ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/355576191

Calibrating DFT formation enthalpy calculations by multi-fidelity machine learning

Preprint · October 2021

DOI: 10.48550/arXiv.2110.13308

CITATIONS
0

READS
631

6 authors, including:

Sheng Gong
ByteDance Inc
49 PUBLICATIONS **2,240** CITATIONS
SEE PROFILE

Shuo Wang
Eastern Institute of Technology, Ningbo
72 PUBLICATIONS **4,898** CITATIONS
SEE PROFILE

Tian Xie
Microsoft
35 PUBLICATIONS **4,060** CITATIONS
SEE PROFILE

David Chae
Massachusetts Institute of Technology
9 PUBLICATIONS **611** CITATIONS
SEE PROFILE

All content following this page was uploaded by Sheng Gong on 25 October 2021.

The user has requested enhancement of the downloaded file.

Calibrating DFT formation enthalpy calculations by multi-fidelity machine learning

Sheng Gong[1], Shuo Wang[2], Tian Xie[3], Woo Hyun Chae[1], Runze Liu[1], and Jeffrey C. Grossman[1,*]

[1]Department of Materials Science and Engineering, Massachusetts Institute of Technology, MA 02139, USA

[2]Department of Materials Science and Engineering, University of Maryland, MD 20742, USA

[3]Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, MA 02139, USA

ABSTRACT

Machine learning materials properties measured by experiments is valuable yet difficult due to the limited amount of experimental data. In this work, we use a multi-fidelity random forest model to learn the experimental formation enthalpy of materials with prediction accuracy higher than the empirically corrected PBE functional (PBEfe) and meta-GGA functional (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 multi-fidelity model is also used as a data-mining approach to find how DFT deviates from experiments by the explaining the model output.

INTRODUCTION

In order to accelerate the design of new materials, accurate computational methods such as Density Functional Theory (DFT)(*1*) have been employed to generate large datasets that contain more than 10^5 entries of materials 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 PBE(*16*), that 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 are reported to be ~0.1 eV/atom(*3, 17*), and due to the sensitivity of phase stability to energy, a difference of 0.1 eV/atom might be the difference between a material that is readily synthesizable and one that is almost impossible to realize(*18-20*). In addition, due to 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*).

In order to improve the accuracy of formation enthalpy calculations, a number of density functionals have been developed, such as modified PBE (PBEfe)(*38*), HSE(*39*) and SCAN(*40*), which have shown significant improvement in accuracy of formation enthalpy calculation(*38, 41*). On the other hand, these more accurate functionals are also computationally more expensive, limiting their utility for generation of large databases(*41, 42*). Empirical corrections represent another, faster approach to try to improve the accuracy of 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(*43*), and in OQMD ΔH_f is corrected by a chemical-potential fitting(*3*) with compounds that have experimentally measured formation enthalpies (exp. ΔH_f). Very recently, Wang *et al.*(*44*) proposed a linear correction scheme with MAE 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(*45*), such corrections are based on human understanding of specific chemistries and relatively simple assumptions and are thus difficult to be transferrable across different chemical spaces(*43*, *45*). It would be beneficial to design prediction schemes that can automatically extract chemistry-property relationship across different chemistries without human intervention, and data-driven ML methods(*18, 24*, *26, 28, 29, 32, 42*) are promising candidates to learn the complex mapping between chemistry and ΔH_f.

One of the biggest challenges in machine learning materials properties is the lack of experimental data(*46*), and as such 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, 47-50*) and multi-fidelity machine learning(*42, 51-53*). The idea of transfer learning (see 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 datasets are highly correlated(*21, 28, 47, 48*), the approach is mostly limited to neural network architectures, and if the correlation is not strong enough, transfer learning will not improve and may even deteriorate the learning performance(*49*). Different from transfer learning where information is passed by transferring network parameters, in multi-fidelity machine learning (see 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)

level(*51*) or in the label (output) level(*42, 52-54*). In other words, the low-fidelity data can be used as 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(*54*). From the handful of previous studies, multi-fidelity machine learning has shown higher predictive power than the single-fidelity ones (directly learning the high-fidelity data) on learning materials properties like band gaps and energies from different density functionals where there are clear systematic shifts between different levels of data(*42, 52-54*). However, there is no previous work that adapt multi-fidelity machine learning in both feature and label level at the same time, and there remains a question whether a combination of transfer learning and multi-fidelity machine learning improves performance.

In this work, we present a comprehensive machine learning study comparing these two strategies for passing information from a large DFT dataset (MP database) to a small experimental dataset, and for multi-fidelity ML we explore the effects of feeding low-fidelity data as feature versus learning the difference between high and low-fidelity data and a combination of the two operations. As for the machine learning architectures, we compare four different models, random forests (RF), multi-layer perceptron (MLP), Representation Learning from Stoichiometry (ROOST)(*26*) and Crystal Graph Convolutional Neural Network (CGCNN)(*32*). We find that multi-fidelity RF in both the feature and label level has the best prediction performance for experimental ΔH_f with a 30% reduction in MAE compared with DFT results from MP, and improved performance compared to recent linear correction schemes(*44*) as well as more sophisticated density functionals like PBEfe(*38*) and SCAN(*40*). We also analyze the effects of machine learning architectures, featurization methods and information transfer strategy on learning experimental ΔH_f and diff. ΔH_f. 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.

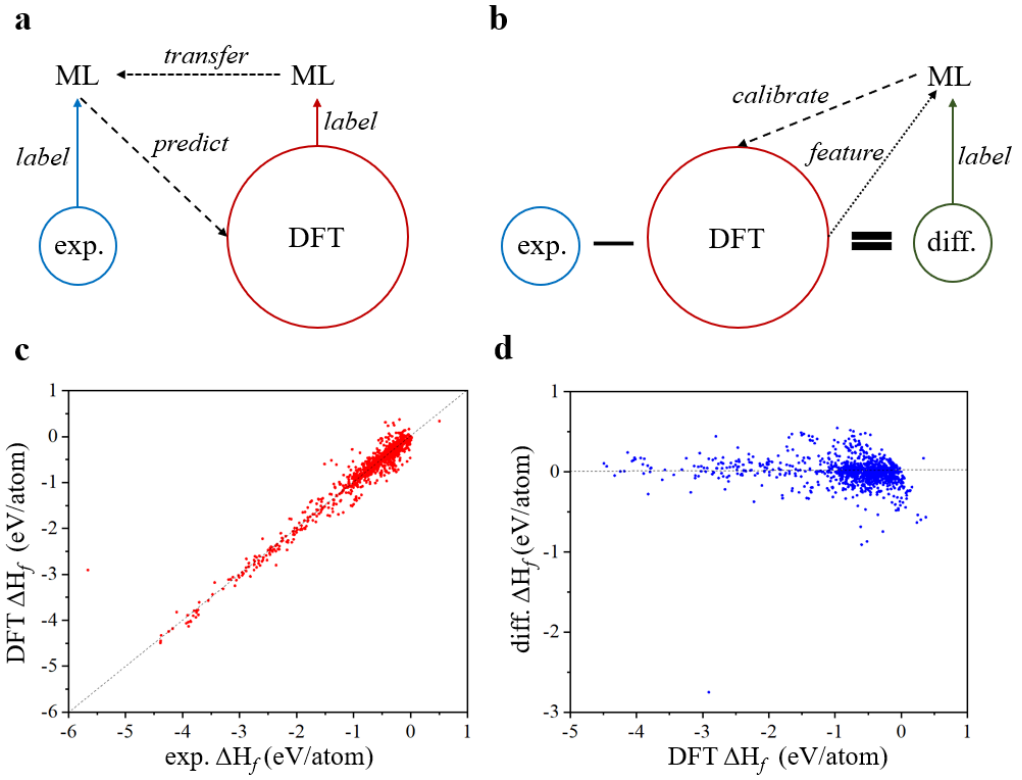

Figure 1. Illustrations of the machine learning frameworks and datasets used in this work. a and **b** Schematics of transfer learning and multi-fidelity machine learning in this work, respectively. In **a**, the DFT ΔH_f (DFT Formation enthalpies from Materials Project) are first used as label to train a ML model, then the weights of the first ML model are transferred to initialize a second ML model, and the exp. ΔH_f (experimental formation enthalpies) are used as label to train the second model, and finally the second model is used to predict exp. ΔH_f of all materials in the large DFT dataset. In **b**, first the dataset of the difference between exp. ΔH_f and DFT ΔH_f are constructed (diff. ΔH_f), then diff. ΔH_f are used as label to train a ML model with the DFT ΔH_f as input feature, and finally the trained model is used to calibrate the different between DFT ΔH_f and exp. ΔH_f for all materials in the large DFT dataset. **c** DFT ΔH_f versus exp.

ΔH_f for materials with both values in the dataset used in this work. **d** Difference between experimental ΔH_f and DFT ΔH_f (diff. ΔH_f) versus DFT ΔH_f.

