Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations

Logan Ward,[1,*] Ruoqian Liu,[2] Amar Krishna,[2] Vinay I. Hegde,[1] Ankit Agrawal,[2] Alok Choudhary,[2] and Chris Wolverton[1]
[1]*Department of Materials Science and Engineering, Northwestern University, Evanston, Illinois 60208, USA*
[2]*Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, Illinois 60208, USA*
(Received 20 March 2017; revised manuscript received 5 June 2017; published 14 July 2017)

While high-throughput density functional theory (DFT) has become a prevalent tool for materials discovery, it is limited by the relatively large computational cost. In this paper, we explore using DFT data from high-throughput calculations to create faster, surrogate models with machine learning (ML) that can be used to guide new searches. Our method works by using decision tree models to map DFT-calculated formation enthalpies to a set of attributes consisting of two distinct types: (i) composition-dependent attributes of elemental properties (as have been used in previous ML models of DFT formation energies), combined with (ii) attributes derived from the Voronoi tessellation of the compound's crystal structure. The ML models created using this method have half the cross-validation error and similar training and evaluation speeds to models created with the Coulomb matrix and partial radial distribution function methods. For a dataset of 435 000 formation energies taken from the Open Quantum Materials Database (OQMD), our model achieves a mean absolute error of 80 meV/atom in cross validation, which is lower than the approximate error between DFT-computed and experimentally measured formation enthalpies and below 15% of the mean absolute deviation of the training set. We also demonstrate that our method can accurately estimate the formation energy of materials outside of the training set and be used to identify materials with especially large formation enthalpies. We propose that our models can be used to accelerate the discovery of new materials by identifying the most promising materials to study with DFT at little additional computational cost.

DOI: 10.1103/PhysRevB.96.024104

I. INTRODUCTION

Especially in the past decade, high-throughput atomistic calculation methods have proven to be powerful tools for discovering new materials [1–6]. These methods generally work by employing an accurate computational tool, often density functional theory (DFT), to predict the properties of large numbers of experimentally observed and hypothetical inorganic compounds created by substituting different elements into known crystal structure types. The results of these predictions are often stored in publicly accessible databases [1,5–11], which makes it possible for many researchers to quickly search for materials that warrant further investigation (e.g., via more accurate and expensive computational methods or via experimental synthesis). This strategy of combinatorial replacement and high-throughput calculations has already enabled the discovery of new materials for a host of applications, including Li-ion batteries, thermoelectrics, water splitting materials, and structural alloys [2,12–21].

While combinatorial searches are evidently useful, they are intrinsically limited by available computational power. Evaluating only the zero temperature, ground state properties of a material using DFT can require hours of processor time per compound. Consequently, the space of possible combinations is too large to evaluate every candidate for some types of compounds. For example, the combinations of every element in a quaternary crystal structure results in at least two million possible compound compositions (more if there are inequivalent sites in the crystal), which outstrips the capability of today's computational resources. For more complex properties (e.g., elastic constants, vibrational properties, defects), evaluating two million compounds is certainly impractical. At some point, it is necessary to selectively evaluate only the parts of the search space that are likely to contain promising candidates.

Machine learning (ML) offers a route for creating fast surrogate models from databases and has proven to be a viable route for estimating the results of DFT calculations [22–37]. These approaches have recently been reviewed in Ref. [38]. One of the first studies in this area, by Curtarolo *et al.* in 2003, built a ML model that predicts the formation enthalpy of binary compound based on the formation enthalpies of the same pair of elements in several other structures [22]. While the model was successfully used to identify previously undiscovered intermetallics [22], it is limited to making predictions to compounds whose structural type appears frequently in available datasets. Later work demonstrated methods for creating ML models from inputs derived from the composition of each training entry [23,27,28], which allow for greater flexibility in using the model but require expensive crystal-structure prediction algorithms to determine the structure of the material when validating the predictions (e.g., using DFT of the predicted compositions) [23]. There has also been work showing how to predict some computationally expensive properties, including elastic constants [32–34], thermal conductivity [24,25], and melting temperature [39], more quickly by using the results from faster DFT calculations as input into a ML model. Additionally, several studies have predicted new materials with a desired crystal structure by training a model on a dataset of compounds with the same stoichiometry or same crystal structure and using that to identify materials that are likely to be stable in

*Present address: Computation Institute, University of Chicago, Chicago, Illinois 60637, USA.

a much larger set [29,40,41]. However, to fully leverage the amount of information available in high-throughput databases to discover new materials, one needs a reliable and fast method for predicting properties given any crystal structure—such a method remains elusive.

Several different strategies for building ML models based on the crystal structure of a material have already been proposed. These methods are composed of two main components: (1) a numerical representation that describes each compound's crystal structure and composition and (2) a choice of ML algorithm. These methods include work by Faber *et al.* [30], Schütt *et al.* [42], and Seko *et al.* [43] that each constructed kernel ridge regression (KRR) models using several different representations. Faber *et al.* trained a ML model on 3938 entries taken from the Materials Project with a Coulomb-matrix (CM)-based representation and achieved a 370 meV/atom mean absolute error (MAE) in cross validation [30]. Schütt *et al.* constructed a ML model to predict the density of states at the Fermi level with a representation based on the partial radial distribution function (PRDF) and showed that it could be used to predict this quantity for crystal structures outside of the original training set [42]. More recently, Seko *et al.* created a model for cohesive energy using a representation based on four different kinds of structural descriptors and observed a root mean squared error (RMSE) of 41 meV/atom in cross validation using a dataset of 18 903 entries consisting only of compounds based on a select set of 49 different structures and 35 elements [43]. These methods are quite promising; the best cross-validation accuracies reported to date are comparable to or lower than various estimates for the error between DFT and experiment for formation enthalpies (50–100 meV/atom [4,44]) and reaction energies of oxides (∼25 meV/atom [45]). However, such exceptional accuracy has yet to be demonstrated on datasets that include as diverse a range of structures and compositions as those in modern DFT databases. Additionally, as we will demonstrate in this paper, these existing methods are impractical to use with the datasets as large as those currently available. Overall, while promising, there is a need for improvements in methods that can link crystal structure and properties with ML.

In this paper, we demonstrate an approach for predicting properties of crystalline compounds using a representation consisting of attributes derived from the Voronoi tessellation of its structure that is both twice as accurate as existing methods and can scale to large training set sizes. Additionally, we designed our representations to be insensitive to changes in the volume of a crystal, which makes it possible to predict the properties of the crystal without needing to compute the DFT-relaxed geometry as input into the model. In this paper, we use a large dataset from the Open Quantum Materials Database (OQMD) to benchmark this new method against existing representations used in the literature (the CM and PRDF methods) using cross validation. Then, to understand the limitations of our approach, we employ cross validation to assess whether the new structural descriptors impact the accuracy of our ML models and to determine which types of compounds yield the highest error rates. Finally, we validate the ability of our model to make predictions of the formation enthalpy of materials outside our currently available training data and to identify materials with strongly negative formation enthalpies given only the structure prototype and composition but not the DFT-relaxed equilibrium geometry and lattice parameters. We envision that this model can be used to screen potential materials based on stability before more expensive calculation techniques are used and, thereby, enable faster high-throughput searches for new materials.

II. METHODOLOGY—CONSTRUCTING THE ML MODEL

Our approach is composed of two distinct steps: (1) representing a compound's composition and crystal structure as a set of quantitative attributes and (2) using ML to extract patterns that relate those attributes to the property of interest. We describe both steps in this section, along with the resource used to provide training data for these models.

A. Training data

All training data for the ML models created in this paper were extracted from the OQMD [4,5]. At the time that the data used here were collected, the OQMD contained the results of DFT calculations for 435 792 unique compounds (i.e., unique combinations of composition and crystal structure) all performed with the Vienna *Ab Initio* Simulation Package (VASP) [46,47]. We employed the crystal-structure matching tools in qmpy to ensure that each entry in the dataset is unique [48]. Detailed settings for VASP used in the OQMD are described in Ref. [4]. The OQMD contains over 30 k entries corresponding to entries from the Inorganic Crystal Structure Database (ICSD) [49], and the remainder are predominantly hypothetical structures created by replacing elements in known crystal structures with different elements. As described in later subsections, we use several unique subsets of this database, which include using only the entries from the ICSD. All datasets are available in the Supplemental Material [50] for this paper.

