A critical examination of stability predictions from machine-learned formation energies

Christopher J. Bartel[1*], Amalie Trewartha[1], Qi Wang[2], Alex Dunn[1,2], Anubhav Jain[2], Gerbrand Ceder[1,3*]

[1]Department of Materials Science & Engineering, University of California, Berkeley, Berkeley, CA 94720, USA
[2]Energy Technologies Area, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[3]Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[*]correspondence to cbartel@berkeley.edu, gceder@berkeley.edu

Version: 3.0
Last edit: 11/13/19; CJB
Target: *Matter* or *npj Computational Materials*

Abstract

The formation energy does not directly determine the stability of an inorganic crystalline solid. Instead, the formation energies must be compared for all compounds within a chemical space to obtain the energy of decomposition with respect to these competing compounds. A number of machine learning models for predicting the formation energy of solid-state materials have emerged in recent years, but none have rigorously addressed the application of these models to stability predictions. In this work, seven machine learning (ML) models for formation energy are tested on stability predictions using the Materials Project database of density functional theory (DFT) calculations for 85,014 unique chemical compositions. These models include five recently published compositional representations, a baseline model using stoichiometry alone, and a structural representation. While the mean absolute error (MAE) on decomposition energies is comparable to the MAE on formation energies and also the MAE of DFT with respect to experiment for formation energies, this accuracy is shown to be insufficient to make accurate stability predictions. Most critically, in sparse chemical spaces where there are many more unstable than stable compounds, only the structural model is capable of efficiently detecting which materials are actually stable. This work demonstrates that accurate predictions of formation energy do not imply accurate predictions of stability, especially for problems that resemble real-world application of these models for materials discovery.

Introduction

Machine learning (ML) is emerging as a standard tool for rapid prediction of material properties.[1–6] In general, these predictions are made by fitting statistical models on thousands of density functional theory (DFT) calculations housed in one of the many open materials databases.[7–11] In principle, once these models are trained on this immense set of quantum chemical data, the determination of properties for new materials can be made in orders-of-magnitude less time using the trained models, rather than computationally expensive DFT calculations.

Of particular interest is the use of machine learning to discover new materials. The combinatorics of materials discovery make for an immensely challenging problem – if we consider the possible combinations of just four elements (*A*, *B*, *C*, *D*), from any of the ~80 elements that are technologically relevant, there are already ~1.6 million quaternary chemical spaces to consider. This is before we consider such factors as stoichiometry ($ABCD_2$, $AB_2C_3D_4$, etc.) or structure (cubic, hexagonal, layered, etc.), each of which add substantially to the combinatoric complexity. The Inorganic Crystal Structure Database (ICSD) of known solid-state materials contains ~10^5 entries,[12] several orders of magnitude less than the 10^{10} quaternary compositions identified as plausible using electronegativity and charge-based rules.[13] This suggests that 1) there is ample opportunity for new materials discovery and 2) the problem of finding stable materials may resemble the needle-in-a-haystack problem, with many unstable compositions for each stable one. The immensity of this problem is a natural fit for high-throughput machine learning techniques.

In this work, we closely examine whether recently published machine learning models for formation energy are capable of distinguishing the relative stability of chemically similar materials and provide a roadmap for doing the same for future models. We show that while the formation energy of compounds from elements can be learned with high accuracy using a variety of machine learning approaches, these learned formation energies do not reproduce DFT-calculated stabilities. While the accuracy of these models for formation energy approach the DFT error (relative to experiment), DFT predictions benefit from a systematic cancellation of error when making stability predictions while ML models do not. Of particular concern is the high rate of materials predicted to be stable that are not confirmed to be stable by DFT, impeding the use of these models to efficiently discover new materials.

Results and Discussion

The relationship between formation energy and stability

A necessary condition for a material to be used for any application is stability (under some conditions); otherwise, any predicted or calculated properties are purely academic, and the material will never find its way into use. The thermodynamic stability of a material is defined by its Gibbs energy of decomposition, ΔG_d, which is the Gibbs formation energy, ΔG_f, of a specified material relative to all other compounds in the relevant chemical space. Temperature-dependent thermodynamics are not yet tractable with high-throughput DFT and have only sparsely been addressed with ML,[14] so material stability is primarily assessed instead using the decomposition enthalpy, ΔH_d.[15–18]

ΔH_d is obtained by a convex hull construction in formation enthalpy (ΔH_f)-composition space. **Figure 1a** shows this construction for a chemical space, *A*-*B*, having three known compounds – A_4B, A_2B, and AB_3. The convex hull of largest area is drawn from all possible points in the composition space (blue line), where stable compositions lie on the convex hull, and unstable compositions lie above the hull. A_4B is unstable (above the hull), so $\Delta H_d > 0$ and is calculated as the distance in ΔH_f between A_4B and the convex hull of stable points. AB_3 is stable (on the hull), so $\Delta H_d < 0$ and is calculated as the distance in ΔH_f between AB_3 and a hypothetical convex hull constructed without AB_3 (dashed line). $|\Delta H_d|$ is therefore the minimum that ΔH_f must decrease for an unstable compound to become stable or the maximum amount that ΔH_f can increase for a stable compound to remain stable. This approach generalizes for chemical spaces comprised of any number of elements.