RESULTS

Illustration of machine learning frameworks and datasets. In this work, we use two different strategies to learn exp. ΔH_f with the assistance of information from the MP dataset, transfer learning and multi-fidelity machine learning (in the following, "DFT ΔH_f" denotes the empirically-corrected PBE ΔH_f from the MP database, queried in March, 2021). As shown in Figure 1a, in transfer learning a neural network is first trained on the large MP dataset with more than 10^5 empirically-corrected PBE formation enthalpies, then weights of the neural network are transferred to initialize a second neural network, and finally part of the weights of the second network are optimized by the small exp. ΔH_f dataset. Once trained, the second neural network can serve to predict exp. ΔH_f of materials in the large MP dataset. In multi-fidelity machine learning, as shown in Figure 1b, first the dataset of the difference between the exp. ΔH_f and DFT ΔH_f (diff. ΔH_f) is built, then machine learning models are trained on diff. ΔH_f dataset, and in the training process, DFT ΔH_f can also serve as an input feature of each material. Once trained, the machine learning model can serve to calibrate the DFT ΔH_f by adding diff. ΔH_f to DFT ΔH_f to get the exp. ΔH_f. The key difference between transfer learning and multi-fidelity machine learning is that in the former two networks are trained and information transfer is achieved by transferring network weights, while in the later only one model is trained and information transfer is done by learning the difference between two datasets and adding the DFT ΔH_f as one of the input features. Note that, in addition to the two basic strategies as shown in Figure 1a and 1b, variations are also tested in this work, including combination of transfer learning and multi-fidelity machine learning by initializing a second network from that trained on DFT ΔH_f and optimizing

the second network by diff. ΔH_f, and multi-fidelity machine learning by only learning diff. ΔH_f and only adding DFT ΔH_f as input feature.

As described above, we choose four different machine learning architectures to realize transfer learning and/or multi-fidelity machine learning, 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 multi-task learning and to enlarge the hypothesis space to search for the best machine learning models for predicting exp. ΔH_f. These ML architectures also provide a variety in terms of aspects 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 input while CGCNN takes both compositions and 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 database as the source of DFT ΔH_f (queried in March, 2021), as it is a frequently updated and widely used large DFT database, ΔH_f is empirically corrected in MP(*43*), 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 dataset(*17*) and SSUB dataset(*55*) and remove the duplicates, leading to 1143 data points with available exp. ΔH_f, DFT ΔH_f, and DFT optimized 3D atomic structures from MP. In addition to the value of exp. ΔH_f, there are also noise level estimations in the IIT dataset(*17*), from which one can see that the mean noise of exp. ΔH_f based on 499 materials is around 0.023 eV/atom, in agreement with that from Wang *et al*.(*44*) (0.022 eV/atom). More details about the data collection procedure are provided in the **METHODS** section. DFT ΔH_f and exp. ΔH_f are compared in Figure 1c, from which one can see that DFT ΔH_f are already quite close to exp. ΔH_f in value,

and there is no clear systematic shift between DFT ΔH_f and exp. ΔH_f, and as shown in Figure 1d, the distribution of diff. ΔH_f is centered around zero, and there is no obvious correlation between diff. ΔH_f and DFT ΔH_f. From Figure 1c and 1d, one can see that diff. ΔH_f has a narrower distribution than exp. ΔH_f with the standard deviation of 0.1718 eV/atom and 0.8000 eV/atom for the diff. ΔH_f dataset and exp. ΔH_f dataset, respectively.

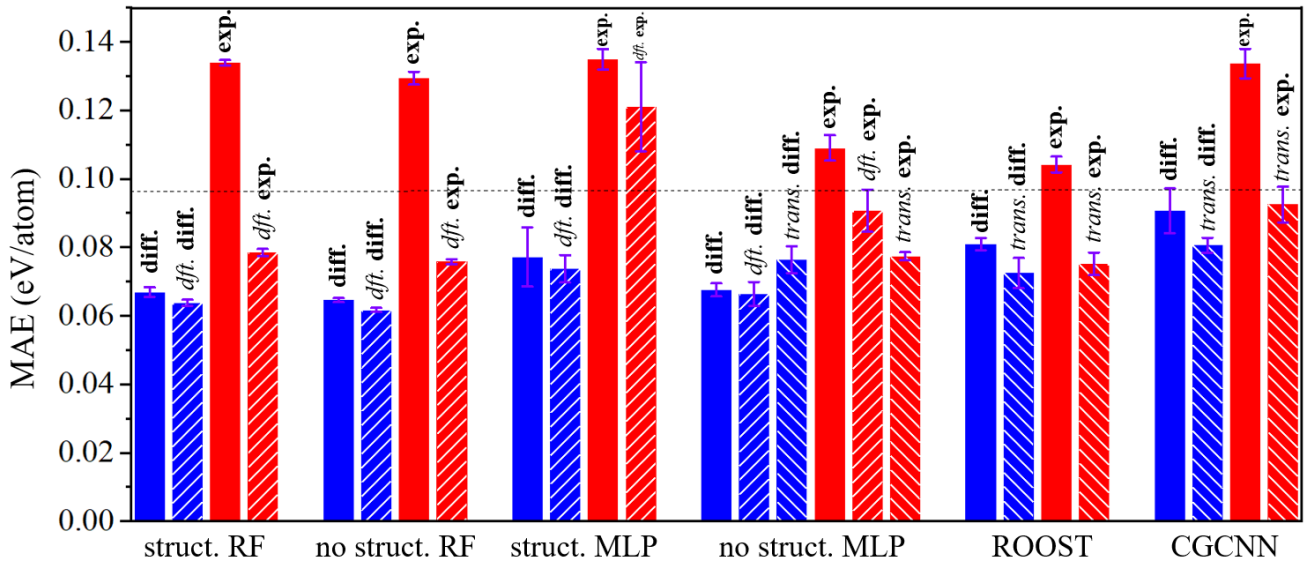

Figure 2. Comparison of machine learning models. Mean average errors (MAE) between predictions of formation enthalpies (ΔH_f) from machine learning models and experimental measurements. MAE values are from the same test set, and each type of machine learning model is trained 10 times to estimate the uncertainty levels. RF denotes random forest, MLP denotes multilayer perceptron, and ROOST(*26*) and CGCNN(*32*) are two deep-learning models that automatically extract materials' fingerprints from compositions and structures, respectively. Here, "struct." means the model is trained with structural features, "no struct." the model is trained with only compositional features, "dft." the model is trained with DFT ΔH_f as input, "trans." the model is trained in a transfer learning manner, "diff." the model is trained on difference between DFT and experimental ΔH_f as label, "exp." the model is directly trained on

experimental ΔH_f as label. The dashed horizontal line corresponds to the MAE of empirically corrected PBE formation enthalpy (DFT ΔH_f).

Predicting exp. ΔH_f by machine learning. For the RF and MLP machine learning architectures, compositional and structural features are provided from matminer(*56*) as input features (a list of features is provided in the **METHOD** 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. In order to test the prediction performance, 20% of the 1143 materials are randomly chosen as the test set. Details about the training procedure are provided in the METHOD section. As a baseline, for the test set, we find that the MAE between DFT ΔH_f and exp. ΔH_f is 0.0955 eV/atom. The test results for all machine learning models are shown in Figure 2 with a detailed list in 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 diff. ΔH_f and has compositional features and DFT ΔH_f as input features (Figure 2). The error for this best case is roughly 30% lower than that of DFT ΔH_f, with a level of error (0.0617 eV/atom) close to "chemical accuracy" of 1 kcal/mol (0.04 eV/atom)(*41*).

We attribute the different performance of the best RF model here and the linear correction scheme from Wang *et al*.(*44*) (0.0617 eV/atom vs. 0.128 eV/atom for the same test set) to the following: i) the RF model takes non-linear effects into account, ii) the compositional descriptors used here capture more information than simple stoichiometry, iii) the learned correction in Wang *et al*.(*44*) is from materials with certain anions and transition metals while in the present work there are no such constraints, and iv) the calibration scheme used here is built on empirically corrected PBE results as opposed to uncorrected PBE

data. Note that the standard deviation of prediction errors from the best RF model is only 0.0006 eV/atom as shown in Figure 2, which means that the RF regressor is very stable in multiple runs and hence we do not analyze the effects of uncertainty of energy correction on compound stability.

In addition to comparison with other machine learning models, prediction performance of the best RF model is also compared with other sophisticated density functionals. As shown in Table S2, the best RF model outperforms the PBEfe(*38*) functional and SCAN(*40*) functional with MAE 0.0582 eV/atom versus 0.1178 eV/atom for RF versus PBEfe based on 44 materials, and 0.0668 eV/atom versus 0.1018 eV/atom for RF versus SCAN based on 85 materials. Note that the materials selected for comparison are not in the training sets for the RF model and PBEfe/SCAN functional. These results are encouraging, because on the one hand the best RF model provides 50% lower error compared with PBEfe and 35% lower error compared with PBEfe, on the other hand, it is much faster than the self-consistent DFT simulations, enabling one to screen ΔH_f of materials accurately in a high-throughput fashion. For example, for more than 10^5 materials that already have corrected-PBE ΔH_f in existing materials databases, more accurate predictions of ΔH_f can be calculated by the RF models within minutes, while that from PBEfe may take months of calculations and even longer time from SCAN as it is usually 2-3 times slower than PBE(*41*). Note that for new materials without low-fidelity ΔH_f predictions yet (such as corrected-PBE), since the best RF model requires the low-fidelity ΔH_f in the prediction, computational cost for the low-fidelity ΔH_f should be added to the total cost of the best RF model. For scenarios where evaluations of ΔH_f for new materials with cost lower than PBE are required, as shown in Figure 2, “no. struct. MLP” with transfer learning and ROOST(*26*) with transfer learning are recommended machine learning models, which are the only two models trained without information of DFT ΔH_f that have better prediction accuracy than empirically corrected PBE.

