

Calibrating DFT Formation Enthalpy Calculations by Multifidelity Machine Learning

Sheng Gong, Shuo Wang, Tian Xie, Woo Hyun Chae, Runze Liu, Yang Shao-Horn,* and Jeffrey C. Grossman*

Cite This: *JACS Au* 2022, 2, 1964–1977

Read Online

ACCESS |

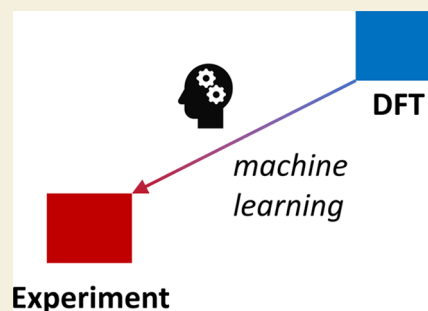
Metrics & More

Article Recommendations

Supporting Information

ABSTRACT: The application of machine learning to predict materials properties measured by experiments are valuable yet difficult due to the limited amount of experimental data. In this work, we use a multifidelity random forest model to learn the experimental formation enthalpy of materials with prediction accuracy higher than the Perdew–Burke–Ernzerhof (PBE) functional with linear correction, PBEsol, and meta-generalized gradient approximation (meta-GGA) functionals (SCAN and r^2 SCAN), and it outperforms the hotly studied deep neural network-based representation learning and transfer learning. We then use the model to calibrate the DFT formation enthalpy in the Materials Project database and discover materials with underestimated stability. The multifidelity model is also used as a data-mining approach to find how DFT deviates from experiments by explaining the model output.

KEYWORDS: formation enthalpy, stability, machine learning, DFT, multifidelity, transfer learning



INTRODUCTION

To accelerate the design of new materials, computational methods such as density functional theory (DFT)¹ have been employed to generate large data sets that contain more than 10^5 entries of material properties, including the Materials Project (MP),² Open Quantum Materials Database (OQMD),³ the Automatic Flow of Materials Discovery Library (AFLOW),⁴ and the Joint Automated Repository for Various Integrated Simulations (JARVIS).⁵ While the availability of such databases has boosted the exploration of novel materials,^{6–15} it is important to note that most of the data is generated with computationally “cheap” DFT functionals such as Perdew–Burke–Ernzerhof (PBE),¹⁶ which can, in turn, lead to non-negligible errors when compared with experimental measurements.

As an example, the formation enthalpy (ΔH_f) is a fundamental property that determines the thermodynamic stability of materials. The mean absolute error (MAE) between the computed ΔH_f in these large DFT databases and experimental measurements is reported to be ~ 0.1 eV/atom.^{3,17} Due to the sensitivity of phase stability to energy, such a difference might be the difference between a material that is readily synthesizable and one that is almost impossible to realize.^{18–20} In addition, because of the limited amount of available experimental data, currently most machine learning (ML) models applied to materials are trained on DFT datasets,^{6,21–35} making any error in the DFT calculations critical to the usefulness of such ML models.^{7,31,36,37}

To improve the accuracy of formation enthalpy calculations, a number of density functionals have been developed, such as

PBEsol,³⁸ SCAN,³⁹ r^2 SCAN,⁴⁰ and HSE,⁴¹ which have shown significant improvement in the accuracy of formation enthalpy calculations.^{42–44} On the other hand, except PBEsol, these more accurate functionals are computationally more expensive, limiting their utility for the generation of large databases.^{43,45} Empirical corrections represent another, faster approach to improve the accuracy of the prediction of ΔH_f . For example, in the MP dataset, ΔH_f of certain materials (including oxides, phosphates, borates, and silicates) is empirically corrected by fitted element corrections,⁴⁶ and in OQMD, ΔH_f is corrected by a chemical-potential fitting.³ Very recently, Wang et al.⁴⁷ proposed a linear correction scheme with an error of 0.051 eV/atom compared with experimental values on a dataset with 222 materials containing certain anions and transition metals. Yet, despite this recent success in lowering the error for some chemical systems,⁴⁸ such corrections are based on the human understanding of specific chemistries and relatively simple assumptions and are thus difficult to be transferable across different chemistries.^{46,48} It would be beneficial to design prediction schemes that can automatically extract the chemistry–property relationship across different chemistries without human intervention, and data-driven ML meth-

Received: April 15, 2022

Revised: May 23, 2022

Accepted: June 1, 2022

Published: September 9, 2022



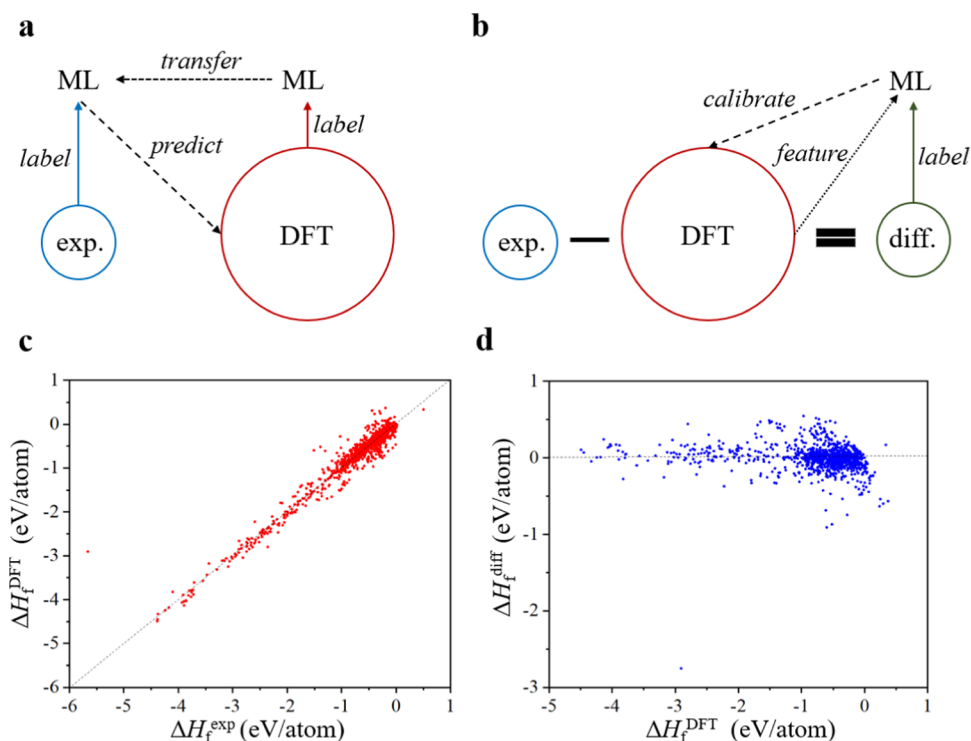


Figure 1. Illustrations of the machine learning frameworks and datasets used in this work. (a, b) Schematics of transfer learning and multifidelity machine learning in this work, respectively. In panel (a), first the ΔH_f^{DFT} is used as a label to train an ML model, then the weights of the first ML model are transferred to initialize a second ML model, and then the ΔH_f^{exp} is used as a label to train the second model; finally, the second model is used to predict ΔH_f^{exp} of all materials in the large DFT dataset. In panel (b), first the dataset of the difference between ΔH_f^{exp} and ΔH_f^{DFT} is constructed (ΔH_f^{diff}), then ΔH_f^{diff} is used as a label to train an ML model with the ΔH_f^{DFT} as an input feature, and finally the trained model is used to calibrate the difference between ΔH_f^{DFT} and ΔH_f^{exp} for all materials in the large DFT dataset. (c) ΔH_f^{DFT} versus ΔH_f^{exp} . (d) ΔH_f^{diff} versus ΔH_f^{DFT} .

ods^{18,24,26,28,29,32,45} are promising candidates to learn the complex mapping between chemistry and ΔH_f .

One of the biggest challenges in machine learning of material properties is the lack of experimental data.⁴⁹ Efforts have been made to improve the performance of learning on small experimental datasets by extracting and transferring information from large DFT datasets. Currently, there are mainly two strategies to achieve the transfer between DFT and experimental datasets, transfer learning^{21,28,50–53} and multifidelity machine learning.^{45,54–56} The idea of transfer learning (Figure 1a) is first learning large DFT datasets (source) using a large neural network and then transferring the weights of the network to the machine learning task of small experimental datasets (target). Although transfer learning has achieved success in problems where the source and target data sets are highly correlated,^{21,28,50,51} the approach is mostly applied to neural network architectures, and if the correlation is not strong enough, transfer learning will not improve and may even deteriorate the learning performance.⁵² Different from transfer learning where information is passed by transferring network parameters, in multifidelity machine learning (Figure 1b), information of cheap and low-fidelity data is directly passed to the learning task of expensive and high-fidelity data, either in the feature (input) level⁵⁴ or in the label (output) level.^{45,55–57} In other words, the low-fidelity data can be used as a feature in the machine learning task of high-fidelity data or the task of machine learning the high-fidelity data can be converted to the task of machine learning the difference between high-fidelity data and low-fidelity data, which is also known as Δ -machine learning.⁵⁷ From the handful of previous studies, multifidelity machine learning has shown higher predictive power than single-fidelity ones (directly

learning the high-fidelity data) on material properties like band gaps and energies from different density functionals.^{45,55–57} However, there is no previous work that adapts multifidelity machine learning at both feature and label levels at the same time.

In this work, we present a comprehensive machine learning study of ΔH_f^{exp} using transfer learning and multifidelity machine learning. For the machine learning architectures, we compare four different models, random forests (RFs), multilayer perceptron (MLP), Representation Learning from Stoichiometry (ROOST),²⁶ and Crystal Graph Convolutional Neural Network (CGCNN).³² We find that multifidelity RF in both the feature and label levels has the best prediction performance for ΔH_f^{exp} with more than 30% reduction in MAE compared with DFT results from MP and improved performance compared to recent linear correction schemes⁴⁷ as well as more sophisticated density functionals like PBEsol,³⁸ SCAN,³⁹ and r²SCAN.⁴⁰ We also analyze the effects of machine learning architectures, featurization methods, and information transfer strategies on learning ΔH_f^{exp} and ΔH_f^{diff} . Further, the more accurate ΔH_f are applied to re-evaluate the thermodynamic stability of materials, and cases with underestimated stability in the MP database are discovered. We also use the machine learning model to find where current DFT deviates from experiments by explaining the model output.