The thermodynamic driving force for (in)stability with respect to competing compounds is quantified by ΔH_d, which arises from the relative ΔH_f for all compounds within a chemical space. Despite this, the standard thermodynamic property that is predicted by ML models is the absolute ΔH_f.[19–27] Using data available in the Materials Project (MP),[15] we applied the convex hull construction to obtain ΔH_d for 85,014 inorganic crystalline solids (the majority of which are in the ICSD) and compare ΔH_d to ΔH_f in **Figure 1b**. It is clear that effectively no linear correlation exists between ΔH_d and ΔH_f, except for the trivial case where only a single compound exists in a chemical space ($\Delta H_d = \Delta H_f$), which is true for only ~3% of materials in MP and is likely indicative of incomplete phase diagrams in the database. While ΔH_f somewhat uniformly spans a wide range of energies (mean ± average absolute deviation = -1.42 ± 0.95 eV/atom), ΔH_d spans a much smaller

energy window (0.06 ± 0.12 eV/atom), suggesting ΔH_d is a more sensitive or subtle quantity to predict (a histogram of ΔH_f and ΔH_d is provided in **Figure S1**). Still, while no linear correlation exists between ΔH_d and ΔH_f, and ΔH_d occurs over a much smaller energy range, it is possible for ΔH_f models to predict ΔH_d as long as the relative differences in ΔH_f within a given chemical space are predicted with accuracy comparable to the range of variation in ΔH_d.

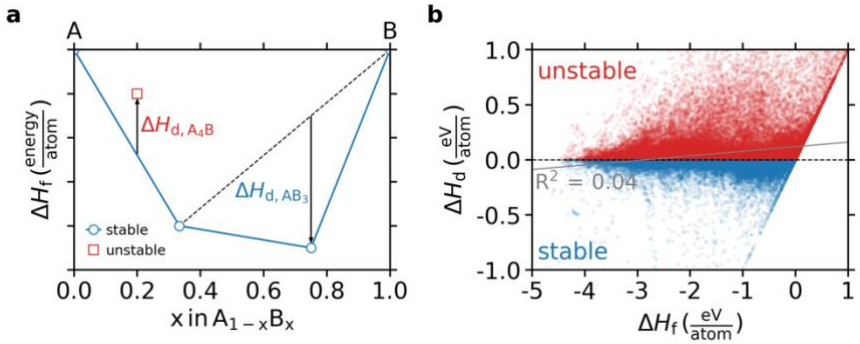

Figure 1. a) Illustration of the convex hull construction to obtain the decomposition enthalpy, ΔH_d, from the formation enthalpy, ΔH_f. **b)** The decomposition enthalpy, ΔH_d, shown against the formation enthalpy, ΔH_f, for 85,014 ground-state entries in Materials Project, indicating effectively no correlation between the two quantities.

Learning formation energy from chemical composition

Machine learning material properties requires that an arbitrary material is "represented" by a set of attributes (features). This representation can be as simple as a vector corresponding with the fractional amount of each element in the compound (e.g., Li_2O = [0, 0, 2/3, 0, 0, 0, 0, 1/3, 0, 0, …], where the length of vector is the number of elements in the periodic table) or a vector that includes substantial physical or chemical information about the material. In the search for new materials, the structure is rarely known *a priori*, and instead a list of compositions with unknown structure is screened for stability. In this case, the material representation is constructed only from the chemical formula without including properties such as the geometric or electronic structures. These models, which take as input the chemical formula and output thermodynamic predictions, are henceforth referred to as compositional models here.

In this work, we assess the potential for five recently introduced compositional models – prb14,[19] prb16,[20] npj16,[21] auto,[22] and arXiv19[23] – to predict the stability of compounds in MP. The models in prb14, prb16, npj16, and auto include chemical information for each element in their material representations from quantities such as atomic electronegativities, radii, and elemental group. arXiv19 differs in that no *a priori* information other than the stoichiometry is

used as input; instead the representation and fit are learned simultaneously using a graph neural network model. In addition, we include a baseline model for comparison, elfrac, where the representation is simply the stoichiometric fraction of each element in the formula. Once the representation is defined, an algorithm (e.g., a neural network) learns the relationship between that representation and the target property of interest (e.g., ΔH_f) during training. Because compositional models necessarily make the same prediction for all structures having the same formula, all analysis in this work is performed using the lowest energy (ground-state) structure for each MP compound. Additional details on the training of each model is available in the **Methods** section.