(2). Training the machine learning models by diff. ΔH_f helps to reduce error compared with training models by exp. ΔH_f directly, as under the same condition (architecture and featurization), the models trained on diff. ΔH_f always have lower MAE than that trained on exp. ΔH_f. Here, we attribute the lower absolute error of learning diff. ΔH_f to the fact that diff. ΔH_f has a narrower distribution than exp. ΔH_f with 5x smaller standard deviation (0.17 eV/atom versus 0.80 eV/atom). As a comparison, the MAEs of models trained on diff. ΔH_f are larger than 1/5 of that trained on exp. ΔH_f, suggesting that diff. ΔH_f is easier to learn absolutely but harder to learn relatively than exp. ΔH_f.

(3). Feeding DFT ΔH_f 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 DFT ΔH_f as one of the input features always have lower error than that without DFT ΔH_f. This effect is more significant when the models are trained on exp. ΔH_f, because as shown in Figure 1c DFT ΔH_f has a strong correlation with exp. ΔH_f, while as shown in Figure 1d the correlation between DFT ΔH_f and diff. ΔH_f is not obvious.

Combining analysis (2) and (3), one can see that adapting the strategy of multi-fidelity machine learning can 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 datasets. Machine learning models with both the modifications of changing label and adding extra input features can outperform that with either single modification.

(4). Similar to (3), transfer learning helps more when transferring from DFT ΔH_f to exp. ΔH_f than from DFT ΔH_f to diff. ΔH_f because of the stronger correlation between DFT ΔH_f and exp. ΔH_f. On the other hand, transfer learning helps more for the deep learning architectures that learn the representations of materials (ROOST and CGCNN) while it helps less or even hurts performance of MLP with *off-the-shelf* featurization, confirming that the power of transfer learning in materials science lies on the transfer of

learned materials representations from larger datasets(*21, 50*), and transferring the mapping function between representations and properties might lead to negative transfer(*57*) if the correlation between source and target is not very strong, such as the case of "no struct. MLP" in Figure 2.

(5) RF performs better than neural network-based models when trained on diff. ΔH_f, while RF and neural-network based models in general have similar performance when trained on exp. ΔH_f. It is not surprising that neural-network based deep learning algorithms don't show superior performance over RF due to the limited dataset size(*49, 58*). The different noise level between diff. ΔH_f and exp. ΔH_f might help to explain why RF performs better than neural network-based models when trained on diff. ΔH_f while there is no such superiority of RF when trained on exp. ΔH_f. As discussed above, diff. ΔH_f has a narrower distribution than exp. ΔH_f, and because diff. ΔH_f = exp. ΔH_f - DFT ΔH_f, if we consider exp. ΔH_f and DFT ΔH_f as two independent random variables, then diff. ΔH_f should have larger noise level than exp. ΔH_f (~0.023 eV/atom as discussed above). Therefore, the dataset of diff. ΔH_f is noisier than exp. ΔH_f as diff. ΔH_f has smaller distribution and larger (absolute) noise level, and the robustness of RF against noise(*58, 59*) might explain the superiority of RF when trained on diff. ΔH_f. The larger noise level of diff. ΔH_f might also help to explain why diff. ΔH_f is harder to learn relatively than exp. ΔH_f as in (2).

(6). Comparing the performance of ROOST, CGCNN and MLP, one can see that the first two architectures that automatically learn the materials representation have similar performance compared with MLP with *off-the-shelf* featurization when trained on exp. ΔH_f, while MLP has slightly better performance when trained on diff. ΔH_f, which implies that although the machine learning architectures learning representations of materials are thought to have a strong ability to learn the complex mapping(*24, 26, 32, 60*), the human-engineered featurization is more robust when the dataset is noisy.

Combining the facts that random forest outperforms neural networks-based models and human-engineered featurization outperforms representation learning when trained on diff. ΔH_f, here we suggest that for machine learning applications in the field of materials science with limited dataset size, without proof of a low noise level of the dataset, deep neural network-based representation learning algorithms(*24, 26, 32, 60*) 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 show that information of local bonding environment can be used to correct formation enthalpies of certain materials like sulfides(*61*), fluorides(*62*) and oxides(*45, 62*), in this work, the machine learning models with only compositions as input outperform those with both compositions and structures as input. This is because there still lacks the data points of polymorphs with the same composition but different exp. ΔH_f in the current dataset, which shows the urgency of building a comprehensive exp. ΔH_f dataset with sufficient entries of polymorphs to comprehensively understand the role of structures in determining exp. ΔH_f.

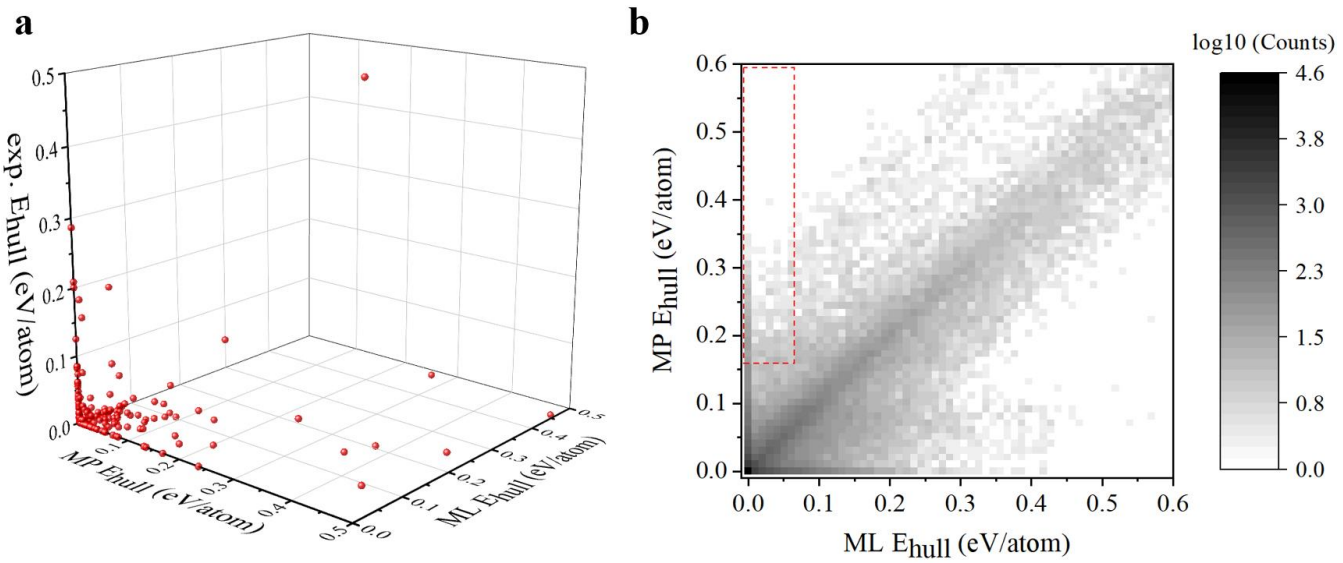

Figure 3. Stability evaluation from energy above hull. a Energy above hull (E_{hull} eV/atom) of materials from experiments versus the MP dataset versus machine learning (ML). Here, E_{hull} is constructed from the 1143 materials in the diff. ΔH_f dataset used as above. **b** Distribution of energy above hull (E_{hull} eV/atom) of all materials in the Materials Project(*2*) database calculated by the corrected-PBE ΔH_f in MP (MP E_{hull}) versus that calculated by the machine learning ΔH_f in this work (ML E_{hull}). Here, E_{hull} is constructed from all materials in the Materials Project(*2*) database, the color scheme is used to show the (log10 of) number of materials within a range of certain ML E_{hull} and MP E_{hull}, and the red rectangle shows the area (MP E_{hull} > 0.16 eV/atom and ML E_{hull} < 0.06 eV/atom) where ML E_{hull} estimates materials to be stable while MP E_{hull} estimates materials to be unstable.

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 the calibrated ΔH_f is provided in the **DATA AVAILABILITY** section, 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(*63*)) of all materials in the

MP database. However, as Bartel *et al.*(*18*) pointed out, although sometimes DFT has large errors for prediction of ΔH_f, DFT ΔH_f of similar materials contain similar systematic errors, and when evaluating phase stability, the cancellation of systematic errors makes DFT more useful for evaluating relative stability between compounds than some machine learning models with similar or even better accuracy with respect to exp. ΔH_f. Therefore, before screening E_{hull} for the full MP dataset, we first evaluate the performance of DFT ΔH_f and ML ΔH_f for constructing E_{hull}. We use exp. ΔH_f, DFT ΔH_f from the MP database, and ML calibrated ΔH_f of the 1143 materials as used above to construct E_{hull} (here, ML ΔH_f of each material is predicted by training a model with the remaining 1142 materials). The MAE between exp. E_{hull} and MP E_{hull} is 0.079 eV/atom while the MAE between exp. E_{hull} and ML E_{hull} is 0.066 eV/atom, which means that, in addition to ΔH_f, the ML model also outperforms DFT on constructing E_{hull}, although the difference in performance is closer for constructing E_{hull} than predicting ΔH_f due to the cancellation of systematic errors of DFT for similar compounds. More importantly, as shown in Figure 3a, ML E_{hull} can be used to discover materials that may have been overlooked in the MP database with exp. E_{hull} close to 0 and considerably positive MP E_{hull}, as there are points in Figure 3a with close-to-zero exp. E_{hull}, considerably positive MP E_{hull} and close-to-zero ML E_{hull}. Such discovery is particularly important based on the fact that large DFT databases such as the Materials Project have boosted the real-world synthesis and/or characterization of many novel materials(*11-15*), and E_{hull} from these large DFT databases has been used as a key criterion to determine whether certain materials would be considered for potential realization(*11-14*). Therefore, increased prediction accuracy for materials with low E_{hull} can uncover new potential materials with previously underestimated stabilities that may have been overlooked as candidates for experimental synthesis and characterization.