B. Representation of crystalline compounds—crystal structure and composition

The representation of a crystalline compound is designed to transform the composition and crystal structure of the compound into a list of quantitative attributes that serve as input into a ML model. Following previous discussions of the desired features of representations for materials [29,42,51–54], we also assert that representations for crystalline compounds should be quick to compute and capture all relevant features of a composition+structure in a compact list of attributes. Additionally, we suggest several other desirable features specific to building representations for crystal structures. First, these attributes should also be insensitive to the choice of a unit cell (i.e., primitive cells, conventional cells, and supercells of the same structure should all have the same representation). Additionally, as our goal in using these models is to estimate the stability of a crystal structure before employing DFT, we also assert that representation should fulfill two other requirements to be predictive. For one, the representation should not rely on knowledge of the DFT-relaxed lattice parameters and internal degrees of freedom and at least be invariant to changes due to simple dilation or contraction of the lattice. Also, the representation should be designed such that small changes in the structure (e.g., perturbations in atomic

position) do not result in unphysical, discontinuous changes in attributes.

Considering all of these constraints, we created a representation for crystalline compounds based on the Voronoi tessellation of the structure [55]. The Voronoi tessellation of a crystal partitions space into the so-called Wigner-Seitz cells of each atom, which encompass the region closer to that atom than any other atom [56]. This tessellation is uniquely defined for a crystal structure and is insensitive to the choice of unit cell (e.g., primitive or conventional). The faces of a Voronoi polyhedron correspond to the nearest neighbors of an atom, which provides an unambiguous way of describing its local environment. To create attributes, we compute many characteristics of the local environment of each atom (described below) and then measure statistics about the distribution of these characteristics across all atoms in the unit cell. These attributes are designed in such a way that they are unaffected by unit cell selection or by changing the volume of the unit cell. Our attributes are dependent on changes in the ratios between lattice parameters (e.g., *c*/*a* for tetragonal structures) and internal degrees of freedom. However, as we will demonstrate later, the effect of changes in these parameters upon relaxation on the output of a ML model is often minor. Furthermore, we also weigh the contribution of each neighboring atom to each attribute according to the area of its corresponding face on the Voronoi cell. In this way, the attributes are stable against discontinuities caused by addition or removal of facets in the tessellation caused by small deformations in the structure, as shown in Fig. S1 in the Supplemental Material [50].

We use the Voronoi tessellation and composition of the structure to create several different categories of attributes. In these descriptions, n is an index to a face of a single cell in the tessellation. Each cell corresponds to the volume around a single atom, and each face of the cell corresponds to a specific nearest neighbor to that atom. To generate attributes, we consider both properties of the face (e.g., area), which are not dependent on composition, and the identities of the neighboring atom, which are affected by composition.

(1) Effective coordination number based on the mean, maximum, minimum, and mean absolute deviation in the effective coordination number of each atom, which is computed using the equation,

$$\mathrm{CN}_{\mathrm{eff}} = \frac{\left(\sum_n A_n\right)^2}{\sum_n A_n^2}, \tag{1}$$

where A_n is the surface area to face n and the sum Σ_n is over all faces of the Voronoi cell. This formula reverts to the number of faces on the cell for cells with equally sized faces (e.g., 12 for fcc) and leads to smaller coordination numbers for structures with unequal faces (e.g., 11.96 rather than 14 for bcc).

(2) Structural heterogeneity attributes that measure the variation in local environments around each atom, including statistics regarding the mean bond length about each atom, the variation in bond length between each neighbor of an atom, and variation in the volume between each Voronoi cell. To make these attributes insensitive to volume changes, the bond lengths are normalized by the mean bond length of all atoms and the cell volumes are normalized by the mean cell volume.

(3) Chemical ordering attributes that are computed using Warren-Cowley-like ordering parameters [57] of the first, second, and third neighbor shells, weighted according to face sizes of each neighboring atom. We define the ordering parameter to be specific to each type of atom in the structure. For the first shell, the ordering parameter is defined as

$$\alpha(t) = 1 - \frac{\sum_n A_n \delta(t - t_n)}{x_t \sum_n A_n}, \tag{2}$$

where $\alpha(t,s)$ is the weighted ordering parameter for type, x_t is the atomic fraction of type t in the crystal, t_n is the type of the atom corresponding to face n, and δ is the delta function. To make the number of attributes the same regardless of the number of elements in the crystal and insensitive to unit cell choice, we measure the mean absolute value of ordering parameters for each atom in the lattice for each type in the crystal. Consequently, crystals with ordered arrangements (e.g., rock salt) will have values of these attributes closer to 1, and more random arrangements will be closer to zero.

To compute this ordering attribute for the second and third neighbor shells, we first compute all nonbacktracking paths of length two or three, respectively, through the network defined by the atoms whose cells share faces in the tessellation. We then assign each step in each path a weight proportional to the fraction of surface area corresponding to the Voronoi face associated with that step (e.g., a face that takes up 10% of the surface area of a cell has a weight of 10%), and each path is assigned a weight equal to the product the weights of each of its step. The ordering parameter is then computed using a similar formula to Eq. (2),

$$\alpha(t,s) = 1 - \frac{\sum_p w_p \delta(t - t_p)}{x_t}, \tag{3}$$

where s is the index of the shell, Σ_n is the overall s-length path, w_p is the weight of each path, and t_p is the type of the atom at the end of the path. In this way, paths that involve small faces have a small contribution to the ordering attribute, which ensures that it is stable against small deformations. Full details of this calculation are available in the Supplemental Material [50].

(4) Maximum packing efficiency, which can be computed by finding the largest sphere that fits inside each Voronoi cell. For example, the maximum packing efficiency for fcc is 0.74 by this definition.

(5) Local environment attributes that are computed by comparing the elemental properties of the element of each atom to those of its nearest neighbors using the relationship

$$\delta_p = \frac{\sum_n A_n |p_n - p_i|}{\sum_n A_n}, \tag{4}$$

where p_n and p_i are values of an elemental property (e.g., electronegativity) of the atom corresponding to face n and central atom, respectively. For this paper, we compute the mean, mean absolute deviation, maximum, minimum, and range of this value for all atoms in a structure for 22 different elemental properties (e.g., atomic number), which are listed in Table S1 in the Supplemental Material [50]. For example, each atom in rock salt NaCl is surrounded by only atoms of the opposite type. The absolute difference between the

electronegativity of each atom and its neighbors is therefore 2.23 (the difference between Na and Cl), and the mean across the entire structure is also 2.23. As all atoms have the same value for this property, the range and mean absolute deviation are both zero.

(6) Composition-based attributes based on the fractions of each element are present in the structure. These attributes are described in recent work by Ward *et al.* [27]:

(a) Stoichiometric attributes that depend on the fractions of each element and not what those elements are.

(b) Elemental-property-based attributes that are based on statistics of the elemental properties of all atoms in the crystal.

(c) Electronic structure attributes, which depend on the fraction of electrons in the s, p, d, and f shells of the constituent elements, normalized by the total number of electrons in the system rather than by the element fractions [as in 6(b) in this list]. These are based on work by Meredig *et al.* [23].

(d) Ionicity attributes derived from differences in electronegativity between constituent elements and whether the material can form a charge-balanced ionic compound if all elements have common oxidation states.

Further details about the attributes are described in the Supplemental Material [50]. In total, our method describes each material with 271 attributes. Each of these attributes can be computed using the Materials-Agnostic Platform for Informatics and Exploration (Magpie) and the Versatile Atomic-Scale Structure Analysis Library (Vassal), which are both freely available under open source licenses [58,59]. Example input files and the datasets used in this paper are also included as Supplemental Material [50].