Parity plots comparing ΔH_f in MP ($\Delta H_{f,MP}$) to machine-learned ΔH_f ($\Delta H_{f,pred}$) for each model are shown in **Figure 2**. It is clear that each published representation substantially improves upon the baseline elfrac model, decreasing the mean absolute error (MAE) by 27-74%. This increased accuracy is attributed to the increased complexity of the representation. The MAE between MP and these ML models is comparable to the expected disagreement between MP and experimentally obtained ΔH_f,[8,15,28–30] implying a substantial amount of the information required to determine ΔH_f is contained in the composition (and not the structure). The success of ML models for predicting ΔH_f is not surprising considering the historical context of simple heuristics that perform relatively well at predicting the driving force for the formation of compounds from elements – e.g., the Miedema model.[31]

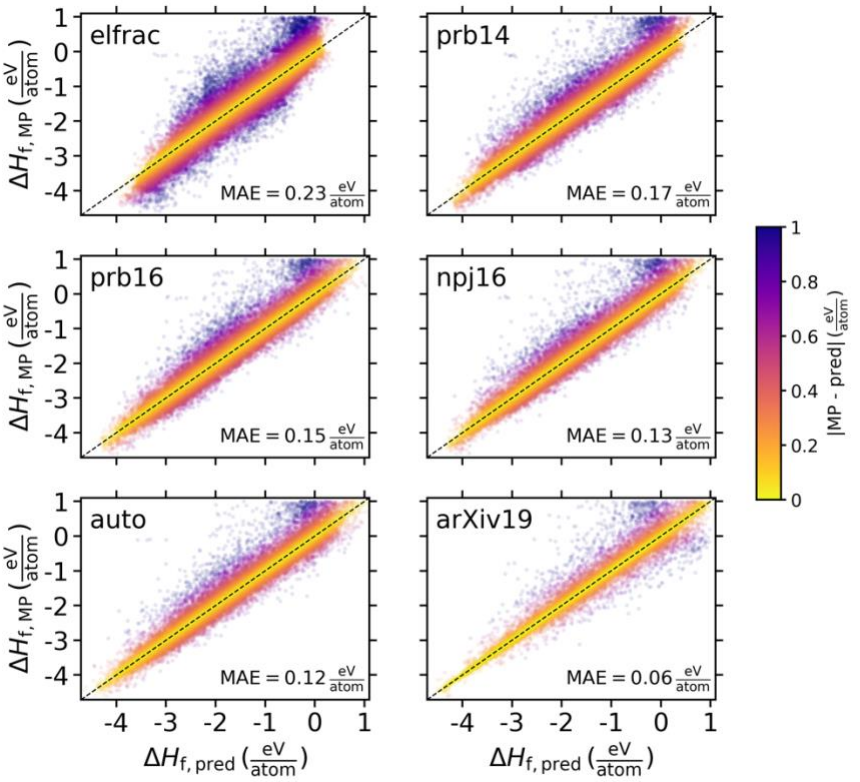

Figure 2. Parity plot for formation enthalpy predictions using six different machine learning models that take as input the chemical formula and output the formation enthalpy. elfrac refers to a baseline model that parametrizes each formula only by the stoichiometric coefficient of each element. prb14, prb16, npj16, auto, and arXiv19 refer to the models in Refs. [19–23], respectively. $\Delta H_{f,pred}$ corresponds with ML predictions for aggregated hold-out sets during 5-fold cross-validation of the Materials Project dataset (see **Methods** for details). $\Delta H_{f,MP}$ refers to the formation energy per atom in the MP database. The absolute error on ΔH_f is shown as the colorbar and the mean absolute error (MAE) is shown within each panel.

Implicit stability predictions from learned formation enthalpies

While ML predictions of DFT-calculated ΔH_f approach the accuracy of DFT compared with experiment for ΔH_f (~0.1-0.2 eV/atom)[8,15,28,29,32] over the entire MP dataset, the use of ΔH_f for stability predictions requires that relative ΔH_f between chemically similar compounds must be accurately predicted. To assess the accuracy of relative ΔH_f, we reconstructed the convex hulls for all chemical spaces in MP with $\Delta H_{f,pred}$ to obtain $\Delta H_{d,pred}$ using each compositional model. Parity plots for ΔH_d are shown in **Figure 3**. In contrast to ΔH_f, where all representations substantially improve the predictive accuracy, for ΔH_d, four of the five models (all except arXiv19) only negligibly improve upon the baseline elfrac model with MAE of ~0.12-0.14 eV/atom. While the MAE for ΔH_d is comparable to or even lower than ΔH_f, the range of values spanned by ΔH_d is also much smaller (**Figure S1**). Importantly, for the purposes of predicting stability, a difference of

~0.1 eV/atom can be the difference between a compound that is readily synthesizable and one that is unlikely to ever be realized.

DFT calculations benefit from a systematic cancellation of errors that leads to much smaller errors for ΔH_d than for ΔH_f, with MAE as low as ~0.04 eV/atom for a substantial fraction of decomposition reactions.[15] Unfortunately, ML models do not similarly benefit from this cancellation of errors and instead appear to learn clusters in material space that have similar ΔH_f but are generally unable to distinguish between stable and unstable compounds within a chemical space. It is notable that the arXiv19 model substantially improves upon the other models. However, there are still strong signatures of inaccurate stability predictions in this parity plot, most notably in the ~vertical line at $\Delta H_{d,pred} = 0$ and ~horizontal line at $\Delta H_{d,MP} = 0$. These two lines indicate substantial disagreement between the actual and predicted stabilities for many compounds, despite the relatively low MAE.