We next re-evaluate materials stability using ML calibrated ΔH_f to construct ML E_{hull} 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, 64*), although E_{hull} = 0 is not a hard threshold for successful synthesis and room-temperature and pressure stability of materials because of other factors such as kinetics(*65*), and in practice empirical heuristics of several room temperature k_BT are used as stability thresholds(*19, 20, 64*). In Figure 3b the distributions of E_{hull} of all materials in the MP database constructed from DFT ΔH_f and ML ΔH_f of all compositions in the MP database are shown, from which one can see that most materials have similar MP E_{hull} and ML E_{hull}, and majority of materials have close-to-zero E_{hull} from both MP E_{hull} and ML E_{hull}. More importantly, as expected above, there are materials with large MP E_{hull} and small ML E_{hull}, which are materials that might have underestimated stabilities in MP. For example, there are 800 materials in the red rectangles in the upper left corner in Figure 3b that have MP E_{hull} > 0.16 eV/atom and ML E_{hull} < 0.06 eV/atom, among which there are around 100 already synthesized materials. (The thresholds are set to be relaxed from 6 times and 2 times of room Temperature k_BT(*19, 20*)). As examples, we list some interesting materials in Table 1 with novel physical properties and/or potential applications with MP E_{hull} > 0.16 eV/atom and ML E_{hull} < 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(*66, 67*), catalysts(*68-70*) to optical(*71-73*), electronic(*74, 75*), magnetic(*76-80*) devices and superconductors(*81, 82*), for which ML E_{hull} succeeds in explaining their synthesizability and stability while MP E_{hull} does not. One extreme example is MnSnIr(*83*), a stable Half-Heusler compound synthesized from a peritectic reaction(*84*), of which MP E_{hull} is considerably high (0.5117 eV/atom) while ML E_{hull} is 0, showing the utility of our machine learning model to identify and correct underestimated stabilities in the MP database. In addition to the already synthesized materials, those unrealized hypothetical materials provide potential opportunities for energy and environmental materials(*85-87*), structural materials(*88*) and electronic devices(*89-91*), and as shown

in Table 1 and Figure 3b, many of these materials that are estimated stable by ML E_{hull} might have underestimated stability in the MP database.

Table 1. Examples of materials that have novel physical properties and/or potential applications with MP E_{hull} > 0.16 eV/atom and ML E_{hull} < 0.06 eV/atom. The materials with experiment as one of the characterization methods are synthesized materials, and others are currently only hypothetical.

Materials	MP E_{hull}	ML E_{hull}	Characterization method(s)	Comment/ novel physical property/ potential application
MnSnIr	0.5117	0	Experiment	Largest difference between MP E_{hull} and ML E_{hull}.
Ta_3Pb	0.3386	0	Experiment	Superconductor(*82*)
AgRh	0.2633	0.0359	Experiment	Electrocatalyst(*68*)
FeCoSn	0.1836	0.0384	Experiment	Tuning phase transitions for isostructural alloying(*92*)
$SmCo_4Ag$	0.1797	0.0493	Experiment	Positively correlated magnetization with temperature(*76*)
$Li_3(FeS_2)_2$	0.1697	0.0180	Experiment	Li-FeS_2 battery electrode(*66*)
$LiMnO_3$	0.1809	0.0567	Experiment	Li-ion battery electrode(*67*)
PdRu	0.2277	0.0032	Experiment	Catalyst(*69*)
Ni_3Ag	0.2332	0	Experiment	Dual-frequency absorption(*71*)
Rb_2NaTaF_6	0.2038	0	Experiment	Large anisotropic shift from both covalent and polarization spin transfer mechanisms(*77*)
Nb_3Tl	0.2083	0	Experiment	Superconductor(*81*)
UPb_3	0.1621	0	Experiment	Sharp metamagnetic transitions(*78*)
Cu_3N	0.1865	0.0464	Experiment	Light recording media(*72*)
$FeNi_2$	0.1858	0.0292	Experiment	Size-dependent catalytic activity(*70*)
$HfCo_7$	0.2098	0.0500	Experiment	Rare-earth-free permanent magnets(*79*)
MnBi	0.2078	0	Experiment/DFT	Half-metallic ferromagnetism(*75*)

$LiCaF_3$	0.1874	0	Experiment/DFT	Deep-ultraviolet light-emitting diodes(*73*)
Be_2Si	0.2352	0.0272	Experiment/DFT	Hybrid nodal-line semimetal(*74*)
Mn_2Hg_5	0.2362	0	Experiment/DFT	π-based covalent magnetism(*80*)
Ta_3Bi	0.3442	0	DFT	Topological Dirac semimetal(*90*)
MnCrSb	0.2564	0	DFT	Half-metallicity(*91*)
LiB_{11}	0.2084	0.0234	DFT	Pseudo-plasticity(*88*)
$NiAg_3$	0.1850	0	DFT	Acetylene adsorbent(*87*)
LiAgSe	0.1746	0.0572	DFT	Topological insulator(*89*)
Li_2VN_2	0.1615	0.0279	DFT	Li-ion battery electrode(*85*)
$LiGdO_3$	0.3476	0.0575	Machine learning	Perovskite with high tolerance factor(*86*)
$LiPmO_3$	0.2815	0	Machine learning	Perovskite with high tolerance factor(*86*)

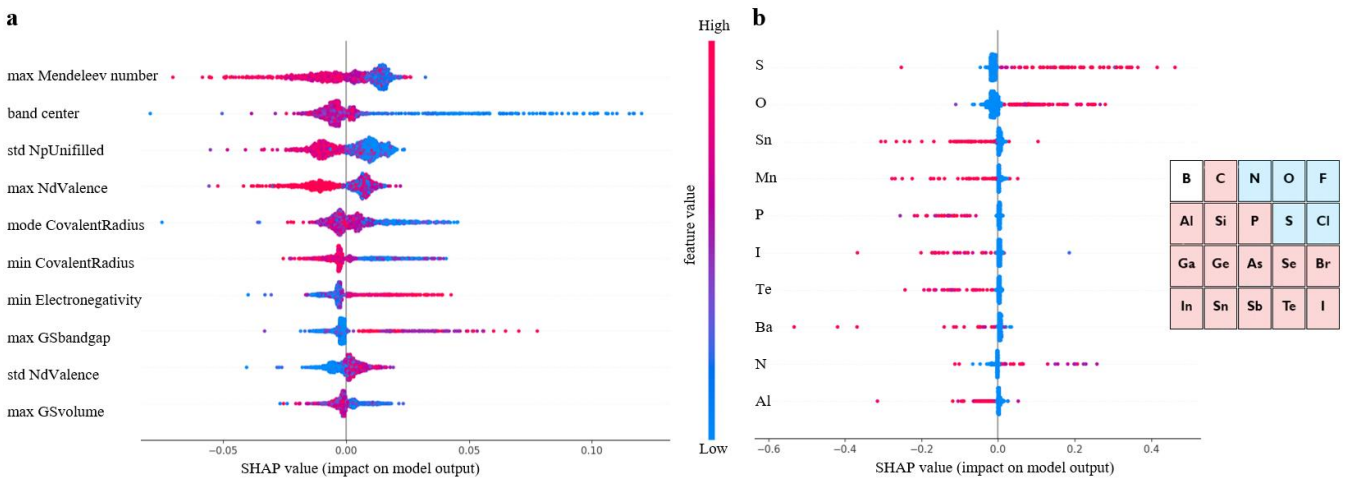

Figure 4. Impact of each feature on model output. a and **b** Distributions of the impacts (SHAP values(*93*)) of compositional and structural features and elemental fractions on the model output (diff. ΔH_f), respectively. The color represents the feature value (red high, blue low), and here only the top 10 features and elemental fractions with the highest sum of absolute SHAP values are shown. The inserted

figure in **b** illustrates the trends of DFT to underestimate or overestimate ΔH_f of materials with certain non-metal 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.

Data-mining where DFT ΔH_f fails by explaining the multi-fidelity model. In addition to predicting more accurate ΔH_f and examining stability of materials, the random forest model trained on diff. ΔH_f (exp. ΔH_f – DFT ΔH_f) with human-engineered features can also serve as a data-mining approach to reveal where and how DFT ΔH_f deviates from exp. ΔH_f (as above, "DFT ΔH_f" refers to the empirically corrected PBE ΔH_f in the Materials Project database), which provides clearer trends than machine learning models trained on DFT ΔH_f only. Here, we analyze the relationship between human-understandable features and diff. ΔH_f by explaining the model, or for each material, calculating the impact of each feature on the model output (known as the SHAP value(*93*)). Previously, the error of DFT ΔH_f is mostly discussed in the context of certain anions(*3, 39, 43*), cations(*3*) and transition metals(*3, 38, 39*). In Figure 4a, the impacts of the top 10 compositional and structural features from matminer(*56*) with the highest sum of absolute SHAP values are shown, from which one can see that, in addition to anion properties ("max GSbandgap", "max GSvolume", the detailed explanations of the descriptors are available in the matminer paper(*56*)) and cation properties ("max NdValence", "min CovalentRadius", "min Electronegativity"), 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 diff. ΔH_f. For example, for smaller "band center" (geometric mean of electronegativity(*56*)), diff. ΔH_f tends to be larger and DFT ΔH_f tends to be smaller than exp. ΔH_f, which means that DFT tends to underestimate ΔH_f of systems with smaller mean electronegativity. Another example is, for larger standard

deviation of number of p valence electrons, diff. ΔH_f tends to be smaller and DFT ΔH_f tends to be larger than exp. ΔH_f, suggesting that DFT tends to overestimate ΔH_f of systems with more dissimilar p valence electron configurations.