C. The ML technique

For the ML algorithm, we chose to use the random forests (RF) algorithm proposed by Breiman due to its superior performance and robustness against overfitting [60]. The RF algorithm works by aggregating the results of several decision trees, each built from a random subset of training examples and attributes. Each decision tree is composed of a series of decision rules (e.g., packing efficiency >0.5) learned by partitioning data into subsets that minimize intrasubset variation of class values, which are formation enthalpies in this case. This partitioning process is repeated recursively (i.e., on each subset generated by the previous rule), forming a tree where each branch is a different decision rule. The leaves of the tree are each assigned a value of formation enthalpy that maximizes fitness to the training set. In the RF algorithm, this decision tree generation process is repeated several times with a different subset of the training set, and the predictions made from all decisions trees are averaged to predict the class value of new data.

In modeling our problem, we used an ensemble of 50 decision trees for all ML models created based on the ICSD dataset and 100 decisions trees for ML models created with the full OQMD dataset. We also investigated increasing the number of trees as the training data increases, but no notable improvement was observed. Models were constructed using the Scikit-Learn library in Python [61] and the Weka ML library in Java [62].

D. Alternate representations—CM and PRDF

In this paper, we compare our new representation against the CM [30] and PRDF [42] matrix approaches. Both methods use KRR as the base ML algorithm, which performs linear regression where the inputs into the linear model are based on the similarity between a new observation and each entry in the training set. This similarity metric is often designed specifically for each problem, and the CM and PRDF methods primarily vary in the choice of metric used to compare two crystal structures.

The PRDF method expresses the similarity between two structures based on a matrix defined by PRDF [42]. Each row of this matrix corresponds to the radial distribution function between a different pair of elements, and the matrix contains all possible pairs of elements. For instance, one row is the Li-Cl RDF, which describes the frequency of Li and Cl atoms a certain distance apart in the structure. To compute the difference between two structures, one generates this matrix for both structures and computes the Frobenius norm of the difference between the two matrices.

The CM method is based on a representation that was originally developed for molecules [63]. In this representation, one computes a matrix that is related to the Coulomb repulsion between the atomic nuclei in the material

$$C_{ij} = \begin{cases} 0.5Z_i^{2.4} & \text{if } i = j \\ \frac{Z_i Z_j}{r_{ij}} & \text{if } i \neq j \end{cases}, \tag{5}$$

where Z_i is the atomic number of atom i and r_{ij} is the distance between atoms i and j. To compare two structures, one first computes the eigenvalues of the CM for both structures and then subtracts the two lists of eigenvalues (padding with zeros to make them the same length). More recently, Faber *et al.* proposed several modifications to the CM to account for periodic boundary conditions [30]. Of their proposed modifications, we use the sine matrix approximation, which they found to lead to the lowest cross-validation error when predicting formation enthalpy.

For both methods, we optimized the hyperparameters for the KRR learning algorithm and, for the PRDF matrix, the cutoff radius and bin size used for the RDF. In both cases, we used a grid search technique. All parameters were varied to maximize the performance of each model at a training set size of 3000 entries. With this technique, we were able to reproduce the observed cross-validation error of the CM reported in Ref. [30]. Our implementation for both of these methods is available as part of Magpie [58].

III. RESULTS—CHARACTERIZING MODEL PERFORMANCE

In this section, we characterize several different aspects of our new ML technique. First, we benchmark our technique to existing methods by comparing their cross-validation accuracy. Then, we analyze the predictions where our model performs least accurately to determine where this model can be best applied. Finally, we study the effect of structural information in our representation to determine whether the model is learning the effect of structural traits on formation energy.

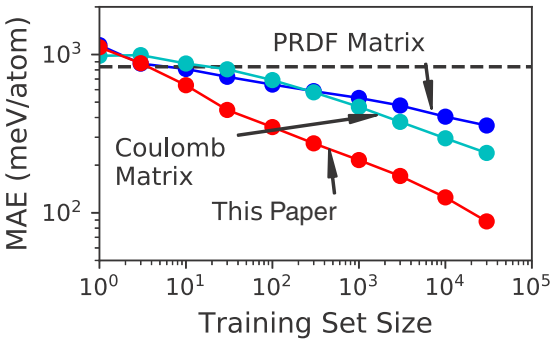

FIG. 1. MAE measured using cross validation of models created using the PRDF [42], CM [30], and the method presented in this paper. Each model was trained on the DFT formation energies of a set of randomly selected compounds from the ICSD and was used to evaluate 1000 distinct compounds that were also selected at random. The black, dashed line indicates the expected error from guessing the mean formation energy of the training set for all structures.

A. Comparison of the Voronoi method to existing techniques

We first use cross validation to study the ability of our technique to model the formation energy of inorganic compounds and compare its performance to existing methods. As a training set, we use the DFT-relaxed structures and formation energies of compounds in the ICSD [49] that are available in the OQMD [4,5]. Our dataset includes 32 111 compounds and represents an unbiased sampling of all known compounds with a primitive cell size smaller than 40 atoms. To assess the effect of increasing training set size, we constructed models using randomly selected training sets with between 1 to 30 000 entries and evaluated the performance of the model on a distinct set of 1000 entries. This test strategy was selected to assess the effect of training set size on model performance. Each cross-validation test was repeated 20 times, and the performance of the model was taken to be the average over all 20 tests.

The comparison of cross-validation error for out Voronoi method with the CM and PRDF models is shown in Fig. 1. We find that the models created using our approach were more accurate than those based on the CM and PRDF methods for all training sets larger than three entries. As shown in Fig. 1, models based on our method have an MAE of 170 meV/atom at a training set size of 3000 entries. In contrast, we find the CM and PRDFs models to have 2.2 times and 2.8 times larger errors, respectively. At a training set size of 30 000 entries, the MAE of our model (88 meV/atom) is still significantly lower than those from the other two methods. Since the error of our models decreases with increasing training set size at a similar rate to those of the CM method and faster than those from the PRDF method, we expect our models to be more accurate even when trained with the largest available DFT formation energy datasets of between 10^5–10^6 compounds [5,7].

Beyond having a lower MAE, models created using our new method also perform better according to more outlier-sensitive performance metrics. We measured the Pearson's correlation coefficient, RMSE, and maximum absolute error for models produced using each method trained identical 30 000-entry training sets and evaluated each model with the same 1000-entry validation set. As shown in Fig. 2, each metric is better for our method than both the PRDF and CM methods. What the better performance according to these metrics suggests is that our method achieves superior accuracy without introducing a larger fraction of outliers.

To determine whether the increased accuracy is a result of the new representation or the use of the RF algorithm, we repeated the comparison between the CM and our Voronoi representations using the same ML algorithm for both. We first test both representations using KRR, and subsequently we test both using RF. For the KRR test, the error for the model using our new representation is significantly higher than when we used the RF algorithm but still lower than the CM+KRR model (see Fig. 3). In contrast, the error rate of models created using our representation is lower than those using the CM by a factor of two when we employed RF as the learning algorithm.

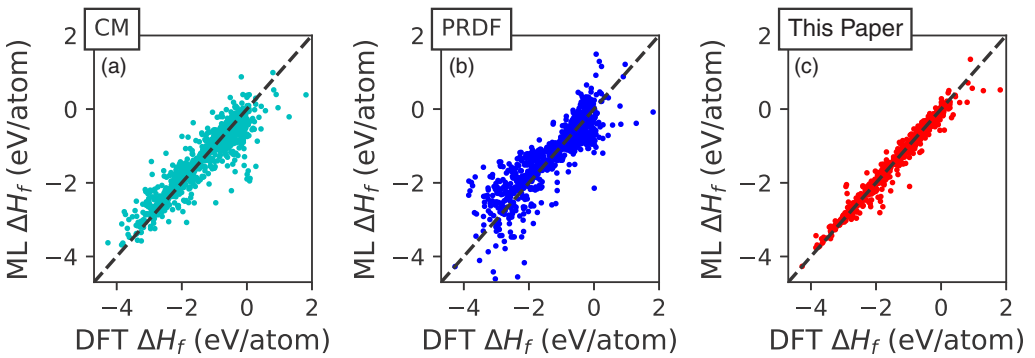

FIG. 2. Formation enthalpy (ΔH_f) computed using DFT and predicted using ML with models created using the (a) CM [30], (b) PRDF [42], and (c) the method proposed in this paper. Each model was trained on the same set of 30 000 entries from the OQMD and evaluated against the same validation set of 1000 compounds. The results from the validation set are shown in this figure along with several different performance metrics.