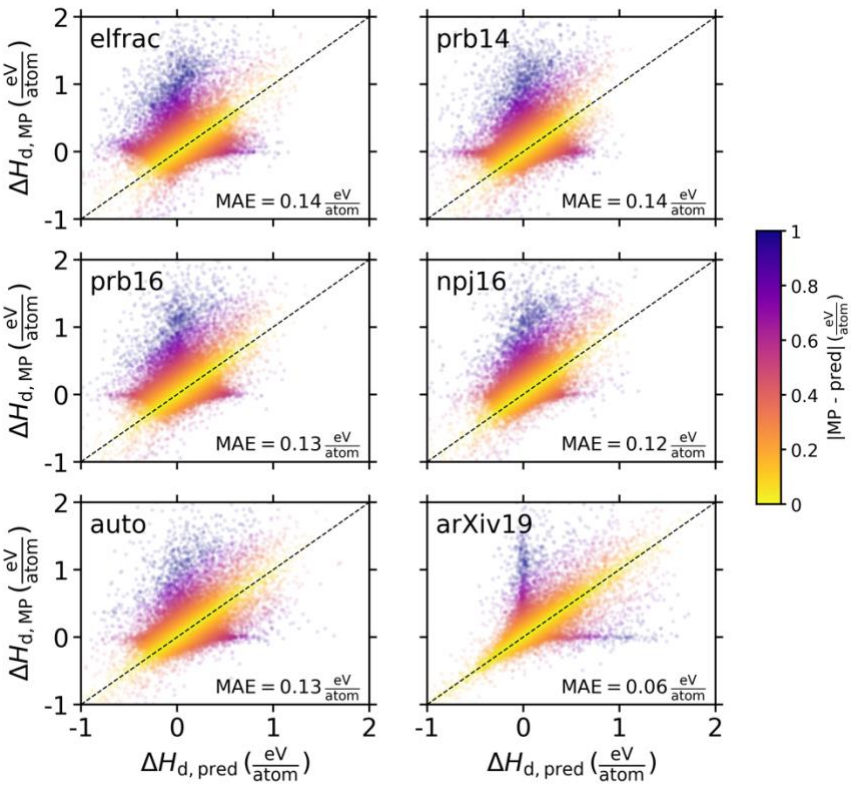

Figure 3. Parity plot for decomposition enthalpy predictions. $\Delta H_{d,MP}$ results from convex hulls constructed with $\Delta H_{f,MP}$ (**Fig. 2**). $\Delta H_{d,pred}$ is obtained from convex hulls constructed with $\Delta H_{f,pred}$ (**Fig. 2**). The annotations are the same as in **Fig. 2**.

The inability for compositional models to properly distinguish relative stability is further demonstrated by assessing how well the models classify materials as stable (on the convex hull) or unstable (above the hull), as shown in **Figure 4**. If all materials were naively predicted to be unstable, the classification accuracy would be 60%. Five of the six models (all except arXiv19) only marginally improve upon this accuracy (58-65%). Strikingly, the arXiv19 model considerably outperforms the other compositional models (76% accuracy), despite using stoichiometry alone as input. Plausibly, this superior performance is due to the arXiv19 model's use of weighted soft attention mechanisms in the manipulation of the representation.[33] While only the nominal chemical composition (element fraction) is used as input, the model learns a more meaningful representation of this inputted composition on a case-by-case basis during training. This is in contrast to the other compositional models, for which the stoichiometric representation is fixed, but hand-picked elemental attributes such as electronegativity are also included in the representation. Notably, the auto model uses a two-step process, first rationally selecting the most relevant elemental attributes from a large list using a decision tree model then fitting a regression model with the reduced feature space, and this optimized process performs approximately equivalently to the other elemental-attribute-based models. This suggests that further improvements to compositional formation energy models will likely result from qualitative changes in model architecture, as in arXiv19, and not from optimizing the selection of elemental attributes.

While the arXiv19 model improves considerably upon other compositional models, the accuracy, F_1 score, and false positive rate taken together do not inspire much confidence that any of these models can accurately predict the stability of solid-state materials. Of particular concern is the high false positive rates of 25-38%. This metric provides the likelihood that a compound predicted to be stable will not actually be. The false positive rate reported here is underestimated compared to the false positive rate that is expected for new materials discovery. The MP database is largely populated with known materials extracted from the ICSD, and this results in ~40% of ground-state entries in MP being stable. Because ~34,000 stable solid-state compounds have already been discovered (and are tabulated in MP), the fraction of all plausible hypothetical compounds that are stable should be orders of magnitude less than 40%. This necessitates that searches for new materials cover a huge number of materials, and false positive rates in excess of 25% can be computationally prohibitive to overcome, presenting a serious challenge for finding stable compounds in new or under-explored chemical spaces.