RESULTS

Illustration of Machine Learning Frameworks and Datasets

In this work, we use two different strategies to learn ΔH_f^{exp} with the assistance of information from the MP dataset, transfer

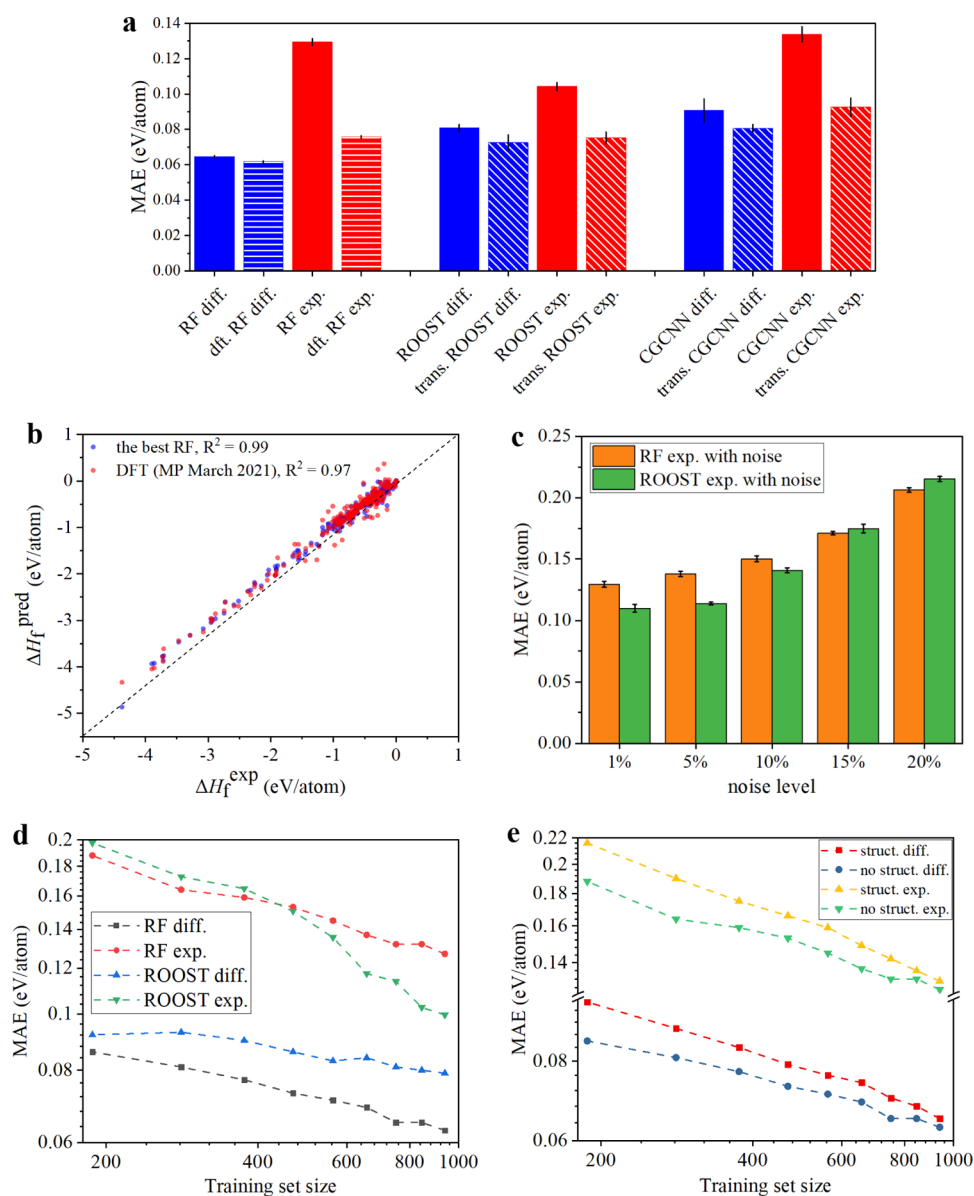


Figure 2. Comparison of machine learning models. (a) Mean average errors (MAEs) between predictions of ΔH_f from some machine learning models and experimental measurements. Each type of machine learning model is trained 10 times to estimate the uncertainty levels. RF denotes random forest with compositional features from matminer,⁵⁹ and ROOST²⁶ and CGCNN³² are two deep learning models that automatically extract materials' fingerprints from compositions and structures, respectively. Here, "dft." in front of "RF" means the model is trained with ΔH_f^{DFT} as an input, "trans." in front of "ROOST" and "CGCNN" model is trained in a transfer learning manner, diff. model is trained on ΔH_f^{diff} , and "exp." model is directly trained on ΔH_f^{exp} . The dashed horizontal line corresponds to the MAE of ΔH_f^{DFT} . (b) ΔH_f^{exp} versus ΔH_f^{ML} from the best RF model (the second from the left in panel (a)) and ΔH_f^{DFT} . (c) MAE of predictions of ΔH_f^{exp} with noise from RF and ROOST. Under each noise level, Gaussian noises with a standard deviation of noise level*0.8 eV/atom (0.8 eV/atom is the standard deviation of the ΔH_f^{exp} dataset) are added to both training set and test set. (d, e) Learning curves of different models. The MAE is for the test set. In panel (e), all of the curves are based on random forest, and "struct." means the model is trained with structural and compositional features, and "no struct." means the model is trained with only compositional features.

learning, and multifidelity machine learning (in the following, " ΔH_f^{DFT} " denotes the empirically corrected PBE ΔH_f by Jain et al.⁴⁶ from the MP database, V2021.03.22). As shown in Figure 1a, in transfer learning, a neural network is first trained on the large MP dataset, then weights of the neural network are transferred to initialize a second neural network, and finally, some of the weights of the second network are optimized by the small ΔH_f^{exp} dataset. Once trained, the second neural network can serve to predict ΔH_f^{exp} of materials in the large MP dataset. In multifidelity machine learning, as shown in Figure 1b, first the dataset of ΔH_f^{diff} ($\Delta H_f^{\text{exp}} - \Delta H_f^{\text{DFT}}$) is built and then machine learning models are trained on the ΔH_f^{diff} dataset, with ΔH_f^{DFT}

also as an input feature of each material. Once trained, the machine learning model can serve to calibrate ΔH_f^{DFT} by adding ΔH_f^{diff} to ΔH_f^{DFT} to obtain ΔH_f^{exp} . The key difference between transfer learning and multifidelity machine learning is that in the former two networks are trained and information transfer is achieved by transferring network weights, while in the latter, only one model is trained and information transfer is achieved by learning the difference between two datasets and adding the ΔH_f^{DFT} as one of the input features. In addition to the two basic strategies as shown in Figure 1a,b, variants are also tested in this work, including a combination of transfer learning and multifidelity machine learning (initializing a network from one

Table 1. Comparison of MAEs between ΔH_f^{exp} and ΔH_f from Different Density Functionals with Different Corrections⁴²

functional (correction)	PBE (random forest)	PBE (linear, Jain et al. ⁴⁶)	PBE (linear, Wang et al. ⁴⁷)	PBEsol (no) ⁴⁴	SCAN (no) ⁴⁴	r ² SCAN (no) ⁴⁴
MAE (eV/atom)	0.0542	0.0935	0.0927	0.0973	0.1024	0.0825

⁴²Different from Figure 2, the reported MAEs here are based on a dataset with 122 materials in the test set that have all of the values of ΔH_f from different sources. The cell of “PBE (random forest)” is the PBE ΔH_f that is first corrected by Jain et al.⁴⁶ and then corrected by the best RF model in this work. “PBE (linear, Jain et al.⁴⁶)” is the one used in the MP database before May 2021 (V2021.03.22) and is the one used as the low-fidelity data in this work (ΔH_f^{DFT}). “PBE (linear, Wang et al.⁴⁷)” is the one used in the MP database after May 2021 (V2021.05.13). “(no)” in the right three cells at the upper row means that no correction is applied to the ΔH_f from the density functional.

trained on ΔH_f^{DFT} and optimizing the newly initialized network by ΔH_f^{diff} and multifidelity machine learning by only learning ΔH_f^{diff} or only adding ΔH_f^{DFT} as an input feature.

As described above, we choose four different machine learning architectures to realize transfer learning and/or multifidelity machine learning, which are RF, MLP, ROOST, and CGCNN. The choice aims to increase the variety of machine learning architectures to fairly evaluate the effect of transfer learning and multifidelity learning and to enlarge the hypothesis space to search for the best machine learning models for predicting ΔH_f^{exp} . These ML architectures also provide varieties in terms of basic algorithms, input information, and featurization: MLP, ROOST, and CGCNN are based on neural networks, while RF is not; ROOST only needs compositions as an input, while CGCNN takes both compositions and three-dimensional (3D) structures as input, and RF and MLP can be trained either with or without structural information; RF and MLP need human-engineered featurization, while ROOST and CGCNN learn fingerprints of materials in the training process.

In this work, we choose the Materials Project (MP) database (V2021.03.22) as the source of ΔH_f^{DFT} because MP is a widely used large DFT database and the difference of ΔH_f between MP and other large DFT databases is not large. For example, the difference between ΔH_f of 563 materials from MP and OQMD is reported to be 0.028 eV/atom.³ As for the experimentally measured ΔH_f , we combine the IIT dataset¹⁷ and SSUB dataset⁵⁸ and remove the duplicates, leading to 1143 data points with available ΔH_f^{exp} , ΔH_f^{DFT} , and DFT-optimized 3D atomic structures from MP. In addition to the value of ΔH_f^{exp} , there are also uncertainty estimations in the IIT dataset,¹⁷ from which one can see that the mean uncertainty of ΔH_f^{exp} based on 499 materials is around 0.023 eV/atom. More details about the data collection procedure are provided in the Methods section. ΔH_f^{DFT} and ΔH_f^{exp} are compared in Figure 1c, from which one can see that ΔH_f^{DFT} are already quite close to ΔH_f^{exp} in value, and there is no clear systematic shift between ΔH_f^{DFT} and ΔH_f^{exp} . As shown in Figure 1d, the distribution of ΔH_f^{diff} is centered around zero, and there is no obvious correlation between ΔH_f^{diff} and ΔH_f^{DFT} . From Figure 1c,d, one can see that ΔH_f^{diff} has a narrower distribution than ΔH_f^{exp} with the standard deviation of 0.1718 and 0.8000 eV/atom for the datasets of ΔH_f^{diff} and ΔH_f^{exp} , respectively.

Predicting ΔH_f^{exp} by Machine Learning

For the RF and MLP, compositional and structural features are provided from matminer⁵⁹ as input features (a list of features is provided in the Methods section), for ROOST only, the compositions are provided as input and it automatically learns the fingerprints of materials, and for CGCNN, the compositions and 3D atomic structures are provided as input and the fingerprints are learned in the training. To test the prediction performance, 20% of the 1143 materials are chosen randomly as the test set. Details about the training procedure are provided in the Methods section. As a baseline, for the test set, we find that

the MAE between ΔH_f^{DFT} and ΔH_f^{exp} is 0.0955 eV/atom. The test results for some machine learning models are shown in Figure 2a with detailed lists of all machine learning models in Figure S3 and Table S1, and here we analyze the results from the following aspects:

- (1) The best performance is achieved with the RF model that is trained on ΔH_f^{diff} and has compositional features and ΔH_f^{DFT} as input features (Figure 2a). The error for this best case, 0.0617 eV/atom, is roughly 30% lower than that of ΔH_f^{DFT} . The parity plot of ΔH_f^{DFT} and ΔH_f^{ML} from the best RF model versus ΔH_f^{exp} of the test set is shown in Figure 2b, from which one can observe that ΔH_f from the best RF model aligns closer to the ΔH_f^{exp} than ΔH_f^{DFT} within the range from -5 to 1 eV/atom. Predictions from the best RF model also have a higher R^2 score (0.99) than that from the DFT calculations in the MP database (0.97).

Recently, Kingsbury et al.⁴⁴ have performed high-throughput calculations for 6000 materials by PBEsol,³⁸ SCAN,³⁹ and r²SCAN functional.⁴⁰ In Table 1, MAEs between experimental ΔH_f and ΔH_f from different density functionals with different empirical corrections are listed. Different from Figure 2, the reported MAEs in Table 1 are based on a dataset with 122 materials that have all of the values of ΔH_f from different sources (these materials are in the test set mentioned above). One can observe that MAE of the best RF model is almost half of that of SCAN,³⁹ PBEsol,³⁸ and also almost half of that of the corrections from Jain et al.⁴⁶ and Wang et al.⁴⁷ The superiority of the best RF model over the meta-generalized gradient approximation (meta-GGA) functionals (SCAN and r²SCAN) is encouraging because (i) the best RF model provides a lower error compared with more sophisticated density functionals, (ii) it is much faster than the self-consistent DFT simulations, especially with meta-GGA functionals, enabling one to screen ΔH_f of materials accurately in a high-throughput fashion. For example, for the 10⁵ materials in large DFT databases such as MP, more accurate predictions of ΔH_f can be calculated by the RF models within minutes, while that from meta-GGA functionals may take months of calculations. For new materials without low-fidelity ΔH_f predictions yet (such as corrected PBE), computational cost for the low-fidelity ΔH_f should be added to the total cost of the best RF model.