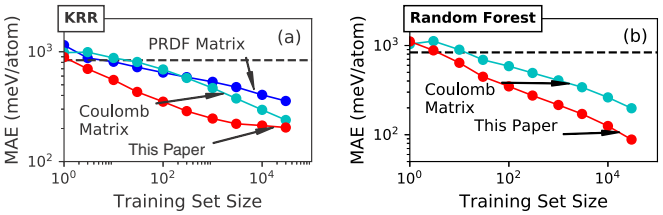

FIG. 3. Performance of ML models for formation enthalpy created with the same ML algorithm but different representations. Each graph shows MAE for (a) KRR model and (b) RF algorithm in a cross-validation test where the model was trained on progressively larger training sets and validated against a separate test set of 1000 entries. For each algorithm, we compare the performance using the Voronoi-tessellation-based representation proposed in this paper against the CM [30] and the PRDF [42] matrix representations.

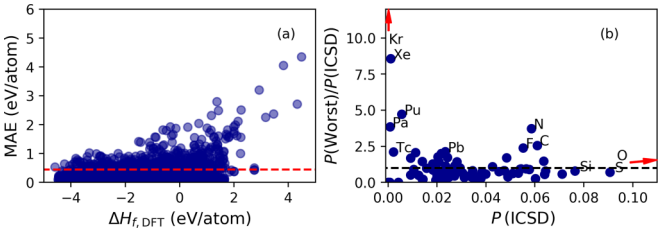

FIG. 5. (a) The DFT-computed formation enthalpy of a compound compared to the MAE between the DFT and machine-learning-predicted formation enthalpy of that compound during a cross-validation test. The red, dashed line indicates the 98th percentile of the MAE. (b) Comparison of the fraction of compounds that contain a certain element in our ICSD training set P(ICSD) to the ratio between the fraction of compounds in the 98th percentile of MAE and the fraction in the training set.

Consequently, we conclude that the improved accuracy of our models is a result of the new representation and not only the choice of the ML algorithm.

Additionally, we find that the training time of our method scales better with increasing training set size and has similar evaluation speed to the PRDF and CM methods. As shown in Fig. 4, as the size of the training set reaches 10 000 and more, the time taken to train and run models created using our method is comparable to the PRDF and CM methods. The training and run time of our model is dominated by the time required to compute the Voronoi tessellation used to generate the attributes, which requires approximately 0.1 s per compound on our test system and accounts for ∼98% of the model training time and >99% of the run time. For our training set sizes, we observe an $O(N)$ scaling (N is the number of compounds in the training set) for training time due to the large, but $O(N)$ is the calculation time for the construction of the representation. The RF ML algorithm scales with $O(N \log N)$, and, hence, we would eventually expect $O(N \log N)$ scaling for large dataset sizes. For small dataset sizes, the time to

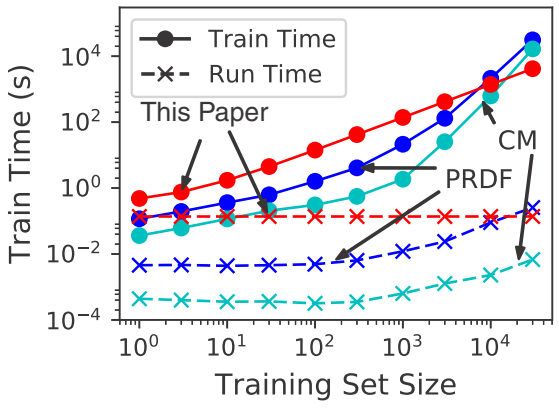

FIG. 4. Comparison of model training and running time of three different techniques to predict the formation energy of inorganic compounds: Coulomb Matrix (CM) [30], Partial Radial Distribution Function (PRDF) [42], and our Voronoi-tessellation-based method. Training time is the sum of attribute generation and model construction with given data. Run time is the average time taken to compute the required attributes and to evaluate the ML model for a single compound.

compute the Voronoi attributes makes it slower to train and run than both competing methods. However, this is not true for large datasets, and we observe parity between the two methods for training set sizes around 10^4. Considering that the training time for the CM and PRDF scale at the faster rate of $O(N^3)$ for KRR, our approach will remain more feasible to train for even larger datasets. For datasets with only 30 000 entries in the training set, our method is faster to train by approximately a factor of 10 and is only slightly slower to run than the CM model—although we find (see Fig. 4) that differences in run speed are also likely to close with increasing training set size.

B. Testing for systematic errors in Voronoi models

To understand where our ML model can be used the most effectively, we ran a cross-validation test and studied the compounds where the model had the highest error rates. In this cross-validation test, we withheld a random selection of 25% of the ICSD dataset used in the previous section for a test set and trained the model on the remaining 75% of the data. We repeated this test 100 times and measured the MAE for each compound over all times it appeared in the test set. Then, we selected the 643 compounds with highest 2% of MAE values (above 446 meV/atom) to determine which compounds are persistently the most difficult to predict accurately. We find that many of these outliers are compounds with positive formation enthalpies [see Fig. 5(a)]. In other words, many of these difficult-to-predict compounds are unstable with respect to decomposition into the elements. The fact that our model performs poorly for very unstable compounds is unsurprising since their formation energies are outliers compared to the rest of the ICSD training set, which are mostly stable.

We also find that compounds containing elements that appear the least frequently in our training set are overrepresented in the compounds with the worst MAEs. Figure 5(b) shows the probability of finding a compound containing a certain element in our entire dataset [P(ICSD)] and the ratio between the probability of finding that element in the entries with the highest MAE [P(Worst)] and the probability of finding it in the entire dataset. Of all elements present in the training set, Kr, Xe, and Pa have the highest overrepresentation (a ratio of 14 for Kr) and are among the least frequently appearing elements

in the original dataset [64]. From these results, we conclude that our model performance is expected to be least predictive for compounds containing elements that appear infrequently in the training data (e.g., Tc, actinides).

The two elements that are both frequently occurring and most overrepresented in our worst-performing materials are C and N. Out of the 643 compounds with the highest error, there are 43 that contain either C or N. This list of 43 C- or N-containing compounds includes many compounds of C or N with rarely observed elements (e.g., ThCN), whose presence in the list can be explained due to the few training examples with the rarely observed elements. Many of the other compounds include C or N covalently bonded with another element, such as materials containing carbonate and nitrate ions. Carbonate ions, for example, are slightly over-represented in the list of compounds with the highest errors, where 0.77% of entries in this list (5/643 compounds) contain carbonate ions compared to only 0.46% compounds in the training set at large. This prevalence of certain classes of materials containing covalent bonds in the worst predictions suggests that our model could be improved by including attributes that capture characteristics such as bond angles or using electron counting rules to characterize the types of bonds present in the structure. Beyond identifying regions to improve this model, our analysis of its failures also identifies where it can be applied with the greatest likelihood of predictive accuracy: compounds with commonly occurring elements (a significant amount of the training data).

C. Assessing the importance of structural information in the ML model

As many of the attributes employed in our representation are not dependent on structure, it is important to determine the impact of the structure-dependent attributes on the accuracy of our ML models. If these structural attributes have a negligible effect, it is possible that the model is only learning from the structurally invariant (i.e., composition-based) attributes. To test the effect of including structure-dependent attributes, we replicated the cross validation described in the previous section and trained a RF algorithm with three sets of attributes: (i) only the composition-based (i.e., structure-independent) attributes, (ii) only the Voronoi-tessellation-based attributes, and (iii) all 271. As a reference, we also include the results of a RF using the CM representation. As shown in Fig. 6(a), there is little difference between the error rate of a model trained using all the attributes and the structure-independent ones. We also find that models created using only the Voronoi-tessellation-based attributes, (ii), have superior performance to the CM representation. Consequently, we conclude that the Voronoi-based attributes carry useful information about a material. However, given the equivalent performance for the composition-only (i) and all-attributes (iii) model in this test, it is not possible to determine whether including structural attributes can lead to an *improved* model compared to a purely composition-dependent model.