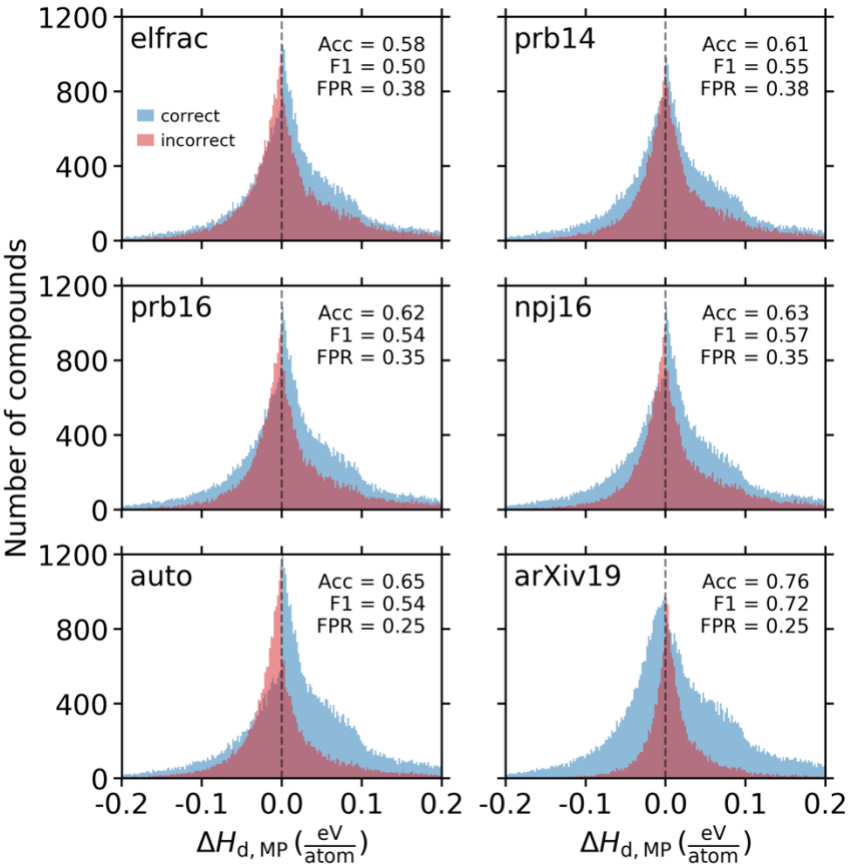

Figure 4. Classification of materials as stable ($\Delta H_d \leq 0$) or unstable ($\Delta H_d > 0$) using each of the six compositional models. The histograms are binned with respect to $\Delta H_{d,MP}$ to indicate how the correct and incorrect predictions vary as a function of the magnitude above or below the convex hull. Acc is the classification accuracy. F1 is the harmonic mean of precision and recall. FPR is the false positive rate.

Predicting stability in sparse chemical spaces

While quantifying the accuracy of ML approaches on the entire MP dataset is instructive about general errors, it does not resemble the materials discovery problem because it assesses only the limited space of compositions that have been previously explored and therefore have many stable compounds. In order to simulate a materials discovery problem, we identified a set of chemical spaces within the MP dataset that are sparse in terms of stable compounds. Lithium transition metal (TM) oxides are used as the cathode material for rechargeable Li-ion batteries and have attracted substantial attention for materials discovery in recent years. In particular, Li-Mn oxides have been considered as an alternative to $LiCoO_2$ utilizing less or no cobalt: e.g., spinel $LiMn_2O_4$,[34] layered $LiMnO_2$,[35] nickel-manganese-cobalt (NMC) cathodes,[36] and disordered rock salt cathodes.[37] For this work, the quaternary space, Li-Mn-TM-O with TM ∈{Ti, V, Cr, Fe, Co,

Ni, Cu}, is an attractive space to test the efficacy of these models, as it contains only 9 stable compounds and 258 more that are unstable in MP. We tested the potential for ML models to discover these stable compounds by excluding all 267 quaternary Li-Mn-TM-O compounds from the MP dataset and repeating the training of each model on ΔH_f with the remaining 84,747 compounds. We then applied each trained model to predict ΔH_f for the excluded Li-Mn-TM-O compounds and assessed their stability.

All the models have a higher accuracy predicting ΔH_f for this subset of materials (**Figure S2**) than for the entire dataset (**Figure 3**). The improved prediction of ΔH_f is likely because the compounds in this subset have strongly negative ΔH_f and are well-represented by the thousands of transition metal and lithium-containing oxides that comprise the MP dataset. Despite this improved accuracy on ΔH_f, the models all have alarmingly poor performance in predicting ΔH_d. In **Figure 5**, we show that none of the models are able to correctly detect more than three of the nine stable compounds, and even for the most successful model by this metric (auto), the three true positives come with 24 false positives. It is noteworthy that in this experiment, the models are given a large head-start towards making these predictions because the composition space under investigation is restricted to those compounds that have DFT calculations tabulated in MP, which is biased towards stability.