As for the superiority of the best RF model over the recent linear correction scheme from Wang et al.,⁴⁷ as shown in Table 1, there are four possible explanations: (i) the RF model takes nonlinear effects into account, (ii) the compositional descriptors used here capture more information than simple stoichiometry used in Wang et al.,⁴⁷ (iii) the learned correction in Wang et al.⁴⁷ is only from materials with certain anions and transition metals,

while in the present work, there is no such constraint, and (iv) the calibration scheme used here is built on empirically corrected PBE results as opposed to uncorrected PBE data in Wang et al.⁴⁷

- (2) Training the machine learning models on ΔH_f^{diff} helps to reduce error compared with training models on ΔH_f^{exp} directly, as, under the same condition (architecture and featurization), the models trained on ΔH_f^{diff} always have lower MAE than that trained on ΔH_f^{exp} . Here, we attribute the lower absolute error of learning ΔH_f^{diff} to the fact that ΔH_f^{diff} has a narrower distribution than ΔH_f^{exp} with 5 times smaller standard deviation (0.17 versus 0.80 eV/atom). One can imagine that, if ΔH_f^{diff} and ΔH_f^{exp} have the same distribution except for a scaling factor of 1/5, then ideally the MAEs of ML models (with proper normalization) trained on ΔH_f^{diff} should also be 1/5 of that trained on ΔH_f^{exp} . However, the MAEs of models trained on ΔH_f^{diff} are all larger than 1/5 of that trained on ΔH_f^{exp} , suggesting that ΔH_f^{diff} is easier to learn absolutely but harder to learn relatively than ΔH_f^{exp} .

To further illustrate the above explanation, we use the R^2 score, a unitless metric invariant to scaling, to show the relative difficulty of predicting ΔH_f^{diff} and ΔH_f^{exp} . The R^2 of predictions of ΔH_f^{diff} by the best RF model is 0.54 (here R^2 of 0.54 is based on predicted ΔH_f^{diff} versus true ΔH_f^{diff} , while the R^2 of 0.99 in Figure 2b is based on predicted ΔH_f^{exp} versus true ΔH_f^{exp}), while the R^2 of predictions of ΔH_f^{exp} by the same RF model is 0.94, suggesting that ΔH_f^{exp} is easier to learn relatively than ΔH_f^{diff} .

- (3) Feeding ΔH_f^{DFT} as one of the input features helps to lower the error, as with the same machine learning architecture (RF or MLP), label, and other features, models with ΔH_f^{DFT} as one of the input features always have lower error than that without ΔH_f^{DFT} . This effect is more significant when the models are trained on ΔH_f^{exp} because, as shown in Figure 1c, ΔH_f^{DFT} has a strong correlation with ΔH_f^{exp} , while, as shown in Figure 1d, the correlation between ΔH_f^{DFT} and ΔH_f^{diff} is not obvious.

Combining analyses (2) and (3), one can observe that adapting the strategy of multifidelity machine learning might help to significantly lower prediction error if the difference between the different fidelity datasets has a narrower distribution than the high-fidelity dataset and/or if there is a strong correlation between the different fidelity data sets. Machine learning models with both the modifications of changing labels and adding extra input features might outperform that with either single modification.

- (4) Similar to (3), transfer learning helps more when transferring from ΔH_f^{DFT} to ΔH_f^{exp} than from ΔH_f^{DFT} to ΔH_f^{diff} because of the stronger correlation between ΔH_f^{DFT} and ΔH_f^{exp} .
- (5) RF with human-engineered features performs better than ROOST and CGCNN, two deep representation learning models, when trained on ΔH_f^{diff} , while RF performs similar to or worse than neural network-based models when trained on ΔH_f^{exp} . Although it is not surprising that neural network-based deep learning algorithms do not show superior performance over RF due to the limited dataset size,^{52,60} the effect of learning targets (ΔH_f^{diff} and ΔH_f^{exp}) on the relative performance of different machine learning models is interesting and worth discussion.

The different uncertainty levels between ΔH_f^{diff} and ΔH_f^{exp} might help to explain why RF performs better than neural network-based models when trained on ΔH_f^{diff} , while there is no such superiority of RF when trained on ΔH_f^{exp} . As discussed above, ΔH_f^{diff} has a narrower distribution than ΔH_f^{exp} . Because $\Delta H_f^{\text{diff}} = \Delta H_f^{\text{exp}} - \Delta H_f^{\text{DFT}}$, if we assume ΔH_f^{exp} and ΔH_f^{DFT} are two independent random variables, then ΔH_f^{diff} would have larger uncertainty than ΔH_f^{exp} . Therefore, the robustness of RF against uncertainty^{60–62} might explain the superiority of RF when trained on ΔH_f^{diff} . The larger uncertainty level of ΔH_f^{diff} might also help to explain why ΔH_f^{diff} is harder to learn relatively than ΔH_f^{exp} as in (2).

To further investigate the effect of uncertainty on the performance of machine learning models, RF and ROOST are employed to learn ΔH_f^{exp} with random noises, a source of uncertainty. In Figure 2a, one can see that RF performs worse than ROOST when trained on ΔH_f^{exp} . In Figure 2c, the MAEs of RF and ROOST and the corresponding noise levels are shown. One can see that, under low noise levels, the errors of RF are still higher than that of ROOST, while under high noise levels, the errors of RF become lower than that of ROOST. The different relative performances of RF and ROOST under different noise levels agree with the superiority of RF against uncertainty^{60–62} and support our hypothesis that the different uncertainty levels of the ΔH_f^{diff} dataset and the ΔH_f^{exp} dataset might explain why RF is better on the ΔH_f^{diff} dataset while ROOST is better on the ΔH_f^{exp} dataset.

In Figure 2e, we plot the learning curves of RF and ROOST on learning ΔH_f^{diff} and ΔH_f^{exp} , respectively. For learning ΔH_f^{exp} , we observe that with a few data points, RF has smaller errors than ROOST, while with more than 400 data points, ROOST outperforms RF, which agrees with previous observations^{26,63} that deep learning is powerful for large datasets, while classic machine learning is more suitable for small datasets. However, for learning ΔH_f^{diff} , we observe that RF performs better than ROOST consistently for all dataset sizes. As for the rate of improvement with respect to dataset size (slope of learning curve), we observe that for RF, the slope on learning ΔH_f^{diff} is slightly smaller than that on ΔH_f^{exp} , while for ROOST, the slope on learning ΔH_f^{diff} is significantly smaller than that on ΔH_f^{exp} , which shows that the slopes of learning curves of machine learning models are affected by the quality of data: higher uncertainty of data, smaller slope of learning curves, and different machine learning models are affected differently: the slope of RF is less affected, while the slope of ROOST is more affected. Further empirical and theoretical studies are necessary to investigate the relation between data quality and the slope of the learning curve for different machine learning models. From the learning curves, we also expect that, with more ΔH_f^{exp} data points in the future, learning ΔH_f^{exp} directly by ROOST might be more powerful than learning ΔH_f^{diff} by random forest.

Based on the fact that when trained on ΔH_f^{diff} , random forest with human-engineered featurization outperforms neural network-based models, especially deep representation learning models, we suggest that for machine learning applications in the field of materials science, with limited data set size and without proof of a low uncertainty level of the data set, deep neural network-based representation learning algorithms^{24,26,32,64} should not be the only type of models employed, and other feature engineering methods and machine learning architectures beyond neural networks should also be tested.

While there are some previous works that show that information on the local bonding environment can be used to

correct formation enthalpies of certain materials like sulfides,⁶⁵ fluorides,⁶⁶ and oxides,^{48,66} in this work, the machine learning models with only compositions as input outperform those with both compositions and structures as input (Figures 2a and S3). One of the possible causes of the phenomenon is that there still lack the data points of polymorphs with the same composition but different ΔH_f^{exp} in the current dataset, which suggests the urgency of building a comprehensive ΔH_f^{exp} dataset with sufficient entries of polymorphs to comprehensively understand the role of structures in determining ΔH_f^{exp} . In Figure 2e, we plot the learning curves for random forest (RF) with and without structural features for learning ΔH_f^{diff} and ΔH_f^{exp} . We can observe that, for learning ΔH_f^{diff} and ΔH_f^{exp} , RF without structural features outperforms RF with structural features, while the slopes of learning curves of the RF models with structural features are larger than that of the RF models without structural features. A possible explanation is that models with structural features have more available information, more degrees of freedom, and therefore easier to overfit small data sets, while those additional information makes models with structural information more powerful and consequently with steeper learning curves. Based on the learning curves, we expect that with more data in the future, models with structural information might outperform those with only compositional information.

For the models only based on compositional information, such as random forest with only compositional features and ROOST, the corrections are the same for the polymorphs. Since the best model in this work is only based on compositional information, in the following sections, the analysis is purely based on compositions. However, there are also models based on structural information trained and tested in this work, such as random forest with compositional and structural features and CGCNN. Corrections from models with structural information are in principle different for polymorphs. In Table S2, we show the corrections from random forest (RF) with both compositional and structural features (“struct. RF diff.” in Figure S3 and Table S1) to three pairs of polymorphs with recorded ΔH_f^{exp} values and not in the training set. We find that, for all materials except CaSiO_3 wol., ΔH_f^{RF} is closer to ΔH_f^{exp} than ΔH_f^{DFT} , showing the ability of the RF model to correct the DFT prediction of ΔH_f . As for relative phase stability, ΔH_f^{DFT} contradicts with ΔH_f^{exp} for SiO_2 and TiO_2 . Unfortunately, for the two systems, corrections from RF cannot reverse the wrong phase stability estimation from DFT. A possible explanation is that the RF model mainly employs compositional information to learn and predict, as we find that compositional features contribute 80% of feature importance, while structural features only contribute 20% of feature importance. Therefore, the RF model predicts similar corrections to different structures with the same composition. More data points of ΔH_f^{exp} , especially that of polymorphs, are necessary to develop machine learning models that rely more on structural information and are capable of reversing the wrong phase stability estimation of polymorphs from DFT.

We summarize the potential drawbacks of using one model versus the others for predicting ΔH_f as follows: (i) RF and MLP rely on off-the-shelf featurization, which means that they cannot capture information unknown to human beings. Therefore, they are typically less powerful than deep representation learning models such as ROOST and CGCNN for large datasets.^{26,63} For predicting ΔH_f , although RF is the best model in this work, with more data points in the future, it is likely that RF will be less powerful than ROOST as shown by the learning curves in Figure

2d. (ii) ROOST and CGCNN are deep representation learning models that learn the features of materials automatically in the training process. Therefore, they are thought to be more powerful than models with off-the-shelf featurization, but their prediction performance might be worse with small datasets,^{26,63} such as this work. (iii) MLP, ROOST, and CGCNN are neural network-based models. Compared with random forest, which is a decision tree-based ensemble model with hundreds of individual models, ensembles of neural networks are typically only composed of around 10 individual models because of the higher computational cost of neural networks.²⁶ Therefore, they might be less powerful than RF in cases where the number of models in ensembles is important,⁶² such as the ΔH_f^{diff} with high uncertainty in this work.⁶²

Discovering Materials with Underestimated Stability in MP

With the best RF model that can significantly lower the error of ΔH_f from the MP database, we can calibrate ΔH_f of all materials in the MP database. The dataset with all of the calibrated ΔH_f is provided in the Supporting Information, and as an application, here we use the calibrated ΔH_f to re-evaluate the thermodynamic stability of all materials in the MP database by constructing the energy above hull (E_{hull} , the energy difference between the candidate compound and the ground-state phase(s) in a compositional space.⁶⁷ More discussions about E_{hull} are provided in the Methods section and Figure S3). However, as Bartel et al.¹⁸ pointed out, although sometimes DFT has large errors for the prediction of ΔH_f , ΔH_f^{DFT} of similar materials contain similar systematic errors, and when evaluating phase stability, the cancellation of systematic errors makes DFT more useful for evaluating the relative stability between compounds than some machine learning models with similar or even better accuracy with respect to ΔH_f^{exp} .