One explanation for the similar performance between a model trained on composition-only and composition-and-structure representations is that the ICSD dataset contains too few examples of multiple structures at the same composition. Consequently, there could be insufficient training data to build

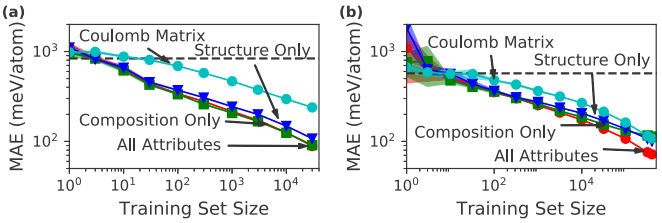

FIG. 6. Performance of ML models trained on different representations in cross-validation tests using data from the (a) ICSD subset of the OQMD and (b) the entire OQMD. These include models trained using all the attributes in our proposed representation and, separately, models created using only the composition-dependent terms and only the structure-dependent terms. The results of a model created using the Coulomb matrix representation with the random forest algorithm is shown for comparison. Shaded regions represent the 90% confidence intervals.

a model that benefits from the additional structural information. To test this hypothesis, we repeated the cross-validation test using a dataset comprised of all nonduplicate entries from the entire OQMD (435 792 entries), which contains dramatically more examples of multiple structures at a single composition. Only 51% of the training entries in this dataset lack another structure at the same composition, which is lower than the 70% of entries without another example structure in the ICSD dataset used previously. With the larger dataset, we observed a significant improvement when using both the structure- and composition-based attributes rather than either subset of attributes alone [as shown in Fig. 6(b)]. At a training set size of 400 000 entries, the model using structural and composition-based attributes has an error rate around 35% lower than the composition only models—showing that there is an advantage to including crystal-structure information into ML models.

The increased accuracy of the all attributes model on the OQMD dataset is not merely an effect of training set size. At a training set size of 10^4, the composition-only model trained on the OQMD dataset (with fewer compositions with only one structure) has a 7% larger MAE than the all attributes model (185 ± 3.5 vs 174 ± 2.0 meV/atom). For the same training set size and the ICSD training set, the composition-only and all attributes model have approximately the same MAE (125 ± 2.2 vs 126 ± 1.8 meV/atom, respectively). The difference between the composition-only and all attributes model in our full OQMD test only becomes larger with increasing sample size. To further test this hypothesis, we performed a cross-validation test using a dataset containing only compositions with multiple structures and find the MAE of the all attributes model to be significantly lower than the composition-only model (105 ± 0.4 vs 158 ± 0.9 meV/atom, respectively). This lower error demonstrates that there is indeed an advantage to introducing structure-based attributes into our ML models. Given the results of our previous tests, the improvement is only significant in datasets where there are sufficient training examples of multiple structures at a single composition.

IV. APPLYING METHOD TO PREDICTING NEW MATERIALS

In this section, we explore using this model to assess the performance of our ML models in two applications: (1)

predicting the formation enthalpy of experimentally observed compounds yet to be included in the OQMD and (2) identifying which materials are most likely to be stable out of a list of compounds studied via a high-throughput search. In both cases, we also seek to determine whether our models can perform well when provided with only the unrelaxed structures that serve as input into DFT calculations. In contrast, we used the fully relaxed structures generated as output from a DFT calculation as input into our ML model in the cross-validation tests in the previous section.

A. Validation with yet-unevaluated materials

One unresolved question from our cross-validation test is whether our models can predict the formation enthalpy of a material without knowledge of the equilibrium structure. To answer this question, we used our model to predict the formation enthalpies of compounds from the ICSD that have yet to be included in the OQMD. The compounds we tested generally have large unit cell sizes, which leads to high computational costs to evaluate with DFT and makes the ability to predict their energies with ML particularly useful [65]. To make our model as accurate as possible, we trained a ML model on the full OQMD dataset. We then used this model to evaluate the 12 667 entries from the ICSD that had not yet been added to the OQMD, which required less than 2 hours on a 2.2 GHz CPU. We then selected a total of 45 entries from this list to validate with DFT using three different strategies: (1) randomly selecting entries, (2) selecting entries predicted to have the most negative ΔH_f, and (3) selecting those predicted that have the largest stability (farthest below the energy of the OQMD convex hull at that composition [66]). By studying these three different strategies separately, we can also assess how best to use our ML model in practice.

As shown in Fig. 7 and Table I, we observed the best performance of the model in the entries that were randomly selected from the dataset—a MAE of 119 ± 47 meV/atom. This is excellent accuracy when considering that these predictions were made before determining the equilibrium structure of the material. The change in the predicted formation enthalpy between the model given the input structure and fully relaxed structure was below 25 meV/atom for 13 out of 15 materials—far below the MAE of 80 meV/atom observed in cross validation. These results show that our ML model can predict the formation enthalpy of unstudied compounds with an accuracy on the order of 100 meV/atom, and the predictions of our model are relatively insensitive to structural relaxations.

The MAE for materials selected by finding those with minimal ΔH_f was generally higher, 177 ± 64 meV/atom, but our model was successful in locating materials with especially large formation enthalpies. The worst-performing entry in this dataset, CeF_4, is likely is an outlier because the DFT calculations in the OQMD treat Ce with only three valence electrons [4]. Consequently, the Ce^{4+} is not modeled correctly, and formation enthalpy for CeF_4 will be more positive than what might be expected based on Ce in other oxidation states and the behavior of other metal-fluoride salts. There are four examples of Ce^{4+} in the list of the worst 2% of predictions described in Sec. III B (CeO_2, $BaCeN_2$, Li_2CeN_2, and Ce_2SeN_2), and the ML predicts a more negative formation

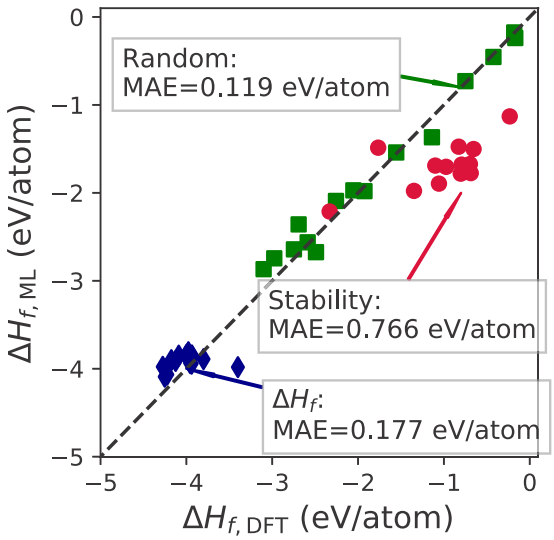

FIG. 7. Comparison of formation enthalpies (ΔH_f) predicted using ML and computed using DFT. The ML model was trained on the formation enthalpies of all 435 792 nonduplicate entries from the OQMD. Each material was selected from a list of 12 667 entries from the ICSD that have yet to be included in the OQMD using three different strategies: (green squares) random selection, (blue diamonds) predictions with the lowest ΔH_f, and (red circles) with the largest, negative difference between the predicted ΔH_f and the OQMD convex hull.

enthalpy than DFT in all cases, just as observed for CeF_4. Provided with enough training examples, it is possible for our model to learn the abnormal behavior of Ce^{4+}, but it apparently lacks the ability with the current training set. If we exclude this compound from the analysis, the error rate in our test is only 148 ± 41 meV/atom. Regardless, the accuracy levels observed in these tests is sufficiently high to successfully identify materials with exceptionally low ΔH_f. All compounds selected based on minimal formation enthalpy are within the 97th percentile (99.4th without CeF_4) and, on average, above the 99.5th percentile of all compounds in the ICSD. While the numerical accuracy of predictions is slightly worse when preferentially selecting large formation enthalpy materials than when randomly selecting materials, we do find it sufficient to identify which materials are most likely to have a large, negative formation enthalpy out of a large dataset.