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 fraction of each element are plotted in Figure 4b. One can see that, for higher atomic fraction of S, O and N, DFT tends to underestimate ΔH_f, while for higher atomic fraction of Sn, Mn, P, I, Te, Ba, Al, DFT tends to overestimate ΔH_f. There are more non-metal 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 non-metal elements: as shown in Figure 4b and Figure S1, for strong oxidizing non-metal elements in the upper-right corner of the periodic table, including F, O, N, S, Cl, DFT tends to underestimate ΔH_f, while for those non-metal elements with weaker oxidizing ability, DFT tends to overestimate ΔH_f. 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, 43*), and on the other hand, the intrinsic limit of the GGA approximation and its underestimation of electron localization might cause the different deviation patterns. Note that in Wang *et al*.(*44*) 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(*94*), while the trend shown here is based on the empirically corrected PBE energies from MP that already take the effect of overestimated energy of diatomic gas molecules into account.

DISCUSSION AND CONCLUSION

In this work, we conduct a comprehensive machine learning study to learn and predict experimental formation enthalpy of materials. We use two different strategies to transfer information from larger DFT dataset to the smaller experimental dataset, transfer learning and multi-fidelity machine learning, and we use four machine learning architectures to realize the two strategies. We find that the random forest model trained on the difference between experimental and DFT formation enthalpy with DFT formation enthalpy as one of the input features can achieve the lowest error, which is 30% lower than that of DFT (empirically corrected PBE), and it also outperforms other more accurate but also computationally intensive density 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, and 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 multi-fidelity machine learning should also be tested in machine learning applications for materials science.

As an application, we employ the found best random forest model to calibrate the formation enthalpy of all materials in the Materials Project database with more accurate formation enthalpies, which are then used to construct energy above hull and discover potential important materials that have underestimated stability in the MP database and have therefore not been considered previously as synthesizable. 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 strong oxidizing elements with higher electronegativity, and overestimate the formation enthalpy of materials with other weak oxidizing non-metal elements.

Note that this work is based on Materials Project database queried in March, 2021. The methodology of this work can also be applied to updated Materials Project database and other large DFT databases, and

it is expected that, with more accurate low fidelity data (DFT formation enthalpy), the method in this work can be used to provide more accurate calibration (exp. formation enthalpy).

One potential limitation of the multi-fidelity model used in this work is that it requires the availability of low-fidelity data for all the materials space of interest, as in this work DFT formation enthalpy is required for learning the difference of formation enthalpy from experiment and DFT and predicting experimental formation enthalpy of materials in the Materials Project database. In cases where low-fidelity data is not available to all the materials, transfer learning might be more appropriate to transfer information between different datasets. Another scenario not considered in the current multi-fidelity machine learning scheme is that, for some materials properties like band gaps, there are 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.

METHODS

Data collection. In this work, we construct the exp. ΔH_f dataset by combining two datasets from IIT(*17*) and SSUB(*55*), and we use the Materials Project(*2*) database to construct the DFT ΔH_f dataset. For the diff. ΔH_f dataset, since the DFT ΔH_f values are provided for some materials in the IIT dataset, diff. ΔH_f values for those materials are obtained by subtracting the provided DFT ΔH_f from the provided exp. ΔH_f, and for materials from the SSUB dataset, since chemical formula is the only identifier, we take the lowest exp. ΔH_f for each formula, and for the DFT ΔH_f of these materials, we assign the lowest DFT ΔH_f to each formula. Note that, for overlaps between the IIT dataset and SSUB dataset, we take the exp. ΔH_f from the IIT database as the IIT database is a more recent one(*17*). Note that the mean absolute difference of exp.

ΔH_f between our dataset and the recent dataset from Wang *et al*. is only 0.007 eV/atom. Codes and a step-by-step instruction for constructing the dataset are provided in the **CODE AVAILABILITY** section.

Machine learning models training procedure. In this work, we first randomly reserve 20% data as the test set for final model evaluation. Then, to determine the best hyper-parameters for each model, with the remaining 80% data, we randomly reserve 20% of the remaining data (20%*80% = 16%) as the validation set to evaluate each specific set of hyper-parameters, and use the 80%*80% = 64% remaining data to train the machine learning model with the given set of hyper-parameters. We screen hyper-parameters by grid search, and tables of search space of hyper-parameters are provided in the Supplementary Information. Finally, with the found best hyper-parameters for each model, we use the 80% of the data (training set + validation set in the hyper-parameter search step) to train machine learning models 10 times with different initialization, and estimate model performance and uncertainty using the 20% data held out at the very beginning (test set). Codes for training the machine learning models are provided in the **CODE AVAILABILITY** section.

In this work, we use four different machine learning architectures to realize transfer learning and/or multi-fidelity machine learning, random forest (RF), multi-layer 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(*95*), and we use the descriptors from matminer(*56*) to feed RF and MLP as features of materials. Modules used to generate compositional features are ElementProperty, ElectronAffinity, BandCenter, CohesiveEnergy, Miedema, TMetalFraction, ValenceOrbital, YangSolidSolution, and modules used to generate structural features are

GlobalSymmetryFeatures, StructuralComplexity, ChemicalOrdering, MaximumPackingEfficiency, MinimumRelativeDistances, StructuralHeterogeneity, AverageBondLength, AverageBondAngle, BondOrientationalParameter, CoordinationNumber, and DensityFeatures.

DATA AVAILABILITY

All data or their collection methods are provided at:

https://github.com/shenggong1996/exp.-formation-enthalpies

CODE AVAILABILITY

All codes for data collection and training machine learning models are provided at:

https://github.com/shenggong1996/exp.-formation-enthalpies

REFERENCES

1. W. Kohn, L. J. Sham, Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **140**, A1133-A1138 (1965).
2. S. P. O. A. Jain*, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
3. S. Kirklin *et al.*, The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **1**, (2015).
4. S. Curtarolo *et al.*, AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* **58**, 227-235 (2012).
5. K. Choudhary *et al.*, JARVIS: An Integrated Infrastructure for Data-driven Materials Design. *arXiv:2007.01831*, (2020).

6. Z. Ahmad, T. Xie, C. Maheshwari, J. C. Grossman, V. Viswanathan, Machine Learning Enabled Computational Screening of Inorganic Solid Electrolytes for Suppression of Dendrite Formation in Lithium Metal Anodes. *ACS Cent Sci* **4**, 996-1006 (2018).
7. M. K. Horton, S. Dwaraknath, K. A. Persson, Promises and perils of computational materials databases. *Nat. Comput. Sci.* **1**, 3-5 (2021).
8. A. D. Sendek *et al.*, Holistic computational structure screening of more than 12 000 candidates for solid lithium-ion conductor materials. *Energy Environ. Sci.* **10**, 306-320 (2017).
9. T. Yang *et al.*, High-Throughput Identification of Exfoliable Two-Dimensional Materials with Active Basal Planes for Hydrogen Evolution. *ACS Energy Lett.* **5**, 2313-2321 (2020).
10. J. Zhou *et al.*, 2DMatPedia, an open computational database of two-dimensional materials from top-down and bottom-up approaches. *Sci Data* **6**, 86 (2019).
11. Q. Yan *et al.*, Solar fuels photoanode materials discovery by integrating high-throughput theory and experiment. *Proc Natl Acad Sci U S A* **114**, 3040-3043 (2017).
12. H. Zhu *et al.*, Computational and experimental investigation of TmAgTe2and XYZ2compounds, a new group of thermoelectric materials identified by first-principles high-throughput screening. *J. Mater. Chem. C* **3**, 10554-10565 (2015).
13. M. T. Dunstan *et al.*, Large scale computational screening and experimental discovery of novel materials for high temperature CO2 capture. *Energy Environ. Sci.* **9**, 1346-1360 (2016).
14. S. Li *et al.*, Data-Driven Discovery of Full-Visible-Spectrum Phosphor. *Chem. Mater.* **31**, 6286-6294 (2019).
15. J. A. Cooley *et al.*, From Waste-Heat Recovery to Refrigeration: Compositional Tuning of Magnetocaloric Mn1+xSb. *Chem. Mater.* **32**, 1243-1249 (2020).
16. J. P. Perdew, Burke, K., & Ernzerhof, M., Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **77**, 3865 (1996).
17. G. Kim, S. V. Meschel, P. Nash, W. Chen, Experimental formation enthalpies for intermetallic phases and other inorganic compounds. *Sci Data* **4**, 170162 (2017).
18. C. J. Bartel *et al.*, A critical examination of compound stability predictions from machine-learned formation energies. *npj Comput. Mater.* **6**, (2020).
19. M. Aykol, Dwaraknath, S. S., Sun, W., & Persson, K. A., Thermodynamic limit for synthesis of metastable inorganic materials. *Sci. Adv.* **4**, eaaq0148 (2018).
20. W. Sun, Dacek, S.T., Ong, S.P., Hautier, G., Jain, A., Richards, W.D., Gamst, A.C., Persson, K.A. and Ceder, G., The thermodynamic scale of inorganic crystalline metastability. *Sci. Adv.* **2**, e1600225 (2016).
21. H. Yamada *et al.*, Predicting Materials Properties with Little Data Using Shotgun Transfer Learning. *ACS Cent Sci* **5**, 1717-1730 (2019).