Therefore, before screening E_{hull} for the full MP dataset, we first evaluate the performance of ΔH_f^{DFT} and ΔH_f^{ML} for evaluating the relative stability between compounds. Since there are only 229 materials in the test set, which are not enough for constructing phase diagrams and E_{hull} , we use the difference between ΔH_f of pairs of compounds in the same chemical system to evaluate the relative stability between compounds. We list all 20 pairs of compounds in the same chemical system in the test set in Table 2, and we also plot the difference from experiments versus that from MP and machine learning (ML) in Figure 3a. One can see that ML outperforms MP in terms of the difference of ΔH_f between compounds in the same chemical system, which shows that the ML model outperforms DFT for relative stability evaluation.

We next re-evaluate material stability using ML calibrated ΔH_f to construct $E_{\text{hull}}^{\text{ML}}$ for all materials in the MP database using all compositions in MP. In chemical intuition, materials with smaller E_{hull} tend to be more thermodynamically synthesizable and stable,^{19,20,68} although $E_{\text{hull}} = 0$ is not a hard threshold for successful synthesis and room-temperature and pressure stabilities of materials because of other factors such as kinetics,⁶⁹ and in practice, empirical heuristics of several room-temperature $k_B T$ are used as stability thresholds.^{19,20,68} In Figure 3b, the distributions of E_{hull} of all materials in the MP database constructed from ΔH_f^{DFT} and ΔH_f^{ML} of all compositions in the MP database are shown, from which one can see that most materials have similar $E_{\text{hull}}^{\text{MP}}$ and $E_{\text{hull}}^{\text{ML}}$, and majority of materials have close-to-zero $E_{\text{hull}}^{\text{MP}}$ and $E_{\text{hull}}^{\text{ML}}$. More importantly, there are materials with large $E_{\text{hull}}^{\text{MP}}$ and small $E_{\text{hull}}^{\text{ML}}$. These materials might have underestimated stabilities in

Table 2. Difference of ΔH_f between Pairs of Compounds in the Same Chemical System from Different Sources^a

pair of compounds	experiment	materials project	machine learning in this work
TiFe ₂ –TiFe	0.0487	–0.1324	–0.1316
BiI ₃ –BiI	0.1075	0.1868	0.1193
LuIr ₂ –LuIr	–0.1502	–0.1664	–0.1826
LaSi–La ₅ Si ₃	0.143	0.1335	0.1229
BMo ₂ –BMo	–0.1858	–0.1856	–0.1972
Na ₂ O–NaO ₂	0.5435	0.5428	0.3328
BW ₂ –B ₃ W ₂	–0.0591	0.5108	0.2408
Co ₃ O ₄ –CoO	0.1229	–0.0302	–0.0553
ZrCo ₂ –Zr ₂ Co	0.0974	0.0574	0.0553
TmAg–TmAg ₂	0.1835	0.0088	0.1187
PrNi ₅ –PrNi	–0.0259	–0.0281	–0.0116
TiAu ₂ –TiAu	0.0179	–0.0026	–0.0243
NdRh–NdRh ₂	0.0446	0.0202	0.0064
CaO ₂ –CaO	–1.0353	–1.1070	–1.075
Zr ₅ Si ₃ –Zr ₃ Si ₄	–0.2094	–0.0855	–0.0964
Zr ₅ Si ₃ –ZrSi ₂	0.1181	0.1654	0.1397
Zr ₅ Si ₄ –ZrSi ₂	0.3275	0.2509	0.2361
Mn ₂ Sb–MnSb	–0.0824	0.3453	0.1428
CrSi–CrSi ₂	0.0090	–0.0783	–0.0280
Mn ₁₁ Si ₁₉ –Mn ₃ Si	0.0596	0.1276	0.0809

^aDifference of ΔH_f is the unit of eV/atom.

MP. For example, there are 800 materials in the blue rectangle in Figure 3b that have $E_{\text{hull}}^{\text{MP}} > 0.16$ eV/atom and $E_{\text{hull}}^{\text{ML}} < 0.06$ eV/atom, among which there are around 100 already synthesized materials. (The thresholds are based on 6 times and 2 times of room-temperature $k_B T$.^{19,20}) As example, we list some materials in Table 3 with novel physical properties and/or potential applications with $E_{\text{hull}}^{\text{MP}} > 0.16$ eV/atom and $E_{\text{hull}}^{\text{ML}} < 0.06$ eV/atom, where there are both synthesized materials and hypothetical materials. One can see that there are a number of materials with various applications ranging from battery electrodes,⁷⁰ catalysts^{71–73} to optical,^{74–76} electronic,^{77,78} magnetic^{79–83} devices, and superconductors^{84,85} for which $E_{\text{hull}}^{\text{ML}}$ succeeds in explaining their synthesizability,

while $E_{\text{hull}}^{\text{MP}}$ suggests a low likelihood of being synthesizable. One extreme example is MnSnIr,⁸⁶ a stable half-Heusler compound synthesized from a peritectic reaction,⁸⁷ of which $E_{\text{hull}}^{\text{MP}}$ is considerably high (0.5117 eV/atom), while $E_{\text{hull}}^{\text{ML}}$ is 0. The huge gap between $E_{\text{hull}}^{\text{MP}}$ and $E_{\text{hull}}^{\text{ML}}$ is mainly because of the large deviation between ΔH_f^{DFT} (0.2945 eV/atom) and ΔH_f^{ML} (–0.2363 eV/atom) of MnSnIr itself. As a comparison, the ΔH_f^{exp} of MnSnIr is –0.3047 eV/atom,¹⁷ which shows that, for this compound, DFT deviates significantly from the experiment, while our machine learning model can calibrate such a huge difference. A possible reason for the large error of ΔH_f^{DFT} of MnSnIr is that, in Materials Project (V2021.03.22), the DFT + U correction is only applied to Mn–F, Mn–O, and Mn–S systems and not applied to systems without F, O, S such as the full ternary compound MnSnIr.⁴⁶ Large deviations between ΔH_f^{DFT} and ΔH_f^{exp} are also observed for other compounds containing Mn and Sn, such as MnSnAu (ΔH_f^{DFT} : –0.0488 eV/atom; ΔH_f^{exp} : –0.5016 eV/atom), MnSn₂ (ΔH_f^{DFT} : 0.1363 eV/atom; ΔH_f^{exp} : –0.0954 eV/atom), and Mn₂SnRu (ΔH_f^{DFT} : 0.0789 eV/atom; ΔH_f^{exp} : –0.1803 eV/atom), which agrees with the observation in Figure 4b shown later that DFT tends to overestimate ΔH_f (more positive) of compounds with Mn and Sn. As a result, in the phase diagram of Mn–Sn, there are no stable intermetallic compounds according to ΔH_f^{DFT} , which disagrees with the experimental phase diagram where there are several stable intermetallics including Mn₃Sn, Mn₃Sn₂, and MnSn₂.⁸⁸

In addition to the already synthesized materials, those unrealized hypothetical materials provide potential opportunities for energy and environmental materials,^{89–91} structural materials,⁹² and electronic devices,^{93,94} and as shown in Table 3 and Figure 3b, many of these materials that are estimated stable by $E_{\text{hull}}^{\text{ML}}$ might have underestimated stability in the MP database.

There are also 1000 materials in the lower-right corner in Figure 3b that have $E_{\text{hull}}^{\text{MP}} < 0.06$ eV/atom and $E_{\text{hull}}^{\text{ML}} > 0.16$ eV/atom. Details of those materials can be obtained in the shared online dataset. An extreme example is LiNbGeO₅,⁹⁵ a

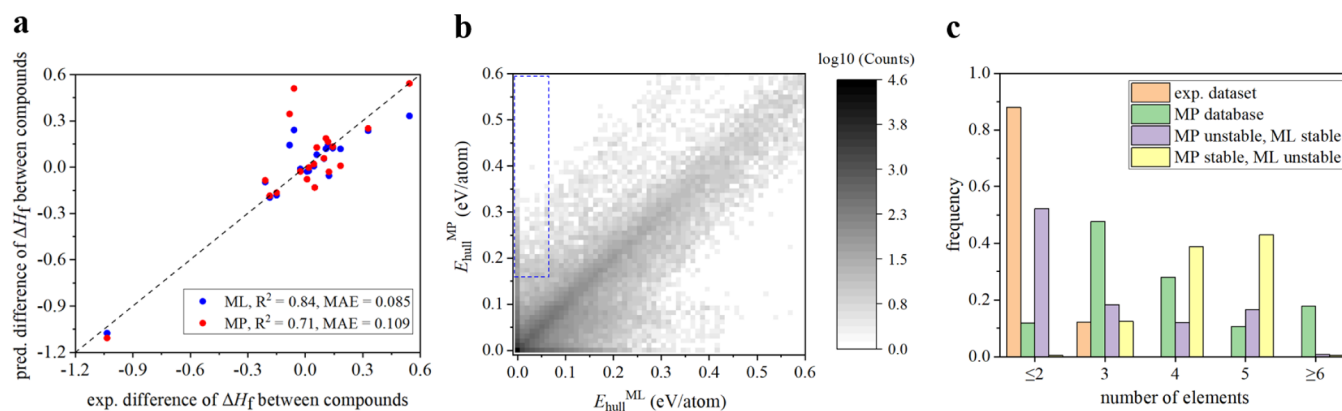


Figure 3. Stability evaluation from energy above hull. (a) Difference of ΔH_f between pairs of compounds in the same chemical system from experiments versus that from MP and machine learning. (b) Distribution of energy above hull (E_{hull} , in eV/atom) of all materials in the Materials Project² database calculated by the corrected PBE ΔH_f in MP ($E_{\text{hull}}^{\text{MP}}$) versus that calculated by the machine learning ΔH_f in this work ($E_{\text{hull}}^{\text{ML}}$). Here, E_{hull} is constructed from all materials in the Materials Project database. The color scheme is used to show the (\log_{10}) number of materials within a range of certain $E_{\text{hull}}^{\text{ML}}$ and $E_{\text{hull}}^{\text{MP}}$, and the blue rectangle shows the area with $E_{\text{hull}}^{\text{MP}} > 0.16$ eV/atom and $E_{\text{hull}}^{\text{ML}} < 0.06$ eV/atom. (c) Appearance frequencies of number of elements of each material in the data sets. Here, “exp. dataset” is the ΔH_f^{exp} used in this work, “MP database” is the set of all materials in the Materials Project database, “MP unstable, ML stable” is the set of materials with $E_{\text{hull}}^{\text{MP}} > 0.16$ eV/atom and $E_{\text{hull}}^{\text{ML}} < 0.06$ eV/atom, and “MP stable, ML unstable” is the set of materials with $E_{\text{hull}}^{\text{MP}} < 0.06$ eV/atom and $E_{\text{hull}}^{\text{ML}} > 0.16$ eV/atom.