Of our three selection strategies, the accuracy of our predictions was worst when selecting materials predicted to be the most stable relative to other compounds. In this test case, our error rates were 766 ± 125 meV/atom, which is approximately the error expected when guessing the mean formation enthalpy of the OQMD training set for all compounds. This poor performance could be a result of the biasing effect described by Faber *et al.* [29]. In their paper, Faber *et al.* observed a low success rate when selecting Elpasolite materials based on the predicted stability with reference to other compounds. They attributed this low success rate to this strategy of selecting materials "systematically favor[ing] those [predictions] with negative ML formation energy errors" [29]. Consistent with their observation, nearly all of our predictions made with this strategy have negative formation enthalpy errors and are well

TABLE I. Performance of the ML algorithm in predicting the formation enthalpy (ΔH_f) of 45 materials outside of the training set that were selected with three different strategies. The DFT computed value is compared to the ML prediction using the structure provided to DFT (before relaxation) and the relaxed, output structure.

			Before Relaxation		After Relaxation		
	Composition	ΔH_f, DFT (eV/atom)	ΔH_f, ML (eV/atom)	Error (eV/atom)	ΔH_f, ML (eV/atom)	Change (eV/atom)	Error (eV/atom)
Random	$CrHg_3Pb_2O_8$	−1.139	−1.368	0.229	−1.370	−0.002	0.231
	$Y_2Co_{14}B$	−0.181	−0.176	0.005	−0.174	0.002	0.006
	$YH_3C_3S_2O_{12}F_9$	−1.555	−1.541	0.014	−1.535	0.006	0.021
	$CuH_{12}C_5S_4N$	−0.168	−0.239	0.071	−0.216	0.023	0.048
	$Rb_2Tc_3Se_6$	−0.750	−0.730	0.020	−0.727	0.003	0.022
	Li_6CaCeO_6	−2.257	−2.092	0.165	−1.373	0.719	0.885
	$Na_5Ti_2VSi_2O_{13}$	−2.747	−2.643	0.104	−2.656	−0.013	0.091
	ErP_5O_{14}	−2.692	−2.359	0.334	−2.368	−0.009	0.325
	$Cs_2USi_6O_{15}$	−3.100	−2.868	0.232	−2.854	0.014	0.246
	NaH_2PO_4	−2.055	−1.972	0.083	−1.976	−0.004	0.079
	$Na_5TbW_4O_{16}$	−2.587	−2.563	0.024	−2.569	−0.006	0.018
	$NaU_2H_5C_4O_{20}$	−1.925	−1.981	0.055	−2.019	−0.038	0.093
	U_2MoO_8	−2.976	−2.744	0.232	−2.737	0.008	0.240
	$DyMnSn_2$	−0.423	−0.455	0.032	−0.454	0.000	0.031
	$RbVP_2O_8$	−2.49	−2.674	0.184	−2.672	0.002	0.182
	Mean	−1.803	−1.760	0.119	−1.713	0.047	0.168
	90% CI	0.470	0.435	0.047	−0.437	0.085	0.102
Largest ΔH_f	$SrMgF_4$	−3.952	−3.876	0.077	−3.862	0.014	0.091
	CeF_4	−3.400	−3.982	0.583	−3.887	0.095	0.488
	Sr_2ScF_7	−4.175	−3.902	0.273	−3.924	−0.022	0.251
	$RbLu_3F_{10}$	−4.275	−3.978	0.297	−4.001	−0.023	0.274
	$BaAlF_5$	−3.956	−3.936	0.020	−3.949	−0.013	0.007
	$ThZrF_8$	−4.223	−4.066	0.157	−4.039	0.027	0.183
	KU_2F_9	−3.800	−3.891	0.091	−3.869	0.022	0.069
	$RbTh_2F_9$	−4.252	−4.091	0.161	−4.104	−0.013	0.148
	Ba_2ZrF_8	−4.125	−3.912	0.213	−3.914	−0.002	0.212
	$Sr_5Al_2F_{16}$	−4.089	−3.851	0.238	−3.854	−0.003	0.235
	KYF_4	−3.976	−3.812	0.164	−3.820	−0.008	0.156
	$SrAlF_5$	−4.002	−3.849	0.153	−3.828	0.021	0.174
	$BaNaZr_2F_{11}$	−3.935	−3.848	0.087	−3.872	−0.024	0.064
	$Ba_6Mg_{11}F_{34}$	−3.931	−3.864	0.066	−3.864	0.001	0.067
	$Ba_7Cl_2F_{12}$	−3.939	−3.943	0.004	−3.943	0.000	0.004
	Mean	−4.005	−3.918	0.177	−3.913	0.004	0.166
	90% CI	0.099	0.038	0.064	0.037	0.014	0.056
Largest Stability	$CeTl_5Fe_2(NO_2)_{12}$	−0.804	−1.782	0.978	−1.754	0.028	0.951
	$YTl_5Cu_2(NO_2)_{12}$	−0.697	−1.673	0.976	−1.652	0.021	0.955
	$Rb_2BiCl_5O_{20}$	−0.655	−1.502	0.847	−1.502	0.000	0.847
	$YTl_5Co_2(NO_2)_{12}$	−0.752	−1.757	1.005	−1.727	0.030	0.974
	$TmAu_2F_9$	−2.331	−2.212	0.119	−2.211	0.002	0.120
	VXe_2F_{34}	−1.348	−1.978	0.629	−2.072	−0.095	0.724
	$CeTl_5Ni_2N_{12}O_{34}$	−0.777	−1.754	0.977	−1.780	−0.026	1.003
	$CsXe_3O_3F_{36}$	−0.687	−1.776	1.088	−1.782	−0.006	1.094
	$ScH_3Cl_2O_{10}$	−1.058	−1.895	0.837	−1.929	−0.034	0.872
	$Lu(H_2ClO_3)_5$	−0.974	−1.704	0.730	−1.669	0.035	0.695
	$Er_5C_2Br_9$	−1.766	−1.486	0.279	−1.459	0.027	0.306
	$SnCl_8O_{25}$	−0.233	−1.131	0.898	−1.121	0.010	0.888
	$NiXe_4F_{28}$	−0.828	−1.475	0.648	−1.385	0.090	0.558
	$Np_2H_8Cl_2O_{13}$	−1.101	−1.690	0.590	−1.700	−0.009	0.599
	$CeAg_6(NO_3)_9$	−0.794	−1.679	0.885	−1.667	0.012	0.873
	Mean	−0.987	−1.700	0.766	−1.694	0.006	0.764
	90% CI	0.231	0.113	0.125	0.123	0.018	0.123

within the 99th percentile of magnitude of errors observed in our cross-validation test. This poor performance suggests that identifying materials based on the difference between ML-predicted formation energy and the energies of competing phases is problematic. Consequently, we recommend either searching for new stable materials by selecting those with large formation energies or directly predicting the stability with reference to other phases.

Overall, this validation test was particularly successful. We observed formation energy errors of approximately 125 meV/atom for randomly selected materials and successfully located materials with exceptionally low formation enthalpies. In these cases, making the ML predictions required only a tiny fraction of the tens of thousands of CPU hours of DFT calculations necessary to validate them for these limited test cases. It is also worth emphasizing that these high accuracies were achieved without knowledge of the equilibrium DFT geometry. Across all 45 predictions, the mean absolute difference between the prediction of our model with the initial guess provided to DFT and with the fully relaxed structure was only 35 ± 27 meV/atom—below the error expected in the prediction from the cross-validation experiment at 80 meV/atom and those observed in this section (354 ± 87 meV/atom). This result demonstrates that our models can be used effectively when only an approximate model of the relaxed geometry is known—a very important feature when searching for new crystalline materials using ML.