22. A. D. Sendek *et al.*, Machine Learning-Assisted Discovery of Solid Li-Ion Conducting Materials. *Chem. Mater.* **31**, 342-352 (2018).
23. E. D. Cubuk, A. D. Sendek, E. J. Reed, Screening billions of candidates for solid lithium-ion conductors: A transfer learning approach for small data. *J Chem Phys* **150**, 214701 (2019).
24. K. T. Schutt, H. E. Sauceda, P. J. Kindermans, A. Tkatchenko, K. R. Muller, SchNet - A deep learning architecture for molecules and materials. *J Chem Phys* **148**, 241722 (2018).
25. T. Xie, J. C. Grossman, Hierarchical visualization of materials space with graph convolutional neural networks. *J Chem Phys* **149**, 174111 (2018).
26. R. E. A. Goodall, A. A. Lee, Predicting materials properties without crystal structure: deep representation learning from stoichiometry. *Nat Commun* **11**, 6280 (2020).
27. O. Isayev *et al.*, Universal fragment descriptors for predicting properties of inorganic crystals. *Nat Commun* **8**, 15679 (2017).
28. D. Jha *et al.*, Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. *Nat Commun* **10**, 5316 (2019).
29. W. Ye, C. Chen, Z. Wang, I. H. Chu, S. P. Ong, Deep neural networks for accurate predictions of crystal stability. *Nat Commun* **9**, 3800 (2018).
30. A. Chandrasekaran *et al.*, Solving the electronic structure problem with machine learning. *npj Comput. Mater.* **5**, 1 (2019).
31. L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2**, (2016).
32. T. Xie, J. C. Grossman, Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys Rev Lett* **120**, 145301 (2018).
33. S. Gong *et al.*, Predicting charge density distribution of materials using a local-environment-based graph convolutional network. *Phys. Rev. B* **100**, 184103 (2019).
34. S. Gong *et al.*, Classifying superheavy elements by machine learning. *Phys. Rev. A* **99**, (2019).
35. Z. Shi *et al.*, Deep elastic strain engineering of bandgap through machine learning. *Proc Natl Acad Sci U S A* **116**, 4117-4122 (2019).
36. B. Meredig *et al.*, Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **89**, (2014).
37. D. Jha *et al.*, ElemNet: Deep Learning the Chemistry of Materials From Only Elemental Composition. *Sci Rep* **8**, 17593 (2018).
38. R. Sarmiento-Perez, S. Botti, M. A. Marques, Optimized Exchange and Correlation Semilocal Functional for the Calculation of Energies of Formation. *J Chem Theory Comput* **11**, 3844-3850 (2015).

39. V. L. Chevrier, S. P. Ong, R. Armiento, M. K. Y. Chan, G. Ceder, Hybrid density functional calculations of redox potentials and formation energies of transition metal compounds. *Phys. Rev. B* **82**, (2010).
40. J. Sun, A. Ruzsinszky, J. P. Perdew, Strongly Constrained and Appropriately Normed Semilocal Density Functional. *Phys Rev Lett* **115**, 036402 (2015).
41. Y. Zhang *et al.*, Efficient first-principles prediction of solid stability: Towards chemical accuracy. *npj Comput. Mater.* **4**, (2018).
42. O. Egorova, R. Hafizi, D. C. Woods, G. M. Day, Multifidelity Statistical Machine Learning for Molecular Crystal Structure Prediction. *J Phys Chem A* **124**, 8065-8078 (2020).
43. A. Jain *et al.*, A high-throughput infrastructure for density functional theory calculations. *Comput. Mater. Sci.* **50**, 2295-2310 (2011).
44. K. Persson *et al.*, A Framework for Quantifying Uncertainty in DFT Energy Corrections. *doi.org/10.26434/chemrxiv.14593476.v1*, (2021).
45. R. Friedrich *et al.*, Coordination corrected ab initio formation enthalpies. *npj Comput. Mater.* **5**, (2019).
46. Y. Zhang, C. Ling, A strategy to apply machine learning to small datasets in materials science. *npj Comput. Mater.* **4**, 1 (2018).
47. T. Xie, Bapst, V., Gaunt, A.L., Obika, A., Back, T., Hassabis, D., Kohli, P. and Kirkpatrick, J., Atomistic graph networks for experimental materials property prediction. *arXiv:2103.13795*, (2021).
48. J. S. Smith *et al.*, Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat Commun* **10**, 2903 (2019).
49. T. Zhu *et al.*, Charting lattice thermal conductivity for inorganic crystals and discovering rare earth chalcogenides for thermoelectrics. *Energy Environ. Sci.* **14**, 3559-3566 (2021).
50. S. Kong, D. Guevarra, C. P. Gomes, J. M. Gregoire, Materials representation and transfer learning for multi-property prediction. *Appl. Phys. Rev.* **8**, (2021).
51. C. Chen, Y. Zuo, W. Ye, X. Li, S. P. Ong, Learning properties of ordered and disordered materials from multi-fidelity data. *Nat. Comput. Sci.* **1**, 46-53 (2021).
52. G. Pilania, J. E. Gubernatis, T. Lookman, Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Comput. Mater. Sci.* **129**, 156-163 (2017).
53. A. Patra *et al.*, A multi-fidelity information-fusion approach to machine learn and predict polymer bandgap. *Comput. Mater. Sci.* **172**, 109286 (2020).
54. R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, Big Data Meets Quantum Chemistry Approximations: The Delta-Machine Learning Approach. *J Chem Theory Comput* **11**, 2087-2096 (2015).
55. I. Hurtado, D. Neuschutz. (Springer, Berlin (1999–2005), 1999).
56. L. Ward *et al.*, Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **152**, 60-69 (2018).

57. K. Weiss, T. M. Khoshgoftaar, D. Wang, A survey of transfer learning. *J. Big Data* **3**, (2016).
58. T. Liu, A. Abd-Elrahman, J. Morton, V. L. Wilhelm, 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. *GIsci Remote Sens* **55**, 243-264 (2018).
59. A. Folleco, Khoshgoftaar, T.M., Van Hulse, J. and Bullard, L., Identifying Learners Robust to Low Quality Data. *2008 IEEE International Conference on Information Reuse and Integration*, 190 (2008).
60. C. W. Park, C. Wolverton, Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Phys. Rev. Mater.* **4**, 063801 (2020).
61. Y. Yu, M. Aykol, C. Wolverton, Reaction thermochemistry of metal sulfides with GGA andGGA+Ucalculations. *Phys. Rev. B* **92**, (2015).
62. M. Aykol, C. Wolverton, Local environment dependentGGA+Umethod for accurate thermochemistry of transition metal compounds. *Phys. Rev. B* **90**, (2014).
63. A. M. Nolan, Y. Zhu, X. He, Q. Bai, Y. Mo, Computation-Accelerated Design of Materials and Interfaces for All-Solid-State Lithium-Ion Batteries. *Joule* **2**, 2016-2046 (2018).
64. W. Sun *et al.*, Thermodynamic Routes to Novel Metastable Nitrogen-Rich Nitrides. *Chem. Mater.* **29**, 6936-6946 (2017).
65. H. J. R. o. P. i. P. Jones, Splat cooling and metastable phases. *Rep. Prog. Phys.* **36**, 1425 (1973).
66. T. Takeuchi *et al.*, Improvement of cycle capability of FeS2 positive electrode by forming composites with Li2S for ambient temperature lithium batteries. *J. Electrochem. Soc.* **159**, A75 (2011).
67. M. Fang *et al.*, The investigation of lithium doping perovskite oxide LiMnO3 as possible LIB anode material. *Ceram.* **44**, 8223-8231 (2018).
68. A. J. Darling, S. Stewart, C. F. Holder, R. E. J. C. Schaak, Bulk ‐ immiscible AgRh Alloy Nanoparticles as a Highly Active Electrocatalyst for the Hydrogen Evolution Reaction. *ChemNanoMat* **6**, 1320-1324 (2020).
69. H. Wang *et al.*, One-pot synthesis of bi-metallic PdRu tripods as an efficient catalyst for electrocatalytic nitrogen reduction to ammonia. *J. Mater. Chem. A* **7**, 801-805 (2019).
70. K.-L. Wu, R. Yu, X.-W. J. C. Wei, Monodispersed FeNi 2 alloy nanostructures: solvothermal synthesis, magnetic properties and size-dependent catalytic activity. *CrystEngComm* **14**, 7626-7632 (2012).
71. C.-C. Lee, Y.-Y. Cheng, H. Y. Chang, D.-H. J. J. o. A. Chen, Synthesis and electromagnetic wave absorption property of Ni–Ag alloy nanoparticles. *J. Alloys Compd.* **480**, 674-680 (2009).
72. M. Asano, K. Umeda, A. J. J. J. o. A. P. Tasaki, Cu3N thin film for a new light recording media. *Jpn. J. Appl. Phys.* **29**, 1985 (1990).