Table 3. Examples of Materials That Have Novel Physical Properties and/or Potential Applications with $E_{\text{hull}}^{\text{MP}} > 0.16$ eV/Atom and $E_{\text{hull}}^{\text{ML}} < 0.06$ eV/Atom^a

materials	MP ID	$E_{\text{hull}}^{\text{MP}}$	$E_{\text{hull}}^{\text{ML}}$	characterization method(s)	comment/novel physical property/potential application
MnSnIr	mp-11480	0.5117	0	experiment	largest difference between $E_{\text{hull}}^{\text{MP}}$ and $E_{\text{hull}}^{\text{ML}}$
Ta ₃ Pb	mp-1187214	0.3386	0	experiment	superconductor ⁸⁵
AgRh	mp-1183233	0.2633	0.0359	experiment	electrocatalyst ⁷¹
FeCoSn	mp-1025124	0.1836	0.0384	experiment	tuning phase transitions for isostructural alloying ⁹⁶
SmCo ₄ Ag	mp-1219086	0.1797	0.0493	experiment	positively correlated magnetization with temperature ⁷⁹
Li ₃ (FeS ₂) ₂	mp-753818	0.1697	0.0180	experiment	Li-FeS ₂ battery electrode ⁷⁰
PdRu	mp-1186459	0.2277	0.0032	experiment	catalyst ⁷²
Ni ₃ Ag	mp-1100764	0.2332	0	experiment	dual-frequency absorption ⁷⁴
Rb ₂ NaTaF ₆	mp-1114459	0.2038	0	experiment	large anisotropic shift from both covalent and polarization spin transfer mechanisms ⁸⁰
Nb ₃ Tl	mp-569366	0.2083	0	experiment	superconductor ⁸⁴
UPb ₃	mp-1184128	0.1621	0	experiment	sharp metamagnetic transitions ⁸¹
Cu ₃ N	mp-1933	0.1865	0.0464	experiment	light recording media ⁷⁵
FeNi ₂	mp-1072076	0.1858	0.0292	experiment	size-dependent catalytic activity ⁷³
HfCo ₇	mp-1105489	0.2098	0.0500	experiment	rare-earth-free permanent magnets ⁸²
MnBi	mp-1185989	0.2078	0	experiment/DFT	half-metallic ferromagnetism ⁷⁸
Be ₂ Si	mp-1009829	0.2352	0.0272	experiment/DFT	hybrid nodal-line semimetal ⁷⁷
Mn ₂ Hg ₅	mp-30720	0.2362	0	experiment/DFT	π -based covalent magnetism ⁸³
Ta ₃ Bi	mp-1187199	0.3442	0	DFT	topological dirac semimetal ⁹³
MnCrSb	mp-1221652	0.2564	0	DFT	half-metallicity ⁹⁴
LiB ₁₁	mp-1180507	0.2084	0.0234	DFT	pseudo-plasticity ⁹²
NiAg ₃	mp-976762	0.1850	0	DFT	acetylene adsorbent ⁹¹
Li ₂ VN ₂	mp-1246112	0.1615	0.0279	DFT	Li-ion battery electrode ⁸⁹
LiGdO ₃	mp-1185401	0.3476	0.0575	machine learning	perovskite with high tolerance factor ⁹⁰
LiPmO ₃	mp-1185388	0.2815	0	machine learning	perovskite with high tolerance factor ⁹⁰

^aThe materials with the experiment as one of the characterization methods are synthesized materials, and others are currently only hypothetical. E_{hull} is in the unit of eV/atom.

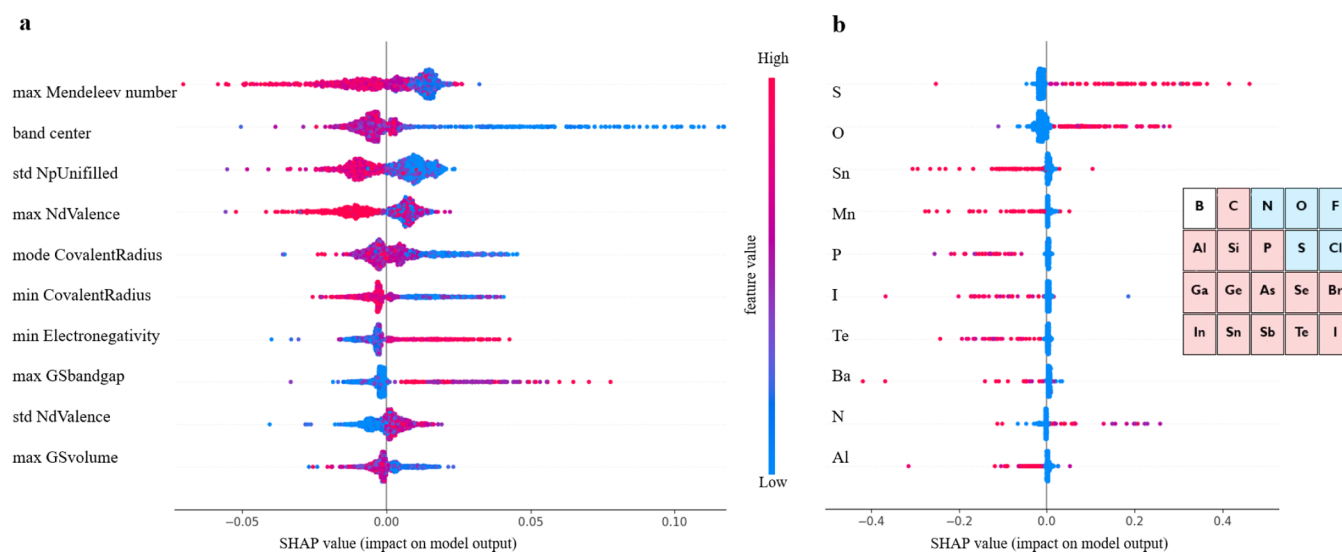


Figure 4. Impact of each feature on model output. (a, b) Distributions of the impacts (SHAP values⁹⁷) of compositional features and elemental fractions on the model output (ΔH_f^{diff}), respectively. The color represents the feature value (red high, blue low), and here only the top 10 features and elemental fractions with the highest sum of absolute SHAP values are shown. The inset figure in panel (b) illustrates the trends of DFT to underestimate or overestimate ΔH_f of materials with certain nonmetal elements. Here, the blue-shaded elements are those for which DFT tends to underestimate ΔH_f , the red-shaded elements are those for which DFT tends to overestimate ΔH_f and boron shows a mixed trend.

synthesized compound with $E_{\text{hull}}^{\text{MP}}$ of 0 and $E_{\text{hull}}^{\text{ML}}$ of 0.4334 eV/atom.

To further investigate how MP and ML disagree with each other, the appearance frequencies of the number of elements in each material in four data sets are plotted in Figure 3c. One can see that in the exp. dataset used as the training set in this work,

around 90% of the materials are binary compounds and 10% of the materials are ternary, while in the MP database, there are about 40% contain more than three elements. Since the training set does not cover material space with more than three elements, the ML predictions for materials with more than three elements are extrapolations and in general less reliable than that for binary

and ternary compounds. For the set of materials unstable by MP and stable by ML, the distribution of the number of elements is similar to that of the exp. dataset where the majority of materials are binary or ternary, while in the set of materials stable by MP and unstable by ML, most materials have four or five elements. Here, the lack of materials with more than three elements in the current ΔH_f^{exp} dataset suggests that the ML predictions for materials with more than three elements should be carefully checked if ML and MP disagree with each other, and it also suggests the urgency of building a comprehensive ΔH_f^{exp} dataset with sufficient entries of materials with more than three elements.

Data Mining Where ΔH_f^{DFT} Fails by Explaining the Multifidelity Model

In addition to predicting more accurate ΔH_f and examining the stability of materials, the random forest model trained on ΔH_f^{diff} ($\Delta H_f^{\text{exp}} - \Delta H_f^{\text{DFT}}$) with human-engineered features can also serve as a data-mining approach to reveal where and how ΔH_f^{DFT} deviates from ΔH_f^{exp} (as above, ΔH_f^{DFT} refers to the empirically corrected PBE ΔH_f by Jain et al.⁴⁶ in the Materials Project database), which provides clearer trends than machine learning models trained on ΔH_f^{DFT} only. Here, we analyze the relationship between human-understandable features and ΔH_f^{diff} by explaining the model, or for each material, calculating the impact of each feature on the model output (known as the SHAP value⁹⁷). Previously, the error of ΔH_f^{DFT} is mostly discussed in the context of certain anions,^{3,41,46} cations,³ and transition metals.^{3,41,42} In Figure 4a, the impacts of the top 10 compositional features from matminer⁵⁹ with the highest sum of absolute SHAP values are shown. One can see that, in addition to anion properties (“max GSbandgap”; the detailed explanations of the descriptors are available in the matminer paper⁵⁹) and cation properties (“max GSvolume”, “max NdValence”, “min CovalentRadius”, “min Electronegativity”), the mean field of elemental properties (“band center”, “mode CovalentRadius”) and standard deviation of elemental properties (“std NpUnifilled”, “std NdValence”) are also among the most impactful properties with respect to ΔH_f^{diff} . For example, with smaller band center (geometric mean of electronegativity⁵⁹), ΔH_f^{diff} tends to be larger and ΔH_f^{DFT} tends to be smaller than ΔH_f^{exp} , which means that DFT tends to underestimate ΔH_f of systems with a smaller mean electronegativity. A possible explanation for this trend is that, for a smaller geometric mean of electronegativity, the ability of atoms to bind the electrons near the atomic nuclei is weaker, and electrons tend to be more delocalized. Since the GGA approximation tends to overestimate the electron delocalization,⁹⁸ ΔH_f^{DFT} tends to be more negative for the systems with delocalized bonds (stronger bonding). Another example is with a larger standard deviation of the number of p valence electrons, ΔH_f^{diff} tends to be smaller and ΔH_f^{DFT} tends to be larger than ΔH_f^{exp} , suggesting that DFT tends to overestimate ΔH_f of systems with greater differences among p electrons. This trend might be explained by the hypothesis that, with more different p electron configurations, in general, the compound is more ionic, and because of the fact that the GGA approximation tends to underestimate the electron localization,⁹⁸ DFT (PBE) ΔH_f tends to be more positive for the systems with localized bonds (weaker bonding).

As for the impacts of certain cations and anions, or impacts of certain elements, we build a decision tree model that takes stoichiometry as input, and the SHAP values of the fraction of each element are plotted in Figure 4b. One can see that, with a

higher atomic fraction of S, O, and N, DFT tends to underestimate ΔH_f , while for a higher atomic fraction of Sn, Mn, P, I, Te, Ba, and Al, DFT tends to overestimate ΔH_f . There are more nonmetal elements (6) in the top 10 most impactful elements than metals (2) and metalloids (2). Particularly, there is an interesting pattern of how DFT treats different nonmetal elements: as shown in Figures 4b and S2, for strong oxidizing nonmetal elements in the upper-right corner of the periodic table, including F, O, N, S, and Cl, DFT tends to underestimate ΔH_f , while for those nonmetal elements with weaker oxidizing ability, DFT tends to overestimate ΔH_f . However, the degree of overestimation or underestimation does not simply correlate with the oxidizing ability. For example, as shown in Figures 4b and S1, F has a stronger oxidizing ability than O and S, but the degree of underestimation of ΔH_f^{DFT} for fluorides is less than that of oxides and sulfides. There are two possible sources of errors that would result in the observed trend: on the one hand, the underestimation or overestimation of ΔH_f of materials with certain elements might come from the element type-based empirical corrections,^{3,46} and, on the other hand, the intrinsic limit of the GGA and GGA + *U* approximation might cause the different deviation patterns. For example, Seo et al.⁹⁹ have proposed that the GGA + *U* method used for transition-metal oxides in the MP database⁴⁶ overestimates the degree of hybridization between the d orbitals of transition metals and p orbitals of oxygen and thus making the calculated ΔH_f more negative.

The trend in Figure 4a also agrees with that in Figure 4b. For example, for max GSbandgap and max GSvolume, they are calculated in the following procedure: first, the ground-state band gaps and ground-state volumes of all of the elements in the compound are listed, and then the maximum values of band gaps and volumes are picked up. Therefore, max GSbandgap and max GSvolume actually relate to the existence of certain elements in the compound. Specifically, max GSbandgap describes the presence of a specific anion in the compound, while max GSvolume describes that of cation. Larger max GSvolume, ΔH_f^{DFT} tends to be larger (more positive) than ΔH_f^{exp} . An explanation for this trend is that with a larger max GSvolume, the cation element tends to have a larger ground-state volume (closer to the bottom-left of the periodic table with the maximum value at Cs). If the cation is closer to the bottom-left of the table, the compound in general will be more ionic. Therefore, ΔH_f^{DFT} tends to be more positive for the systems with more ionic bonds, as mentioned above. On the other hand, for larger max GSbandgap, ΔH_f^{DFT} tends to be smaller (more negative) than ΔH_f^{exp} . This phenomenon might be explained by the fact that, with a larger max GSbandgap, the anionic element is closer to the upper-right corner of the periodic table, the maximum value at N, and according to Figure 4b, the compound tends to have more negative ΔH_f^{DFT} .