B. Application to combinatorial searches

To test how our models could be applied to the high-throughput materials discovery process, we simulated the results of searching for new compounds based on several common crystal structures. First, we trained each model using data from all 32 111 compounds in the OQMD that are based on entries from the ICSD. We used only the ICSD entries as a training set because it is not computationally feasible to train the PRDF and CM models on the entire OQMD. Then, we used this model to evaluate the formation enthalpies of all entries in the OQMD with the B2, $L1_0$, and orthorhombically distorted perovskite crystal structures. To simulate how this model would be used in practice, we evaluated the formation energy of the compound using the same input geometry provided to the DFT calculation: simply the original prototype structure with new elements substituted in. In contrast to Sec. IV A, the inputs to this model emulate a "structure prediction" use case for ML, where no experimental data about the structure is known. These three structural prototypes were chosen as separate test cases to sample structures that have a variety of local environments and that are known to be stable for compounds with both metallic and ionic bonding. Furthermore, the B2 and $L1_0$ datasets were created by generating all possible combinations of elements into the structure, which is useful for testing the ability of the model to evaluate a broad range of chemistries. In contrast, the orthorhombic perovskite dataset is limited to only ABO_3 metal oxides and predominately includes materials with negative formation enthalpies, which will allow us to evaluate the performance over a more-restricted space. Additionally, the orthorhombic perovskite is the structure with the largest number of structural degrees of freedom available in the large numbers from the OQMD, which will allow us to better test the effect of structural relaxation on prediction accuracy.

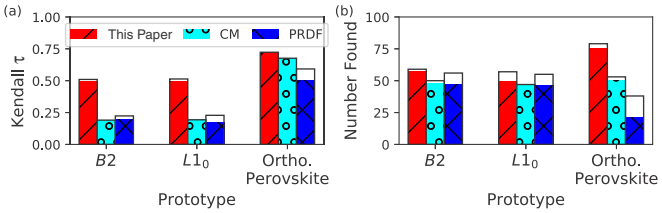

FIG. 8. Comparison of the ability of different ML methods to rank different types of compounds based on DFT formation energy, measured using two different metrics. (a) The Kendall τ ranking correlation coefficient, which is based on how well the model ranks the entire dataset. A correlation value of 1.0 corresponds to perfect ranking. (b) The number of the 100 compounds with the lowest DFT formation energy that were predicted by the model to be within the lowest 100 compounds. Each model was trained on the DFT-predicted formation energy of 32 111 inorganic compounds from the ICSD. The solid bar indicates the ranking performance using the input structure provided to DFT. Black outline around each bar indicates the ranking accuracy when provided with the fully relaxed output from DFT.

To evaluate the ability of each ML algorithm to rank compounds from most to least stable, we measured the Kendall τ ranking correlation coefficient between the predicted and actual formation enthalpies for each prototype structure. The Kendall τ, which is defined as the difference in the fraction of pairs in a list that are correctly and incorrectly ordered [67], allows us to understand how well the algorithm could be employed to identify compounds that are likely to be stable. As shown in Fig. 8(a), the model created using our new method has the highest ranking correlation coefficient for all three considered test cases. For the $L1_0$ structure, our model performs twice as well as the CM model and almost three times better than the PRDF model; the differences are similarly large in the B2 test case.

The performance for all three ML methods was best for the perovskite test case, where the dataset was restricted to metal oxides with mostly (99.3%) negative formation enthalpies. In that example, our model had very strong ranking performance—an 85% success rate. This exceptional ranking performance is likely a result of the dataset containing mostly materials that have negative formation enthalpies. If we repeat the ranking test for B1 and $L1_0$ with only compounds with negative formation enthalpies, we observe improved performance for all three ML techniques. The improved performance on a dataset containing only materials with negative formation enthalpies is consistent with our previous finding that the model performs worst for materials with positive formation enthalpies [see Fig. 5(a)]. Consequently, we propose that the selection performance of each model could be improved by first screening the space based on heuristic chemical rules (e.g., are the elements in reasonable oxidation states?). This could eliminate compounds that are more likely to be extremely unstable at the risk of potentially missing exciting materials with elements in surprising oxidation states (as in Ref. [29]).

One factor leading to improved performance of our method is the insensitivity of our representation to changes in volume. In the case of B2, the only degree of freedom in the crystal structure is the volume. Consequently, our predictions are not dependent on the quality of the initial guess for the equilibrium volume. Incidentally, the accuracy of the CM method is also only negligibly affected. In contrast, the predictive accuracy of the PRDF method increases significantly when we use the final, fully relaxed geometry as input to the model. For our other two test cases—$L1_0$ and orthorhombic perovskite—the predicted enthalpies depend upon relaxation because there is more than one degree of freedom in the structure. Even so, the mean change between the initial and final structures in the predicted ΔH_f is approximately 65 meV/atom, and the correlation coefficient between the two predictions is approximately 99% for both structure types. Correspondingly, the ranking performance only changes slightly. Considering both this fact and the highest Kendall τ ranking coefficient, we conclude our model is the best choice for this ranking task.

In practice, these ML models might only be used to select the entries with the lowest predicted formation enthalpy. To measure the ability of each model to identify entries with the largest formation enthalpies, we measured the number of entries predicted by our ML model to have the 100 largest formation enthalpies that were within the top 100 of the test set. As shown in Fig. 8(b), the model created using our method performs the best per this metric for all three cases, and over half of the predictions made with our model are actually within the top 100. What this high predictive accuracy suggests is that it is possible to use a ML model trained on data with dissimilar crystal structures (e.g., the entire OQMD) to predict stable compounds with a target crystal structure without having to first create a new, problem-specific training set—as is common practice in previous ML-assisted searches for stable compounds [29,40,41,68]. Our method can also be used to predict the stability of compounds with infrequently observed structure types, a limitation of crystal structure prediction methods that search for correlations between energies of commonly occurring prototypes [22,69–71]. By using existing data and our ML technique, we can quickly make predictions of which materials are most likely to be stable and use that knowledge to accelerate high-throughput DFT searches for new materials.

V. CONCLUSIONS

In this paper, we present a strategy for predicting the formation energy of crystalline, inorganic compounds using characteristics derived from the Voronoi tessellation of its structure and ML. We demonstrate that these models are more accurate in cross validation and better at ranking unseen compounds from most to least stable than those produced using the CM [30] and PRDF [42] methods and equivalently as fast. Furthermore, we show that our model is learning the effect of structure on formation enthalpy and can accurately predict the formation enthalpy of materials without knowledge of the fully equilibrated crystal structure. Provided the high predictive accuracy of this method and the ability to utilize large training datasets, we envision it will be possible to employ this method to identify new, stable materials at a low computational cost.

ACKNOWLEDGMENTS

This work was performed under the following financial assistance Award 70NANB14H012 from U.S. Department of Commerce, National Institute of Standards and Technology as part of the Center for Hierarchical Materials Design (CHiMaD). V.I.H. was supported by NSF Grant No. DMR-1309957. The authors also thank Gus Hart for an insightful discussion on deformation stability and Voronoi tessellations and Vancho Kocevski for his helpful comments on the manuscript.

[1] 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, APL Mater. **1**, 011002 (2013).
[2] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, Nat. Mater. **12**, 191 (2013).
[3] G. Ceder and K. Persson, Sci. Am. **309**, 36 (2013).
[4] 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).
[5] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, JOM **65**, 1501 (2013).
[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, J. Phys. Chem. Lett. **2**, 2241 (2011).
[7] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, and O. Levy, Comput. Mater. Sci. **58**, 227 (2012).
[8] X. Qu, A. Jain, N. N. Rajput, L. Cheng, Y. Zhang, S. P. Ong, M. Brafman, E. Maginn, L. A. Curtiss, and K. A. Persson, Comput. Mater. Sci. **103**, 56 (2015).
[9] G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari, and B. Kozinsky, Comput. Mater. Sci. **111**, 218 (2016).
[10] http://nomad-repository.eu/cms/.
[11] S. S. Borysov, R. M. Geilhufe, and A. V. Balatsky, PLoS One **12**, e0171501 (2017).
[12] H. Chen, G. Hautier, A. Jain, C. Moore, B. Kang, R. Doe, L. Wu, Y. Zhu, Y. Tang, and G. Ceder, Chem. Mater. **24**, 2009 (2012).
[13] S. Kirklin, B. Meredig, and C. Wolverton, Adv. Energy Mater. **3**, 252 (2013).
[14] R. Gautier, X. Zhang, L. Hu, L. Yu, Y. Lin, T. O. L. Sunde, D. Chon, K. R. Poeppelmeier, and A. Zunger, Nat. Chem. **7**, 308 (2015).
[15] G. L. W. Hart, S. Curtarolo, T. B. Massalski, and O. Levy, Phys. Rev. X **3**, 041035 (2013).
[16] I. E. Castelli, T. Olsen, S. Datta, D. D. Landis, S. Dahl, K. S. Thygesen, and K. W. Jacobsen, Energy Environ. Sci. **5**, 5814 (2012).
[17] W. Chen, J.-H. Pöhls, G. Hautier, D. Broberg, S. Bajaj, U. Aydemir, Z. M. Gibbs, H. Zhu, M. Asta, G. J. Snyder, B. Meredig, M. A. White, K. Persson, and A. Jain, J. Mater. Chem. C **4**, 4414 (2016).