To account for the MP stability bias and more closely simulate a real materials discovery problem, we assessed the potential for these models to identify the stable MP compounds when considering a much larger composition space. Using the approach defined in Ref. 13, we produced 13,392 additional quaternary compounds in these seven Li-Mn-TM-O chemical spaces that obey simple electronegativity- and valence-based rules. For this expanded space of quaternary compounds, we used each compositional model (trained on all of MP minus the 267 Li-Mn-TM-O compounds) to predict ΔH_f and assessed their stability (**Table 1**). The compositional models each predict ~4-5% of these compounds to be stable, and all of the models fail to accurately predict the stability of more than one of the nine stable compounds in MP. 167 compounds are predicted to be stable by all four models and 1,325 unique compounds are predicted to be stable by at least one model. While it is likely that the space of stable quaternary compounds in the Li-Mn-TM-O space has not yet been fully explored in MP (or by extension, the ICSD), it is highly unlikely that the number of new stable materials in this well-studied space is orders of magnitude larger than the number of known stable materials. The false positive rates on the entire MP dataset shown in

Figure 4 suggest ~25-38% of these predicted stable compounds are not actually stable, and this rate is likely underestimated, as discussed previously. The magnitude of compounds predicted to be stable and the reported false positive rates pose a serious challenge to materials discovery because confirming the predicted stability requires the application of crystal structure prediction algorithms, which rely on many DFT calculations per compound of interest.[38]

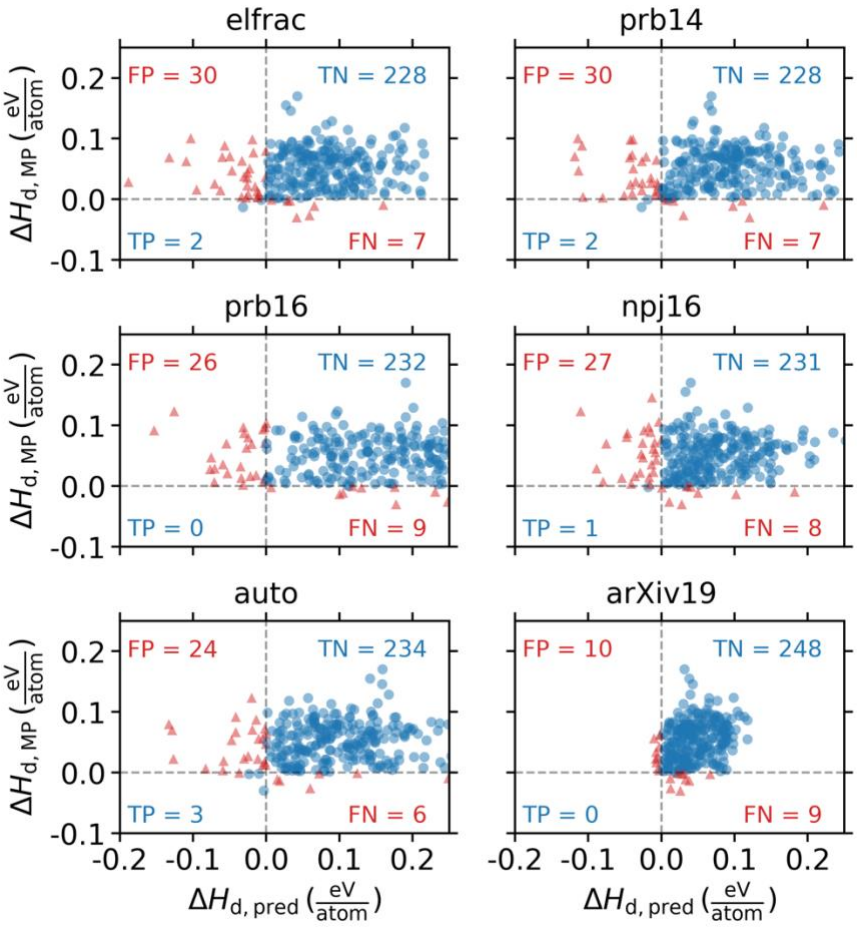

Figure 5. Re-training each model on all of MP minus 267 quaternary compounds in the Li-Mn-TM-O chemical space (TM ∈ {Ti, V, Cr, Fe, Co, Ni, Cu}) and obtaining ΔH_d using the predicted ΔH_f for each of the excluded compounds ($\Delta H_{d,pred}$) and comparing to stabilities available in MP, $\Delta H_{d,MP}$. FP = false positive, TP = true positive, TN = true negative, FN = false negative.

Table 1. Predictions in the expanded Li-Mn-TM-O (TM ∈ {Ti, V, Cr, Fe, Co, Ni, Cu}) composition space. Candidate compounds were generated by combining all quaternary MP compounds in this space along with quaternary compounds generated by the approach described in Ref. 13, resulting in 13,392 candidates. Among these candidates, 9 compounds are calculated to be stable in MP. The stability of all candidates was assessed using each compositional model for ΔH_f.

	elfrac	**prb14**	**prb16**	**npj16**	**auto**	**arXiv19**
candidate compounds	13,659	13,659	13,659	13,659	13,659	13,659
stable compounds in MP	9	9	9	9	9	9
compounds predicted stable	685	528	562	619	541	507
% predicted stable	5.0	3.9	4.1	4.5	4.0	3.7
pred. stable and stable in MP	1	1	1	1	1	1