In Wang et al.⁴⁷ all anionic corrections are negative, which is because their correction is applied to the original PBE results and PBE tends to overestimate the energy of diatomic gas molecules,¹⁰⁰ while the trend shown here is based on the empirically corrected PBE energies from MP that already take the effect of the overestimated energy of diatomic gas molecules into account.

DISCUSSION AND CONCLUSIONS

In this work, we conduct a comprehensive machine learning study to learn and predict the experimental formation enthalpy of materials. We use two different strategies to transfer

information from a larger DFT dataset to a smaller experimental dataset, transfer learning and multifidelity machine learning, and we use four machine learning architectures to realize the two strategies (RF, MLP, ROOST, CGCNN). We find that the random forest (RF) model trained on the difference between experimental and DFT formation enthalpies with the DFT formation enthalpy as one of the input features can achieve the lowest error, which is almost half of that of DFT (empirically corrected PBE), and it also outperforms other more accurate but more computationally expensive density functionals, such as meta-GGA functionals. Beyond identifying the best model, we suggest that the deep neural network-based representation learning algorithms and transfer learning should not be the only machine learning architecture and information transfer strategy considered. Other feature engineering methods such as human-engineered features, machine learning architectures beyond neural networks such as random forest, and information transfer strategy such as multifidelity machine learning should also be tested in machine learning applications for materials science.

As an application, we employ the best-found random forest model to calibrate the formation enthalpy of all materials in the Materials Project database, which are then used to construct the energy above hull and discover potentially important materials that have an underestimated stability in the MP database. Further, we use the machine learning model as a data-mining approach to identify patterns in the performance of DFT, for example, in its tendency to underestimate the formation enthalpy of materials with elements in the upper-right corner of the periodic table.

This work is based on the Materials Project database queried on March 2021 (V2021.03.22). The methodology of this work can also be applied to the updated Materials Project database (such as V2021.05.13) and other large DFT databases. It is expected that, with more accurate low-fidelity data (DFT formation enthalpy), such as the very recent datasets by Kingsbury et al.⁴⁴ and by Schmidt et al.¹⁰¹ with thousands of materials calculated by meta-GGA functionals, the method in this work can be used to provide more accurate calibration (exp. formation enthalpy).

One potential limitation of the multifidelity model used in this work is that it requires the availability of low-fidelity data for the whole material space of interest, as in this work, DFT formation enthalpy is required for learning the difference of formation enthalpy from the experiment and DFT. In cases where low-fidelity data is not available to all of the materials, transfer learning might be more appropriate to transfer information between different data sets. Another scenario not considered in the current multifidelity machine learning scheme is that, for some properties, there might be datasets with multiple levels of fidelity available. In such cases, in addition to incorporating different fidelity data into the input, the learning of differences might be conducted multiple times to enlarge the availability of high-fidelity data gradually.

More broadly, for machine learning applications with small datasets, choosing proper models and strategies is critical to the usefulness of the machine learning models. On the one hand, with small datasets, one should carefully compare the performance of deep representation learning and classic machine learning models based on off-the-shelf featurization and make the choice for production. Typically, with more than 10,000 data points, deep representation learning might be more powerful; with less than 500 data points, classic machine learning models might be more suitable; and with more than 500 but less than

10,000 data points, a careful comparison is necessary for employing a suitable model for production. On the other hand, if larger low-fidelity data sets are available, then information transfer might be useful to improve the learning and prediction of the high-fidelity data. There are two strategies, transfer learning and multifidelity learning, for the information transfer. Although there still lacks a theoretical guarantee or quantitative metric to estimate whether information transfer would help or not, empirically, the two strategies are worth trying if the high- and low-fidelity datasets are strongly correlated.

METHODS

Data Collection

In this work, we construct the ΔH_f^{exp} dataset by combining two sources from IIT¹⁷ and SSUB,⁵⁸ and we use the Materials Project² database (V2021.03.22) to construct the ΔH_f^{DFT} dataset. For the ΔH_f^{diff} dataset, since the ΔH_f^{DFT} values are provided for some materials in the IIT dataset, ΔH_f^{diff} values for those materials are obtained by subtracting the provided ΔH_f^{DFT} from the provided ΔH_f^{exp} , and for materials from the SSUB dataset, since the chemical formula is the only identifier, we take the lowest ΔH_f^{exp} for each formula, and for the ΔH_f^{DFT} of these materials, we assign the lowest ΔH_f^{DFT} to each formula. For overlaps between the IIT dataset and SSUB dataset, we take the ΔH_f^{exp} from the IIT database as the IIT database is a more recent one.¹⁷ Note that the mean absolute difference of ΔH_f^{exp} between our dataset and the recent dataset from Wang et al.⁴⁷ is only 0.007 eV/atom. Codes and a step-by-step instruction for constructing the dataset are provided in the [Supporting Information](#).

Machine Learning Model Training Procedure

In this work, the dataset of the 1143 ΔH_f^{exp} is used for three purposes: (1) hyperparameter tuning for each machine learning model, (2) model evaluation, and (3) production or prediction of ΔH_f of all materials in the Materials Project database (MP). For purposes (1) and (2), we first randomly reserve 20% data as the test set for model selection (these 20% data are also excluded in the larger MP dataset for transfer learning). Then, to determine the best set of hyperparameters for each model, with the remaining 80% data, we randomly reserve 20% of the remaining data ($20\% \times 80\% = 16\%$ of total data) as the validation set to evaluate each specific set of hyperparameters and use 80% of the remaining data ($80\% \times 80\% = 64\%$ of total data) to train the machine learning model with the given set of hyperparameters. We screen hyperparameters by grid search, and tables of search space of hyperparameters are provided in the [Supporting Information](#). Finally, with the best-found hyperparameters for each model, we use the 80% of the data (training set + validation set in the hyperparameter search step) to train machine learning models 10 times with different initializations and evaluate model performance and uncertainty using the 20% data held out at the very beginning (test set). For purpose (3), production, for the best prediction performance, all available 1143 data points are used to train the best-found model with the best-found hyperparameter.

In this work, we use four different machine learning architectures to realize transfer learning and/or multifidelity machine learning, random forest (RF), multilayer perceptron (MLP), Representation Learning from Stoichiometry (ROOST),²⁶ and Crystal Graph Convolutional Neural Network (CGCNN).³² For ROOST, we feed the compositions of materials as input, and it learns the representations of materials, and for CGCNN, we feed the 3D atomic structures of materials as input, and it also learns the representations. RF and MLP are realized by scikit-learn,¹⁰² and we use the descriptors from matminer⁵⁹ to feed RF and MLP as features of materials. Modules used to generate compositional features are ElementProperty, ElectronAffinity, BandCenter, CohesiveEnergy, Miedema, TMetalFraction, ValenceOrbital, and YangSolidSolution, and modules used to generate structural features are GlobalSymmetryFeatures, StructuralComplexity, ChemicalOrdering, MaximumPackingEfficiency, MinimumRelativeDistances, StructuralHeterogeneity, AverageBondLength, AverageBondAngle, BondOrientationalParameter, CoordinationNumber, and DensityFeatures.

Since the dataset size is not large in this work, training of machine learning models is not time-consuming. Using a PC, for RF and MLP, the training time is within 10 min, and for ROOST and CGCNN, the training time is within an hour. However, if transfer learning is employed, then the training of models on the large DFT dataset is very time-consuming, requiring days of training time for all of the three neural network models. For the speed of prediction, although it is critical in machine learning-based interatomic potentials,¹⁰³ for the prediction of properties, it is typically not a crucial factor in choosing models. After initialization of features or graphs, which are time-consuming (hours of time) but reusable for multiple runs and tasks, all of the models in this work cost less than an hour to predict ΔH_f of $\sim 100,000$ materials in the Materials Project database.

Energy above Hull

In the Materials Project (MP), the energy above hull (E_{hull}) is defined as the energy of decomposition of a material into the set of most stable materials at this chemical composition.² The decomposition is tested against all potential chemical combinations that result in the material's composition. A positive E_{hull} indicates that this material is unstable with respect to decomposition, and a zero E_{hull} indicates that this compound is stable with respect to decomposition. In this work, the energy above hull is defined in the same way as MP. A graphical illustration of E_{hull} is provided in Figure S2. PhaseDiagram module in Pymatgen¹⁰⁴ is used to calculate the E_{hull} . The inputs required by the PhaseDiagram module are the compositions and formation enthalpies, and the corresponding output is the energies vs composition diagram, from which the decomposition energies and E_{hull} can be calculated.

■ ASSOCIATED CONTENT

SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/jacsau.2c00235>.

Impact of top 50 impactful elements on ΔH_f^{diff} , illustration of energy above hull, comparison of errors from predictions of ΔH_f^{exp} , corrections to polymorphs, and search space of hyperparameters (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Yang Shao-Horn — Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; Email: shaohorn@mit.edu

Jeffrey C. Grossman — Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; orcid.org/0000-0003-1281-2359; Email: jcg@mit.edu

Authors

Sheng Gong — Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; orcid.org/0000-0002-7457-7959

Shuo Wang — Department of Materials Science and Engineering, University of Maryland, College Park, Maryland 20742, United States

Tian Xie — Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; orcid.org/0000-0002-0987-4666

Woo Hyun Chae — Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; orcid.org/0000-0001-8385-9491

Runze Liu — Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; orcid.org/0000-0001-6040-6940

Complete contact information is available at: <https://pubs.acs.org/doi/10.1021/jacsau.2c00235>

Author Contributions

S.G., S.W., and J.C.G. conceived the idea, S.G. performed the machine learning analysis, Y.S.H. and J.C.G. supervised this work, and all authors contributed to the analysis of data and writing and editing of the paper.

Notes

The authors declare no competing financial interest.

All datasets, including the dataset of 1143 materials with experimental and MP formation enthalpies, the dataset of 122 materials used for comparison between different functionals with different corrections, and the dataset of 98,338 materials in the MP database with machine learning-predicted formation enthalpies and energy above hull, are provided at: <https://doi.org/10.6084/m9.figshare.19100645.v1>.

All scripts with their requirements and trained machine learning models are provided at: <https://doi.org/10.6084/m9.figshare.19100645.v1>.

■ ACKNOWLEDGMENTS

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

■ REFERENCES

- (1) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133–A1138.
- (2) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, No. 011002.
- (3) Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **2015**, *1*, No. 15010.
- (4) Curtarolo, S.; Setyawan, W.; Wang, S.; Xue, J.; Yang, K.; Taylor, R. H.; Nelson, L. J.; Hart, G. L. W.; Sanvito, S.; Buongiorno-Nardelli, M.; Mingo, N.; Levy, O. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* **2012**, *58*, 227–235.
- (5) Choudhary, K.; Garrity, K. F.; Reid, A. C.; DeCost, B.; Biacchi, A. J.; Walker, A. R. H.; Trautt, Z.; Hatrick-Simpers, J.; Kusne, A. G.; Centrone, A. et al. JARVIS: An Integrated Infrastructure for Data-Driven Materials Design. 2020, arXiv:2007.01831. arXiv.org e-Print archive. <https://arxiv.org/abs/2007.01831>.
- (6) Ahmad, Z.; Xie, T.; Maheshwari, C.; Grossman, J. C.; Viswanathan, V. Machine Learning Enabled Computational Screening of Inorganic Solid Electrolytes for Suppression of Dendrite Formation in Lithium Metal Anodes. *ACS Cent. Sci.* **2018**, *4*, 996–1006.
- (7) Horton, M. K.; Dwaraknath, S.; Persson, K. A. Promises and perils of computational materials databases. *Nat. Comput. Sci.* **2021**, *1*, 3–5.
- (8) Sendek, A. D.; Yang, Q.; Cubuk, E. D.; Duerloo, K.-A. N.; Cui, Y.; Reed, E. J. Holistic computational structure screening of more than 12

000 candidates for solid lithium-ion conductor materials. *Energy Environ. Sci.* **2017**, *10*, 306–320.

