] A. M. Deml, R. O'Hayre, C. Wolverton, and V. Stevanović, Predicting density functional theory total energies and enthalpies of formation of metal-nonmetal compounds by linear regression, Phys. Rev. B **93**, 085142 (2016).

[18] 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).

[19] 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).

[20] 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).

[21] J. Lee, A. Seko, K. Shitara, K. Nakayama, and I. Tanaka, Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques, Phys. Rev. B **93**, 115104 (2016).

[22] 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).

[23] G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis, and T. Lookman, Machine learning bandgaps of double perovskites, Sci. Rep. **6**, 19375 (2016).

[24] P. Dey, J. Bible, S. Datta, S. Broderick, J. Jasinski, M. Sunkara, M. Menon, and K. Rajan, Informatics-aided bandgap engineering for solar materials, Comput. Mater. Sci. **83**, 185 (2014).

[25] A. Seko, T. Maekawa, K. Tsuda, and I. Tanaka, Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single-and binary-component solids, Phys. Rev. B **89**, 054303 (2014).

[26] A. Seko, H. Hayashi, K. Nakayama, A. Takahashi, and I. Tanaka, Representation of compounds for machine-learning prediction of physical properties, Phys. Rev. B **95**, 144110 (2017).

[27] G. Pilania, J. E. Gubernatis, and T. Lookman, Structure classification and melting temperature prediction in octet AB solids via machine learning, Phys. Rev. B **91**, 214302 (2015).

[28] 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).

[29] C. S. Kong, S. R. Broderick, T. E. Jones, C. Loyola, M. E. Eberhart, and K. Rajan, Mining for elastic constants of intermetallics from the charge density landscape, Physica B: Condens. Matter **458**, 1 (2015).

[30] M. De Jong, W. Chen, R. Notestine, K. Persson, G. Ceder, A. Jain, M. Asta, and A. Gamst, A statistical learning framework for materials science: Application to elastic moduli of k-nary inorganic polycrystalline compounds, Sci. Rep. **6**, 34256 (2016).

[31] A. O. Furmanchuk, A. Agrawal, and A. Choudhary, Predictive analytics for crystalline materials: Bulk modulus, RSC Adv. **6**, 95246 (2016).

[32] 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 28*, edited by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Curran Associates, Inc., Red Hook, NY, 2015), pp. 2224–2232.

[33] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley, Molecular graph convolutions: Moving beyond fingerprints, J. Comput.-Aided Mol. Des. **30**, 595 (2016).

[34] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, Neural message passing for quantum chemistry, in *Proceedings of the 34th International Conference on Machine Learning*, ICML'17, edited by D. Precup and Y. W. Teh (JMLR.org, Sydney, NSW, Australia, 2017), Vol. 70, pp. 1263–1272.

[35] 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).

[36] 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).

[37] 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).

[38] T. Xie and J. C. Grossman, Hierarchical visualization of materials space with graph convolutional neural networks, J. Chem. Phys. **149**, 174111 (2018).

[39] D. Jha, L. Ward, A. Paul, W. K. Liao, A. Choudhary, C. Wolverton, and A. Agrawal, Elemnet: Deep learning the chemistry of materials from only elemental composition, Sci. Rep. **8**, 17593 (2018).

[40] C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, Graph networks as a universal machine learning framework for molecules and crystals, Chem. Mater. **31.9**, 3564 (2019).

[41] 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 (2013).

[42] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevMaterials.4.063801 for further details, which includes Refs. [16,47].

[43] J. Dshemuchadse and W. Steurer, More statistics on intermetallic compounds–ternary phases, Acta Cryst. A **71**, 335 (2015).

[44] P. Villars and K. Cenzual, *Pearson's Crystal Data: Crystal Structure Database for Inorganic Compounds (on Digital Versatile Disc)* (ASM International, Materials Park, OH, 2013).

[45] R. Hoffmann and C. Zheng, Making and breaking bonds in the solid state: The thorium chromium silicide ($ThCr_2Si_2$) structure, J. Phys. Chem. **89**, 4175 (1985).

[46] K. Kim, L. Ward, J. He, A. Krishna, A. Agrawal, and C. Wolverton, Machine-learning-accelerated high-throughput materials screening: Discovery of novel quaternary Heusler compounds, Phys. Rev. Mater. **2**, 123801 (2018).

[47] 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**(12), 5090 (2017).

[48] G. Bergerhoff, R. Hundt, R. Sievers, and I. D. Brown, The inorganic crystal structure data base, J. Chem. Inf. Comput. Sci. **23**, 66 (1983).

[49] A. Belsky, M. Hellenbrandt, V. L. Karen, and P. Luksch, New developments in the Inorganic Crystal Structure Database (ICSD): Accessibility in support of materials research and design, Acta Cryst. B **58**, 364 (2002).

[50] G. Kresse and J. Hafner, *Ab initio* molecular dynamics for liquid metals, Phys. Rev. B **47**, 558 (1993).

[51] G. Kresse and D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Phys. Rev. B **59**, 1758 (1999).

[52] https://github.com/txie-93/cgcnn.