73. S. Ono *et al.*, Band-structure design of fluoride complex materials for deep-ultraviolet light-emitting diodes. *Jpn. J. Appl. Phys.* **44**, 7285 (2005).
74. Z. Li, W. Wang, P. Zhou, Z. Ma, L. J. N. J. o. P. Sun, New type of hybrid nodal line semimetal in Be2Si. *New J. Phys.* **21**, 033018 (2019).
75. Y.-Q. Xu, B.-G. Liu, D. J. P. R. B. Pettifor, Half-metallic ferromagnetism of MnBi in the zinc-blende structure. *Phys. Rev. B* **66**, 184435 (2002).
76. M. Gjoka, I. Panagiotopoulos, D. J. J. o. m. p. t. Niarchos, Structure and magnetic properties of Sm (Co1− xMx) 5 (M= Cu, Ag) alloys. *J. Mater. Process. Technol.* **161**, 173-175 (2005).
77. B. R. McGarvey, A. Reuveni, in *Magnetic Resonance and Related Phenomena*. (Springer, 1979), pp. 121-121.
78. K. Sugiyama *et al.*, High-field magnetization of USn 3 and UPb 3. *J. Phys. Soc. Japan* **71**, 326-331 (2002).
79. B. Balamurugan, B. Das, W. Zhang, R. Skomski, D. J. J. J. o. P. C. M. Sellmyer, Hf–Co and Zr–Co alloys for rare-earth-free permanent magnets. *J. Phys. Condens. Matter* **26**, 064204 (2014).
80. V. J. Yannello, E. Lu, D. C. J. I. C. Fredrickson, At the Limits of Isolobal Bonding: π-Based Covalent Magnetism in Mn2Hg5. *Inorg. Chem.* **59**, 12304-12313 (2020).
81. L. W. Kammerdiner, "Film deposition of Nb-based A15 superconductors," (California Univ., 1975).
82. V. Volodin, E. Zhakanbaev, A. Z. Tuleushev, Y. J. V. N. n. Y. T. R. K. Tuleushev, Synthesis and structure of new intermetallic compound Ta 3 Pb. *Vestnik Natsional'nogo Yadernogo Tsentra Respubliki Kazakhstan* **4**, 49-54 (2005).
83. H. Masumoto, K. J. J. o. t. P. S. o. J. Watanabe, 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.* **32**, 281-281 (1972).
84. M. Yin, P. J. T. J. o. C. T. Nash, Standard enthalpies of formation of selected XYZ half-Heusler compounds. *J. Chem. Thermodyn.* **91**, 1-7 (2015).
85. J. Xu *et al.*, 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* **7**, 26858-26866 (2019).
86. X. Li *et al.*, Computational Screening of New Perovskite Materials Using Transfer Learning and Deep Learning. *Appl. Sci.* **9**, (2019).
87. Y. Zhou *et al.*, Adsorption of acetylene on ordered NixAg1-x/Ni (111) and effect of Ag-dopant: A DFT study. *Appl. Surf. Sci.* **435**, 521-528 (2018).
88. I. Dudenkov, K. J. R. j. o. i. c. Solntsev, Theoretical prediction of the new high-density lithium boride LiB 11 with polymorphism and pseudoplasticity. *Russ. J. Inorg. Chem.* **54**, 1261-1272 (2009).
89. H.-J. Zhang *et al.*, Topological insulators in ternary compounds with a honeycomb lattice. *Phys. Rev. Lett.* **106**, 156402 (2011).

90. W. Hou *et al.*, Prediction of crossing nodal-lines and large intrinsic spin Hall conductivity in topological Dirac semimetal Ta3As family. *npj Comput. Mater.* **7**, (2021).
91. L. Guan-Nan, J. J. C. P. L. Ying-Jiu, First-principles study on the half-metallicity of half-Heusler alloys: XYZ (X= Mn, Ni; Y= Cr, Mn; Z= As, Sb). *Chin. Phys. Lett.* **26**, 107101 (2009).
92. Y. Li *et al.*, Structural, magnetic properties, and electronic structure of hexagonal FeCoSn compound. *Chinese Phys. B* **27**, 026101 (2018).
93. S. M. Lundberg *et al.*, From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell* **2**, 56-67 (2020).
94. S. Grindy, B. Meredig, S. Kirklin, J. E. Saal, C. Wolverton, Approaching chemical accuracy with density functional calculations: Diatomic energy corrections. *Phys. Rev. B* **87**, (2013).
95. F. Pedregosa *et al.*, Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825-2830 (2011).

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.

Supporting Information for:

Calibrating DFT formation energy calculations by multi-fidelity machine learning

Sheng Gong[1], Shuo Wang[2], Tian Xie[3], Woo Hyun Chae[1], Runze Liu[1], and Jeffrey C. Grossman[1,*]

[1]Department of Materials Science and Engineering, Massachusetts Institute of Technology, MA 02139, USA
[2]Department of Materials Science and Engineering, University of Maryland, MD 20742, USA
[3]Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, MA 02139, USA

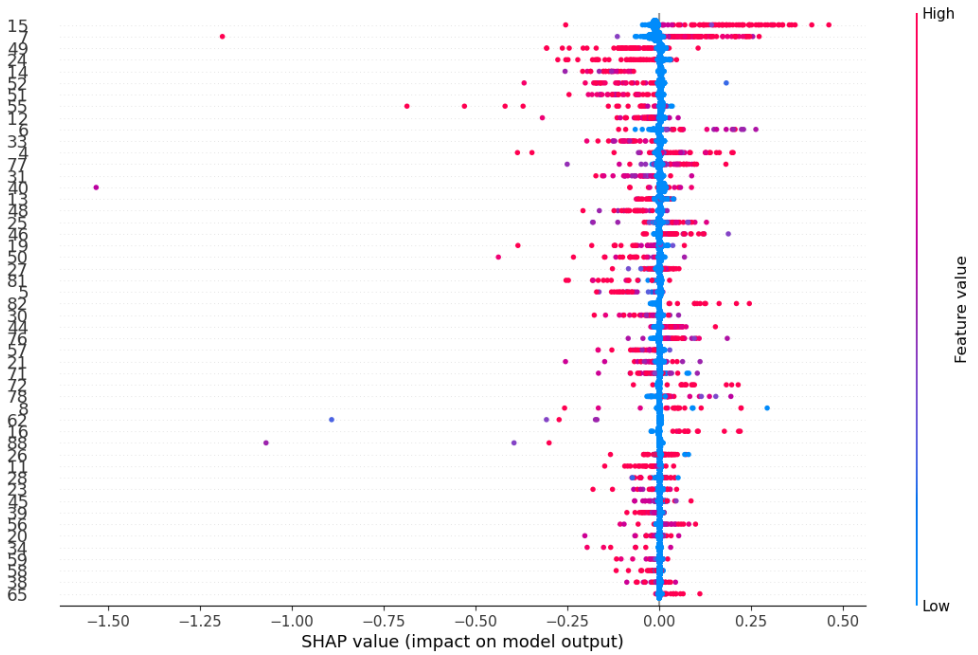

Figure S1. Impacts of top 50 impactful elements on diff. ΔH_f. Here the numbers in the left are the atomic numbers minus 1 (for example, 15 corresponds to S and 7 to O), and the elements are arranged in a sequence with decreasing impacts from the top to the bottom.

Table S1. Comparison of mean average errors (MAE) and standard deviations (std) between predictions of formation enthalpies and experimental measurements. MAE and std values are from the same test set, and each machine learning model is trained 10 times. RF denotes random forest, MLP denotes multilayer perceptron, and ROOST(*26*) and CGCNN(*32*) are two deep-learning models that automatically extract materials' fingerprints from compositions and structures, respectively. Here, "struct." means the model is trained with structural features, "no struct." the model is trained with only compositional features, "dft." the model is trained with DFT formation enthalpies as input, "trans" the model is trained in a transfer learning manner, "diff." the model is trained on difference between DFT and experimental formation enthalpies, "exp." the model is directly trained on experimental formation enthalpies.

Models	MAE (eV/atom)	std (eV/atom)

DFT	0.0955	-
struct. RF diff.	0.0669	0.0014
dft. struct. RF diff.	0.0638	0.0009
no struct. RF diff.	0.0646	0.0006
dft. no struct. RF diff.	0.0617	0.0006
struct. RF exp.	0.1339	0.0008
dft. struct. RF exp.	0.0785	0.0010
no struct. RF exp.	0.1294	0.0019
dft. no struct. RF exp.	0.0758	0.0007
struct. MLP diff.	0.0772	0.0086
dft. struct. MLP diff.	0.0737	0.0040
no struct. MLP diff.	0.0676	0.0019
dft. no struct. MLP diff.	0.0664	0.0035
trans. no struct. MLP diff.	0.0764	0.0039
struct. MLP exp.	0.1350	0.0030
dft. struct. MLP exp.	0.1210	0.0130
no struct. MLP exp.	0.1090	0.0037
dft. no struct. MLP exp.	0.0907	0.0062
trans. no struct. MLP exp.	0.0774	0.0012
ROOST diff.	0.0809	0.0018
trans. ROOST diff.	0.0725	0.0044
ROOST exp.	0.1042	0.0023
trans. ROOST exp.	0.0752	0.0033

CGCNN diff.	0.0907	0.0065
trans. CGCNN diff.	0.0806	0.0022
CGCNN exp.	0.1337	0.0043
trans. CGCNN exp.	0.0925	0.0052

Table S2. Comparison between exp. ΔH_f, empirically corrected PBE ΔH_f (co-PBE ΔH_f), machine learning (the best RF model) calibrated ΔH_f from this work and ΔH_f from more sophisticated density functionals (PBEfe(*38*) and SCAN(*40*)). ΔH_f is in the unit of eV/atom. Here, the value of ML calibrated ΔH_f for each material is obtained by excluding that material and using all remaining data to train a model and predict the excluded material.