(9) Yang, T.; Zhou, J.; Song, T. T.; Shen, L.; Feng, Y. P.; Yang, M. High-Throughput Identification of Exfoliable Two-Dimensional Materials with Active Basal Planes for Hydrogen Evolution. *ACS Energy Lett.* **2020**, *5*, 2313–2321.

(10) Zhou, J.; Shen, L.; Costa, M. D.; Persson, K. A.; Ong, S. P.; Huck, P.; Lu, Y.; Ma, X.; Chen, Y.; Tang, H.; Feng, Y. P. 2DMatPedia, an open computational database of two-dimensional materials from top-down and bottom-up approaches. *Sci. Data* **2019**, *6*, No. 86.

(11) Yan, Q.; Yu, J.; Suram, S. K.; Zhou, L.; Shinde, A.; Newhouse, P. F.; Chen, W.; Li, G.; Persson, K. A.; Gregoire, J. M.; Neaton, J. B. Solar fuels photoanode materials discovery by integrating high-throughput theory and experiment. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, 3040–3043.

(12) Zhu, H.; Hautier, G.; Aydemir, U.; Gibbs, Z. M.; Li, G.; Bajaj, S.; Pöhls, J.-H.; Broberg, D.; Chen, W.; Jain, A.; White, M. A.; Asta, M.; Snyder, G. J.; Persson, K.; Ceder, G. Computational and experimental investigation of TmAgTe₂ and XYZ₂ compounds, a new group of thermoelectric materials identified by first-principles high-throughput screening. *J. Mater. Chem. C* **2015**, *3*, 10554–10565.

(13) Dunstan, M. T.; Jain, A.; Liu, W.; Ong, S. P.; Liu, T.; Lee, J.; Persson, K. A.; Scott, S. A.; Dennis, J. S.; Grey, C. P. Large scale computational screening and experimental discovery of novel materials for high temperature CO₂ capture. *Energy Environ. Sci.* **2016**, *9*, 1346–1360.

(14) Li, S.; Xia, Y.; Amachraa, M.; Hung, N. T.; Wang, Z.; Ong, S. P.; Xie, R.-J. Data-Driven Discovery of Full-Visible-Spectrum Phosphor. *Chem. Mater.* **2019**, *31*, 6286–6294.

(15) Cooley, J. A.; Horton, M. K.; Levin, E. E.; Lapidus, S. H.; Persson, K. A.; Seshadri, R. From Waste-Heat Recovery to Refrigeration: Compositional Tuning of Magnetocaloric Mn₁ + xSb. *Chem. Mater.* **2020**, *32*, 1243–1249.

(16) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865.

(17) Kim, G.; Meschel, S. V.; Nash, P.; Chen, W. Experimental formation enthalpies for intermetallic phases and other inorganic compounds. *Sci. Data* **2017**, *4*, No. 170162.

(18) Bartel, C. J.; Trewartha, A.; Wang, Q.; Dunn, A.; Jain, A.; Ceder, G. A critical examination of compound stability predictions from machine-learned formation energies. *npj Comput. Mater.* **2020**, *6*, No. 97.

(19) Aykol, M.; Dwaraknath, S. S.; Sun, W.; Persson, K. A. Thermodynamic limit for synthesis of metastable inorganic materials. *Sci. Adv.* **2018**, *4*, No. eaaq0148.

(20) Sun, W.; Dacek, S. T.; Ong, S. P.; Hautier, G.; Jain, A.; Richards, W. D.; Gamst, A. C.; Persson, K. A.; Ceder, G. The thermodynamic scale of inorganic crystalline metastability. *Sci. Adv.* **2016**, *2*, No. e1600225.

(21) Yamada, H.; Liu, C.; Wu, S.; Koyama, Y.; Ju, S.; Shiomi, J.; Morikawa, J.; Yoshida, R. Predicting Materials Properties with Little Data Using Shotgun Transfer Learning. *ACS Cent. Sci.* **2019**, *5*, 1717–1730.

(22) Sendek, A. D.; Cubuk, E. D.; Antoniuk, E. R.; Cheon, G.; Cui, Y.; Reed, E. J. Machine Learning-Assisted Discovery of Solid Li-Ion Conducting Materials. *Chem. Mater.* **2019**, *31*, 342–352.

(23) Cubuk, E. D.; Sendek, A. D.; Reed, E. J. Screening billions of candidates for solid lithium-ion conductors: A transfer learning approach for small data. *J. Chem. Phys.* **2019**, *150*, No. 214701.

(24) Schütt, K. T.; Sauceda, H. E.; Kindermans, P. J.; Tkatchenko, A.; Müller, K. R. SchNet - A deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, No. 241722.

(25) Xie, T.; Grossman, J. C. Hierarchical visualization of materials space with graph convolutional neural networks. *J. Chem. Phys.* **2018**, *149*, No. 174111.

(26) Goodall, R. E. A.; Lee, A. A. Predicting materials properties without crystal structure: deep representation learning from stoichiometry. *Nat. Commun.* **2020**, *11*, No. 6280.

(27) Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. Commun.* **2017**, *8*, No. 15679.

(28) Jha, D.; Choudhary, K.; Tavazza, F.; Liao, W. K.; Choudhary, A.; Campbell, C.; Agrawal, A. Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. *Nat. Commun.* **2019**, *10*, No. 5316.

(29) Ye, W.; Chen, C.; Wang, Z.; Chu, I. H.; Ong, S. P. Deep neural networks for accurate predictions of crystal stability. *Nat. Commun.* **2018**, *9*, No. 3800.

(30) Chandrasekaran, A.; Kamal, D.; Batra, R.; Kim, C.; Chen, L.; Ramprasad, R. Solving the electronic structure problem with machine learning. *npj Comput. Mater.* **2019**, *5*, No. 22.

(31) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2016**, *2*, No. 16028.

(32) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120*, No. 145301.

(33) Gong, S.; Xie, T.; Zhu, T.; Wang, S.; Fadel, E. R.; Li, Y.; Grossman, J. C. Predicting charge density distribution of materials using a local-environment-based graph convolutional network. *Phys. Rev. B* **2019**, *100*, No. 184103.

(34) Gong, S.; Wu, W.; Wang, F. Q.; Liu, J.; Zhao, Y.; Shen, Y.; Wang, S.; Sun, Q.; Wang, Q. Classifying superheavy elements by machine learning. *Phys. Rev. A* **2019**, *99*, No. 022110.

(35) Shi, Z.; Tsymbalov, E.; Dao, M.; Suresh, S.; Shapeev, A.; Li, J. Deep elastic strain engineering of bandgap through machine learning. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 4117–4122.

(36) Meredig, B.; Agrawal, A.; Kirklın, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **2014**, *89*, No. 094104.

(37) Jha, D.; Ward, L.; Paul, A.; Liao, W. K.; Choudhary, A.; Wolverton, C.; Agrawal, A. ElemNet: Deep Learning the Chemistry of Materials From Only Elemental Composition. *Sci. Rep.* **2018**, *8*, No. 17593.

(38) Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E.; Constantin, L. A.; Zhou, X.; Burke, K. Restoring the density-gradient expansion for exchange in solids and surfaces. *Phys. Rev. Lett.* **2008**, *100*, No. 136406.

(39) Sun, J.; Ruzsinszky, A.; Perdew, J. P. Strongly Constrained and Appropriately Normed Semilocal Density Functional. *Phys. Rev. Lett.* **2015**, *115*, No. 036402.

(40) Furness, J. W.; Kaplan, A. D.; Ning, J.; Perdew, J. P.; Sun, J. Accurate and Numerically Efficient r(2)SCAN Meta-Generalized Gradient Approximation. *J. Phys. Chem. Lett.* **2020**, *11*, 8208–8215.

(41) Chevrier, V. L.; Ong, S. P.; Armiento, R.; Chan, M. K. Y.; Ceder, G. Hybrid density functional calculations of redox potentials and formation energies of transition metal compounds. *Phys. Rev. B* **2010**, *82*, No. 075122.

(42) Sarmiento-Pérez, R.; Botti, S.; Marques, M. A. Optimized Exchange and Correlation Semilocal Functional for the Calculation of Energies of Formation. *J. Chem. Theory Comput.* **2015**, *11*, 3844–3850.

(43) Zhang, Y.; Kitchaev, D. A.; Yang, J.; Chen, T.; Dacek, S. T.; Sarmiento-Pérez, R. A.; Marques, M. A. L.; Peng, H.; Ceder, G.; Perdew, J. P.; Sun, J. Efficient first-principles prediction of solid stability: Towards chemical accuracy. *npj Comput. Mater.* **2018**, *4*, No. 9.

(44) Kingsbury, R.; Gupta, A. S.; Bartel, C. J.; Munro, J. M.; Dwaraknath, S.; Horton, M.; Persson, K. A. Performance comparison of r2SCAN and SCAN metaGGA density functionals for solid materials via an automated, high-throughput computational workflow. *Phys. Rev. Mater.* **2022**, *6*, No. 013801.

(45) Egorova, O.; Hafizi, R.; Woods, D. C.; Day, G. M. Multifidelity Statistical Machine Learning for Molecular Crystal Structure Prediction. *J. Phys. Chem. A* **2020**, *124*, 8065–8078.

(46) Jain, A.; Hautier, G.; Moore, C. J.; Ping Ong, S.; Fischer, C. C.; Mueller, T.; Persson, K. A.; Ceder, G. A high-throughput infrastructure