Direct training on decomposition energy

An alternative approach to consider is training directly on ΔH_d instead of using ML-predicted ΔH_f to obtain ΔH_d through the convex hull construction. We repeated the analysis shown in **Figures 3-5** and **Table 1** but training each model directly on ΔH_d. The performance of each model on the MP Li-Mn-TM-O dataset is shown in **Figure 6**, the performance on the expanded Li-Mn-TM-O space in **Table S1**, and results for the entire MP dataset in **Figures S3-S4**. While the prediction accuracy on the entire MP dataset is typically comparable to or slightly better when training on ΔH_d instead of ΔH_f (**Figures S3-S4**), the capability of the trained model to predict stability in sparse chemical spaces is even worse than when training on ΔH_f (**Figure 6**, **Table S1**). In the case of the 267 Li-Mn-TM-O compounds in MP, none of the models are able to identify even one of the nine MP-stable quaternary compounds, and in fact every model predicts all 267 Li-Mn-TM-O compounds to be unstable (**Figure 6**). It is especially notable that for all models except arXiv19, the predictions for all 267 quaternary compounds fall in a very small window ($0.040 \text{ eV/atom} < \Delta H_{d,pred} < 0.082 \text{ eV/atom}$), suggesting the models only learn that all compounds in this space should be within the vicinity of the convex hull and do nothing to distinguish between chemically similar compounds. When the space of potential compounds is expanded to 13,659 compounds, only the arXiv19 model predicts any compound to be stable, but again, none of the nine MP stable compounds appear among this set (**Table S1**). Beyond the poor performance associated with these models, the direct prediction of ΔH_d is difficult to physically motivate because unlike ΔH_f, ΔH_d is not an intrinsic property of a material, and instead depends on the completeness of a given phase diagram. As new materials are discovered in a chemical space, ΔH_d

is subject to change for all compounds in that space, complicating the application of ML models trained on ΔH_d.

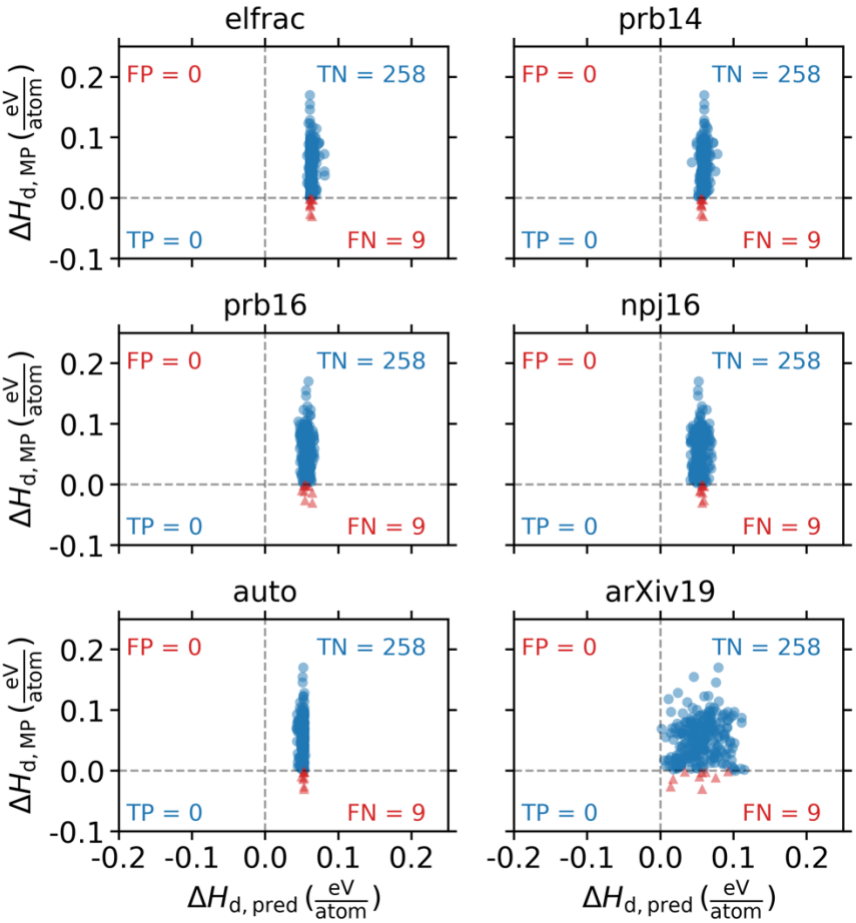

Figure 6. Repeating the analysis in **Figure 5** but training directly on ΔH_d instead of ΔH_f. All annotations are the same as in **Figure 5**.

Revisiting stability predictions with structural representations

In addition to compositional models, representations that rely on the crystal structure for predicting formation energy have also received substantial attention in recent years.[24–27,39–41] These models necessarily perform a different task than compositional models because they evaluate the property of a material given both the composition and the structure. Nevertheless, it is interesting to assess whether these structural models can predict stability with improved accuracy relative to models that are only given composition.