[18] A. A. Emery, J. E. Saal, S. Kirklin, V. I. Hegde, and C. Wolverton, Chem. Mater. **28**, 5621 (2016).

[19] A. Bhatia, G. Hautier, T. Nilgianskul, A. Miglio, J. Sun, H. J. Kim, K. H. Kim, S. Chen, G.-M. Rignanese, X. Gonze, and J. Suntivich, Chem. Mater. **28**, 30 (2016).

[20] J. He, M. Amsler, Y. Xia, S. S. Naghavi, V. I. Hegde, S. Hao, S. Goedecker, V. Ozoliņš, and C. Wolverton, Phys. Rev. Lett. **117**, 046602 (2016).

[21] J. E. Saal and C. Wolverton, Scr. Mater. **67**, 798 (2012).

[22] S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, and G. Ceder, Phys. Rev. Lett. **91**, 135503 (2003).

[23] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, Phys. Rev. B **89**, 094104 (2014).

[24] J. Carrete, W. Li, N. Mingo, S. Wang, and S. Curtarolo, Phys. Rev. X **4**, 011019 (2014).

[25] A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput, and I. Tanaka, Phys. Rev. Lett. **115**, 205901 (2015).

[26] G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis, and T. Lookman, Sci. Rep. **6**, 19375 (2016).

[27] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, npj Comput. Mater. **2**, 16028 (2016).

[28] A. M. Deml, R. O'Hayre, C. Wolverton, and V. Stevanovic, Phys. Rev. B **93**, 085142 (2016).

[29] F. A. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, Phys. Rev. Lett. **117**, 135502 (2016).

[30] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, Int. J. Quantum Chem. **115**, 1094 (2015).

[31] E. O. Pyzer-Knapp, G. N. Simm, and A. Aspuru Guzik, Mater. Horiz. **3**, 226 (2016).

[32] C. S. Kong, S. R. Broderick, T. E. Jones, C. Loyola, M. E. Eberhart, and K. Rajan, Phys. B Condens. Matter **458**, 1 (2015).

[33] M. de Jong, W. Chen, R. Notestine, K. Persson, G. Ceder, A. Jain, M. Asta, and A. Gamst, Sci. Rep. **6**, 34256 (2016).

[34] A. Furmanchuk, A. Agrawal, and A. Choudhary, RSC Adv. **6**, 95246 (2016).

[35] T. Moot, O. Isayev, R. W. Call, S. M. McCullough, M. Zemaitis, R. Lopez, J. F. Cahoon, and A. Tropsha, Mater. Discov. **6**, 9 (2016).

[36] H. Wu, A. Lorenson, B. Anderson, L. Witteman, H. Wu, B. Meredig, and D. Morgan, Comput. Mater. Sci. **134**, 160 (2017).

[37] A. Agrawal and A. Choudhary, APL Mater. **4**, 053208 (2016).

[38] L. Ward and C. Wolverton, Curr. Opin. Solid State Mater. Sci. **21**, 167 (2017).

[39] A. Seko, T. Maekawa, K. Tsuda, and I. Tanaka, Phys. Rev. B **89**, 054303 (2014).

[40] A. O. Oliynyk, E. Antono, T. D. Sparks, L. Ghadbeigi, M. W. Gaultois, B. Meredig, and A. Mar, Chem. Mater. **28**, 7324 (2016).

[41] G. Pilania, P. V Balachandran, C. Kim, and T. Lookman, Front. Mater. **3**, 19 (2016).

[42] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, and E. K. U. Gross, Phys. Rev. B **89**, 205118 (2014).

[43] A. Seko, H. Hayashi, K. Nakayama, A. Takahashi, and I. Tanaka, Phys. Rev. B **95**, 144110 (2017).

[44] V. Stevanović, S. Lany, X. Zhang, and A. Zunger, Phys. Rev. B **85**, 115104 (2012).

[45] G. Hautier, S. P. Ong, A. Jain, C. J. Moore, and G. Ceder, Phys. Rev. B **85**, 155208 (2012).

[46] G. Kresse and J. Hafner, Phys. Rev. B **47**, 558 (1993).

[47] G. Kresse and D. Joubert, Phys. Rev. B **59**, 1758 (1999).

[48] https://github.com/wolverton-research-group/qmpy.

[49] A. Belsky, M. Hellenbrandt, V. L. Karen, and P. Luksch, Acta Crystallogr. Sect. B **58**, 364 (2002).

[50] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevB.96.024104 for a detailed description of the attributes used in this study, and the scripts, datasets, and software necessary to replicate this paper.

[51] V. Botu and R. Ramprasad, Int. J. Quantum Chem. **115**, 1074 (2015).

[52] L. Yang, S. Dacek, and G. Ceder, Phys. Rev. B **90**, 054102 (2014).

[53] A. P. Bartók, R. Kondor, and G. Csányi, Phys. Rev. B **87**, 184115 (2013).

[54] A. Jain, G. Hautier, S. P. Ong, and K. Persson, J. Mater. Res. **31**, 977 (2016).

[55] G. Voronoi, J. reine angew. Math. **134**, 198 (1908).

[56] E. Wigner and F. Seitz, Phys. Rev. **43**, 804 (1933).

[57] J. Cowley, Phys. Rev. **77**, 669 (1950).

[58] https://bitbucket.org/wolverton/magpie.

[59] https://bitbucket.org/wolverton/vassal.

[60] L. Breiman, Mach. Learn. **45**, 5 (2001).

[61] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, J. Mach. Learn. Res. **12**, 2825 (2011).

[62] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, ACM SIGKDD Explor. Newsl. **11**, 10 (2009).

[63] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).

[64] Several other infrequently appearing elements (He, Ne, Ar, Pm, Ac) violate this trend because they appear infrequently in both the training set and in the list of worst predictions. In the case of the noble gas elements in this surprisingly good category, they only appear as elemental compounds in the training set, and our model correctly identifies those compounds as having near-zero formation enthalpies.

[65] The median size of the compounds we tested is 64. In contrast, the median size of compounds in the OQMD is 4. Considering the cost DFT scales at least $O(N^2 \log N)$ with the number of atoms (N), the typical calculation in this set is at least 768 times more expensive than the typical OQMD entry.

[66] A. R. Akbarzadeh, V. Ozoliņš, and C. Wolverton, Adv. Mater. **19**, 3233 (2007).

[67] G. S. Shieh, Stat. Probab. Lett. **39**, 17 (1998).

[68] G. H. Jóhannesson, T. Bligaard, A. V. Ruban, H. L. Skriver, K. W. Jacobsen, and J. K. Nørskov, Phys. Rev. Lett. **88**, 255506 (2002).

[69] C. C. Fischer, K. J. Tibbetts, D. Morgan, and G. Ceder, Nat. Mater. **5**, 641 (2006).

[70] G. Hautier, C. C. Fischer, A. Jain, T. Mueller, and G. Ceder, Chem. Mater. **22**, 3762 (2010).

[71] C. S. Kong, W. Luo, S. Arapan, P. Villars, S. Iwata, R. Ahuja, and K. Rajan, J. Chem. Inf. Model. **52**, 1812 (2012).