- for density functional theory calculations. *Comput. Mater. Sci.* **2011**, *50*, 2295–2310.
- (47) Wang, A.; Dwaraknath, S.; Ong, S. P.; Jain, A.; Horton, M.; McDermott, M.; Kingsbury, R.; Wang, A. A Framework for Quantifying Uncertainty in DFT Energy Corrections. *Sci. Rep.* **2021**, *11*, No. 15496.
- (48) Friedrich, R.; Usanmaz, D.; Oses, C.; Supka, A.; Fornari, M.; Buongiorno Nardelli, M.; Toher, C.; Curtarolo, S. Coordination corrected ab initio formation enthalpies. *npj Comput. Mater.* **2019**, *5*, No. 59.
- (49) Zhang, Y.; Ling, C. A strategy to apply machine learning to small datasets in materials science. *npj Comput. Mater.* **2018**, *4*, No. 25.
- (50) Xie, T.; Bapst, V.; Gaunt, A. L.; Obika, A.; Back, T.; Hassabis, D.; Kohli, P.; Kirkpatrick, J. Atomistic Graph Networks For Experimental Materials Property Prediction. 2021, arXiv:2103.13795. arXiv.org e-Print archive. <https://arxiv.org/abs/2103.13795>.
- (51) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **2019**, *10*, No. 2903.
- (52) Zhu, T.; He, R.; Gong, S.; Xie, T.; Gorai, P.; Nielsch, K.; Grossman, J. C. Charting lattice thermal conductivity for inorganic crystals and discovering rare earth chalcogenides for thermoelectrics. *Energy Environ. Sci.* **2021**, *14*, 3559–3566.
- (53) Kong, S.; Guevarra, D.; Gomes, C. P.; Gregoire, J. M. Materials representation and transfer learning for multi-property prediction. *Appl. Phys. Rev.* **2021**, *8*, No. 021409.
- (54) Chen, C.; Zuo, Y.; Ye, W.; Li, X.; Ong, S. P. Learning properties of ordered and disordered materials from multi-fidelity data. *Nat. Comput. Sci.* **2021**, *1*, 46–53.
- (55) Paliana, G.; Gubernatis, J. E.; Lookman, T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Comput. Mater. Sci.* **2017**, *129*, 156–163.
- (56) Patra, A.; Batra, R.; Chandrasekaran, A.; Kim, C.; Huan, T. D.; Ramprasad, R. A multi-fidelity information-fusion approach to machine learn and predict polymer bandgap. *Comput. Mater. Sci.* **2020**, *172*, No. 109286.
- (57) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Delta-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087–2096.
- (58) Hurtado, I.; Neuschütz, D. *Thermodynamic Properties of Inorganic Materials, Compiled by SGTE*; Springer: Berlin, 1999; Vol. 19.
- (59) Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N. E. R.; Bajaj, S.; Wang, Q.; Montoya, J.; Chen, J.; Bystrom, K.; Dylla, M.; Chard, K.; Asta, M.; Persson, K. A.; Snyder, G. J.; Foster, I.; Jain, A. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **2018**, *152*, 60–69.
- (60) Liu, T.; Abd-Elrahman, A.; Morton, J.; Wilhelm, V. L. Comparing fully convolutional networks, random forest, support vector machine, and patch-based deep convolutional neural networks for object-based wetland mapping using images from small unmanned aircraft system. *GIScience Remote Sens.* **2018**, *55*, 243–264.
- (61) Folleco, A.; Khoshgoftaar, T. M.; Van Hulse, J.; Bullard, L. In *Identifying Learners Robust to Low Quality Data*, 2008 IEEE International Conference on Information Reuse and Integration, 2008; p 190.
- (62) Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *Adv. Neural Inf. Process. Syst.* **2017**, 6402–6413.
- (63) Dunn, A.; Wang, Q.; Ganose, A.; Dopp, D.; Jain, A. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Comput. Mater.* **2020**, *6*, No. 138.
- (64) Park, C. W.; Wolverton, C. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Phys. Rev. Mater.* **2020**, *4*, No. 063801.
- (65) Yu, Y.; Aykol, M.; Wolverton, C. Reaction thermochemistry of metal sulfides with GGA and GGA + U calculations. *Phys. Rev. B* **2015**, *92*, No. 195118.
- (66) Aykol, M.; Wolverton, C. Local environment dependent GGA + U method for accurate thermochemistry of transition metal compounds. *Phys. Rev. B* **2014**, *90*, No. 115105.
- (67) Nolan, A. M.; Zhu, Y.; He, X.; Bai, Q.; Mo, Y. Computation-Accelerated Design of Materials and Interfaces for All-Solid-State Lithium-Ion Batteries. *Joule* **2018**, *2*, 2016–2046.
- (68) Sun, W.; Holder, A.; Orvañanos, B.; Arca, E.; Zakutayev, A.; Lany, S.; Ceder, G. Thermodynamic Routes to Novel Metastable Nitrogen-Rich Nitrides. *Chem. Mater.* **2017**, *29*, 6936–6946.
- (69) Jones, H. Splat cooling and metastable phases. *Rep. Prog. Phys.* **1973**, *36*, 1425.
- (70) Takeuchi, T.; Kageyama, H.; Nakanishi, K.; Inada, Y.; Katayama, M.; Ohta, T.; Senoh, H.; Sakaebe, H.; Sakai, T.; Tatsumi, K.; Kobayashi, H. Improvement of cycle capability of FeS₂ positive electrode by forming composites with Li₂S for ambient temperature lithium batteries. *J. Electrochem. Soc.* **2011**, *159*, A75.
- (71) Darling, A. J.; Stewart, S.; Holder, C. F.; Schaak, R. E. Bulk-immiscible AgRh Alloy Nanoparticles as a Highly Active Electrocatalyst for the Hydrogen Evolution Reaction. *ChemNanoMat* **2020**, *6*, 1320–1324.
- (72) Wang, H.; Li, Y.; Li, C.; Deng, K.; Wang, Z.; Xu, Y.; Li, X.; Xue, H.; Wang, L. One-pot synthesis of bi-metallic PdRu tripods as an efficient catalyst for electrocatalytic nitrogen reduction to ammonia. *J. Mater. Chem. A* **2019**, *7*, 801–805.
- (73) Wu, K.-L.; Yu, R.; Wei, X.-W. Monodispersed FeNi₂ alloy nanostructures: solvothermal synthesis, magnetic properties and size-dependent catalytic activity. *CrystEngComm* **2012**, *14*, 7626–7632.
- (74) Lee, C.-C.; Cheng, Y.-Y.; Chang, H. Y.; Chen, D.-H. Synthesis and electromagnetic wave absorption property of Ni–Ag alloy nanoparticles. *J. Alloys Compd.* **2009**, *480*, 674–680.
- (75) Asano, M.; Umeda, K.; Tasaki, A. Cu₃N thin film for a new light recording media. *Jpn. J. Appl. Phys.* **1990**, *29*, 1985.
- (76) Ono, S.; El Ouenzerfi, R.; Quema, A.; Murakami, H.; Sarukura, N.; Nishimatsu, T.; Terakubo, N.; Mizuseki, H.; Kawazoe, Y.; Yoshikawa, A.; Fukuda, T. Band-structure design of fluoride complex materials for deep-ultraviolet light-emitting diodes. *Jpn. J. Appl. Phys.* **2005**, *44*, 7285.
- (77) Li, Z. H.; Wang, W.; Zhou, P.; Ma, Z.; Sun, L. New type of hybrid nodal line semimetal in Be₂Si. *New J. Phys.* **2019**, *21*, No. 033018.
- (78) Xu, Y.-Q.; Liu, B.-G.; Pettifor, D. G. Half-metallic ferromagnetism of MnBi in the zinc-blende structure. *Phys. Rev. B* **2002**, *66*, No. 184435.
- (79) Gjoka, M.; Panagiotopoulos, I.; Niarchos, D. Structure and magnetic properties of Sm (Co_{1-x}Mx)₅ (M = Cu, Ag) alloys. *J. Mater. Process. Technol.* **2005**, *161*, 173–175.
- (80) McGarvey, B. R.; Reuveni, A. ¹⁹F NMR Studies of Rare Earth Elpasolites. *Magnetic Resonance and Related Phenomena*; Springer, 1979; pp 121–121.
- (81) Sugiyama, K.; Iizuka, T.; Aoki, D.; Tokiwa, Y.; Miyake, K.; Watanabe, N.; Kindo, K.; Inoue, T.; Yamamoto, E.; Haga, Y.; Onuki, Y. High-field magnetization of USn₃ and UPb₃. *J. Phys. Soc. Jpn.* **2002**, *71*, 326–331.
- (82) Balamurugan, B.; Das, B.; Zhang, W.; Skomski, R.; Sellmyer, D. J. Hf–Co and Zr–Co alloys for rare-earth-free permanent magnets. *J. Phys.: Condens. Matter* **2014**, *26*, No. 064204.
- (83) Yannello, V. J.; Lu, E.; Fredrickson, D. C. At the Limits of Isolated Bonding: π -Based Covalent Magnetism in Mn₂Hg₅. *Inorg. Chem.* **2020**, *59*, 12304–12313.
- (84) Kammerdiner, L. W. *Film Deposition of Nb-Based A15 Superconductors*; California University, 1975.
- (85) Volodin, V. N.; Zhakanbaev, E. A.; Tuleushev, A. Z.; Tuleushev, Y. Z. Synthesis and structure of new intermetallic compound Ta₃Pb. *Vestn. Nats. Yad. Tsentra Resp. Kaz.* **2005**, *4*, 49–54.
- (86) Masumoto, H.; Watanabe, K. New compounds of the Clb, Cl types of RhMnSb, IrMnSn and IrMnAl, New L21 (Heusler) type of Ir₂MnAl and Rh₂MnAl alloys, and magnetic properties. *J. Phys. Soc. Jpn.* **1972**, *32*, 281–281.
- (87) Yin, M.; Nash, P. Standard enthalpies of formation of selected XYZ half-Heusler compounds. *J. Chem. Thermodyn.* **2015**, *91*, 1–7.

(88) Aljarrah, M.; Obeidat, S.; Fouad, R. H.; Rababah, M.; Almagableh, A.; Itradat, A. Thermodynamic calculations of the Mn–Sn, Mn–Sr and Mg–Mn–{Sn, Sr} systems. *IET Sci., Meas. Technol.* **2015**, *9*, 681–692.

(89) Xu, J.; Wang, D.; Liu, Y.; Lian, R.; Gao, X.; Chen, G.; Wei, Y. Theoretical prediction and atomic-scale investigation of a tetra-VN 2 monolayer as a high energy alkali ion storage material for rechargeable batteries. *J. Mater. Chem. A* **2019**, *7*, 26858–26866.

(90) Li, X.; Dan, Y.; Dong, R.; Cao, Z.; Niu, C.; Song, Y.; Li, S.; Hu, J. Computational Screening of New Perovskite Materials Using Transfer Learning and Deep Learning. *Appl. Sci.* **2019**, *9*, No. 5510.

(91) Zhou, Y.; Sun, W.; Chu, W.; Zheng, J.; Gao, X.; Zhou, X.; Xue, Y. Adsorption of acetylene on ordered NixAg1-x/Ni (111) and effect of Ag-dopant: A DFT study. *Appl. Surf. Sci.* **2018**, *435*, 521–528.

(92) Dudenkov, I. V.; Solntsev, K. A. Theoretical prediction of the new high-density lithium boride LiB 11 with polymorphism and pseudoplasticity. *Russ. J. Inorg. Chem.* **2009**, *54*, 1261–1272.

(93) Hou, W.; Liu, J.; Zuo, X.; Xu, J.; Zhang, X.; Liu, D.; Zhao, M.; Zhu, Z.-G.; Luo, H.-G.; Zhao, W. Prediction of crossing nodal-lines and large intrinsic spin Hall conductivity in topological Dirac semimetal Ta3As family. *npj Comput. Mater.* **2021**, *7*, No. 37.

(94) Guan-Nan, L.; Ying-Jiu, J. First-principles study on the half-metallicity of half-Heusler alloys: XYZ (X = Mn, Ni; Y = Cr, Mn; Z = As, Sb). *Chin. Phys. Lett.* **2009**, *26*, No. 107101.

(95) Manaa, H.; Moncorgé, R.; Butashin, A. V.; Mill, B.; Kaminskii, A. A. Luminescence Properties of Cr-Doped LiNbGeO₃ Laser Crystal. In *Advanced Solid State Lasers*; Pinto, A.; Fan, T., Eds.; Optical Society of America: New Orleans, Louisiana, 1993; p TL13.

(96) Li, Y.; Dai, X.-F.; Liu, G.-D.; Wei, Z.-Y.; Liu, E.-K.; Han, X.-L.; Du, Z.-W.; Xi, X.-K.; Wang, W.-H.; Wu, G.-H. Structural, magnetic properties, and electronic structure of hexagonal FeCoSn compound. *Chin. Phys. B* **2018**, *27*, No. 026101.

(97) Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S. I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67.

(98) Gerrits, N.; Smeets, E. W. F.; Vuckovic, S.; Powell, A. D.; Doblhoff-Dier, K.; Kroes, G. J. Density Functional Theory for Molecule-Metal Surface Reactions: When Does the Generalized Gradient Approximation Get It Right, and What to Do If It Does Not. *J. Phys. Chem. Lett.* **2020**, *11*, 10552–10560.

(99) Seo, D.-H.; Urban, A.; Ceder, G. Calibrating transition-metal energy levels and oxygen bands in first-principles calculations: Accurate prediction of redox potentials and charge transfer in lithium transition-metal oxides. *Phys. Rev. B* **2015**, *92*, No. 115118.

(100) Grindy, S.; Meredig, B.; Kirklin, S.; Saal, J. E.; Wolverton, C. Approaching chemical accuracy with density functional calculations: Diatomic energy corrections. *Phys. Rev. B* **2013**, *87*, No. 075150.

(101) Schmidt, J.; Wang, H. C.; Cerqueira, T. F. T.; Botti, S.; Marques, M. A. L. A dataset of 175k stable and metastable materials calculated with the PBEsol and SCAN functionals. *Sci. Data* **2022**, *9*, No. 64.

(102) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(103) Musaelian, A.; Batzner, S.; Johansson, A.; Sun, L.; Owen, C. J.; Kornbluth, M.; Kozinsky, B. Learning Local Equivariant Representations for Large-Scale Atomistic Dynamics. 2022, arXiv:2204.05249v1. arXiv.org e-Print archive. <https://arxiv.org/abs/2204.05249v1>.

(104) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **2013**, *68*, 314–319.