Materials	exp.(*17, 55*)	co-PBE(*2*)	ML calibrated	PBEfe(*38*)	SCAN(*41*)
Ag_2O	-0.1075	-0.3277	-0.2892	-0.252	-0.189
Ag_2S	-0.1133	-0.2762	-0.1652	-0.159	-0.069
Ag_2Se	-0.1503	-0.0672	-0.1815	-0.185	-0.079
AlAs	-0.6028	-0.4846	-0.6272		-0.595
$AlCl_3$	-1.8283	-1.9513	-1.8600	-1.634	-1.783
AlF_3	-3.9136	-3.8921	-3.8649	-3.900	-4.113
AlN	-1.6478	-1.5959	-1.6601		-1.728
Al_2O_3	-3.4735	-3.4357	-3.4398	-3.442	-3.504
Al_2S_3	-1.3442	-1.4546	-1.3801	-1.276	-1.233
Al_2Se_3	-1.1752	-0.8400	-1.0259	-0.994	-0.981
Al_2Te_3	-0.6609	-0.4513	-0.6395	-0.548	-0.496
Al_2NiO_4	-2.8367	-2.7576	-2.7781	-2.841	

BaO	-2.8398	-2.8313	-2.8379		-2.829
BaO_2	-2.1913	-2.1710	-2.1622		-2.124
BaS	-2.4356	-2.4103	-2.3752		-2.352
$BaCrO_4$	-2.4978	-2.4739	-2.4114	-3.163	
Be_3N_2	-1.2194	-1.2273	-1.2295	-1.383	-1.399
BeO	-3.1580	-3.1105	-3.2496		-3.215
BeS	-1.2230	-1.3884	-1.4432		-1.338
$CaCl_2$	-2.7479	-2.7927	-2.7248	-2.558	-2.725
CaF_2	-4.2424	-4.2456	-4.3458		-4.456
CaO	-3.2902	-3.3164	-3.3134		-3.353
CaS	-2.4615	-2.4832	-2.4676		-2.432
CoF_2	-2.3251	-2.4496	-2.3557	-2.172	-2.137
CoF_3	-2.0521	-2.2397	-2.1508	-2.035	-1.901
Co_3S_4	-0.5152	-0.8174	-0.5636	-0.541	-0.350
CrO_2	-2.0092	-2.0628	-1.9934	-2.306	-2.337
Cr_2O_3	-2.3643	-2.3814	-2.3250	-2.435	-2.392
CrN	-0.6063	-0.7028	-0.7196		-0.673
Cr_2FeO_4	-2.1480	-2.2614	-2.2257	-2.168	
CuS	-0.2850	-0.5254	-0.3612	-0.324	-0.250
CuO	-0.8073	-0.9504	-0.9883		-0.791
Cu_2O	-0.5894	-0.6496	-0.6042	-0.656	-0.501
CuSe	-0.2168	-0.1234	-0.2819	-0.211	-0.143
FeO	-1.4147	-1.7280	-1.6540		-1.191

FeS	-0.5250	-0.8385	-0.6638		-0.365
Fe_3O_4	-1.6596	-1.8587	-1.7574	-1.632	-1.706
Fe_2NiO_4	-1.6057	-1.6825	-1.5866	-1.539	
GaS	-1.0841	-0.9835	-0.9336	-0.708	-0.699
GaSb	-0.2155	-0.1566	-0.2711		-0.153
GaSe	-0.8239	-0.5912	-0.7986	-0.596	-0.606
Ga_2O_3	-2.2615	-2.2782	-2.1330	-2.053	-2.187
Ga_2S_3	-1.0702	-1.0605	-0.9648	-0.724	-0.746
Ga_2Se_3	-0.8412	-0.5844	-0.7782	-0.592	-0.632
$CdCl_2$	-1.3530	-1.4887	-1.3333		-1.322
CdO	-1.3421	-1.3807	-1.3235		-1.325
CdS	-0.7742	-0.9724	-0.9035		-0.736
CdSe	-0.7504	-0.6280	-0.7380		-0.692
CoO	-1.2333	-1.4100	-1.3603		-0.909
CoS	-0.5078	-0.7416	-0.4996		-0.146
$CoSb_3$	-0.1626	-0.1877	-0.2338		-0.156
Co_3O_4	-1.3562	-1.3799	-1.2652		-1.071
GeS	-0.3171	-0.5846	-0.3896	-0.338	-0.222
GeSe	-0.3578	-0.2091	-0.3295	-0.246	-0.171
HfN	-1.9132	-1.9386	-1.8941		-1.982
HfO_2	-3.8541	-4.0300	-3.8774	-3.949	-4.026
HgS	-0.3057	-0.4593	-0.2579		-0.137
HgSe	-0.3078	-0.2058	-0.2399		-0.195

HgTe	-0.2349	-0.1634	-0.2003		-0.168
LaN	-1.5719	-1.5128	-1.5365		-1.627
LaS	-2.3633	-2.4948	-2.4683		-2.378
InS	-0.6938	-0.8231	-0.7485	-0.600	-0.646
InSb	-0.1581	-0.1027	-0.1694		-0.174
In_2O_3	-1.9132	-2.0062	-1.9403	-1.799	-1.952
In_2S_3	-0.7372	-0.9310	-0.8354	-0.621	-0.692
KSb	-0.4727	-0.4308	-0.5128	-0.542	-0.53
KSb_2	-0.3708	-0.2990	-0.3966	-0.371	-0.354
K_2O	-1.2496	-1.2530	-1.2713		-1.27
K_2S	-1.3009	-1.2762	-1.2734		-1.253
K_2Se	-1.3298	-1.0966	-1.2079		-1.284
La_2O_3	-3.7219	-3.8833	-3.8612	-3.805	
La_2S_3	-2.5328	-2.5906	-2.5197	-2.404	
Li_2O	-2.0655	-2.0667	-2.3858		-2.113
Li_2S	-1.5423	-1.5572	-1.6849		-1.527
Li_2Se	-1.3877	-1.2605	-1.4105		-1.436
Li_3N	-0.5117	-0.4625	-0.5178		-0.510
MgO	-3.1178	-3.0624	-3.0701		-3.121
MgS	-1.8034	-1.7593	-1.7280		-1.658
MgSe	-1.4137	-1.2499	-1.4038		-1.440
MgTe	-1.0841	-0.8773	-1.0539		-0.991
Mg_3Sb_2	-0.6222	-0.3645	-0.6352	-0.474	

NaSb	-0.3425	-0.3280	-0.4201		-0.404
Na_2O	-1.4440	-1.4503	-1.4731		-1.476
Na_2S	-1.2937	-1.2916	-1.2282		-1.271
Na_2Se	-1.1853	-1.0687	-1.2209		-1.257
Na_3Sb	-0.5117	-0.4360	-0.5064	-0.552	
NiO	-1.2422	-1.0374	-0.9415		-0.765
Ni_3S_2	-0.4484	-0.6362	-0.4794		-0.393
NbO_2	-2.7464	-2.9078	-2.7571	-2.785	-2.785
Ni_3S_2	-0.4484	-0.6362	-0.4794	-0.454	
PdS_2	-0.2703	-0.7134	-0.4302	-0.342	-0.369
RbS	-1.2026	-1.2146	-1.2262		-1.198
PtS	-0.4315	-0.7374	-0.4449		-0.613
PtS_2	-0.3816	-0.7918	-0.4402		-0.494

Table S3. Search space of hyper-parameters for random forest. Setting of other parameters not specified in this table are available in the provided code.

Parameters	Values searched
n_estimators	100, 200
max_depth	None, 16, 64
min_samples_split	2, 4, 8

Table S4. Search space of hyper-parameters for multi-layer perceptron. Setting of other parameters not specified in this table are available in the provided code.

Parameters	Values searched
depth	3, 4, 5, 6, 7, 8, 9, 10
width	199, 299, 399

Table S5. Search space of hyper-parameters for multi-layer perceptron with transfer learning. Setting of other parameters not specified in this table are available in the provided code.

Parameters	Values searched
ratio of layers transferred	0.1, 0.3, 0.5, 0.7, 0.9
learning_rate	1e-6, 1e-5, 1e-4, 1e-3, 1e-2

Table S6. Search space of hyper-parameters for ROOST(*28*). Setting of other parameters not specified in this table are available in the provided code.

Parameters	Values searched
n-graph	2, 3, 4
elem-fea-len	8, 16, 32, 64
weight-decay	0, 1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1

Table S7. Search space of hyper-parameters for ROOST(*28*) with transfer learning. Setting of other parameters not specified in this table are available in the provided code.

Parameters	Values searched
ratio of layers transferred	0.2, 0.4, 0.6, 0.8
learning-rate	9e-6, 5e-5, 3e-4

weight-decay	0, 1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1

Table S8. Search space of hyper-parameters for CGCNN(*32*). Setting of other parameters not specified in this table are available in the provided code.

Parameters	Values searched
n-conv	2, 3, 4
n-h	1, 2
atom-fea-len	8, 16
h-fea-len	16, 32
weight-decay	0, 1e-5

Table S9. Search space of hyper-parameters for CGCNN(*32*). Setting of other parameters not specified in this table are available in the provided code.

Parameters	Values searched
number of layers transferred	20, 22, 24, 26, 28
learning_rate	1e-6, 1e-5, 1e-4, 1e-3, 1e-2

View publication stats