Here we take the crystal graph convolutional neural network (CGCNN)[26] as a representative example of existing structural models. CGCNN is a flexible framework that uses message passing over the atoms and bonds of a crystal. In **Figure 7**, we show the performance of

CGCNN on the same set of analyses as were shown for the compositional models in **Figures 2-5**. It is clear that CGCNN improves substantially upon the direct prediction of ΔH_f (**Figure 7a**) and the implicit prediction of ΔH_d (**Figure 7b**), reducing the MAE by ~50% compared with the best performing compositional model (arXiv19). The extent of extremely inaccurate predictions of ΔH_d is also no longer present with CGCNN, as indicated by the lack of a vertical line at $\Delta H_{d,pred} = 0$ or a horizontal line at $\Delta H_{d,MP} = 0$ (**Figure 7b**) and the narrow distribution of incorrect stability predictions, centered around $\Delta H_{d,MP} = 0$ (**Figure 7c**). Most impressively, CGCNN is relatively successful at finding the needles in the excluded Li-Mn-TM-O haystack, recovering five of the nine stable compounds with only six false positives (**Figure 7d**). In addition to the improved predictive accuracy, the parity plot for this excluded set looks fundamentally different than for the compositional models. In the compositional models (**Figure 4**), the parity plot is scattered, and there is effectively no linear correlation between the actual and predicted ΔH_d, whereas for CGCNN, there is a strong linear correlation, and the parity plot looks almost as though ΔH_d was predicted directly using the model.

The non-incremental improvement in stability predictions that arises from including structure in the representation is a strong endorsement for structural models and also sheds insight into the structural origins of material stability. While the thermodynamic driving force for forming a compound from its elements (formation energy) can be learned with high accuracy from only the composition, the structure dictates the subtle differences in thermodynamic driving force between chemically similar compounds and enables accurate machine learning predictions of material stability (decomposition energy). However, the glaring limitation of this approach is that it requires the structure as input, and the structure of new materials that are yet to be discovered is not known *a priori*. For example, because we do not know the ground-state structure for an arbitrary composition, we cannot repeat the test where we assess the ability of the ML model to find the stable Li-Mn-TM-O compounds among a large set of candidate compositions. While CGCNN shows substantially improved performance in predicting material stability, these results are obtained using the DFT-optimized ground-state crystal structures as input. The performance of these models on stability predictions for unexplored compounds has not yet been demonstrated.

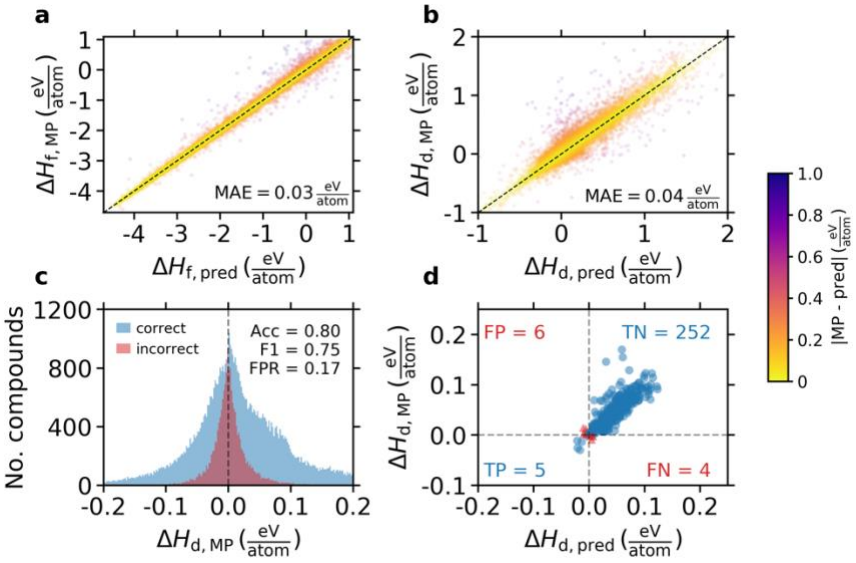

Figure 7. a) Repeating the analysis shown in **Figure 2** using CGCNN. Annotations are the same as in **Figure 2**. **b)** Repeating the analysis shown in **Figure 3** using CGCNN. Annotations are the same as in **Figure 3**. **c)** Repeating the analysis shown in **Figure 4** using CGCNN. Annotations are the same as in **Figure 4**. **d)** Repeating the analysis shown in **Figure 5** using CGCNN. Annotations are the same as in **Figure 5**.

Outlook

There have been a number of recent successes in the application of machine learning for materials science problems. However, the critical question of solid-state material stability had not been rigorously assessed. In this work, we show that while existing ML models can predict ΔH_f with relatively high accuracy from the chemical formula, they are insufficient to accurately distinguish stable from unstable compounds within an arbitrary chemical space. The error in predicting DFT-calculated ΔH_f by ML models is often compared favorably to the error DFT makes in predicting ΔH_f relative to experimentally obtained values. This comparison neglects the fact that the DFT-calculated ΔH_f differs systematically with experiment, whereas ML predictions do not. For DFT calculations, this leads to substantially lower errors for stability predictions (ΔH_d) than for ΔH_f. A similar fortuitous cancellation of errors does not occur for ML models and the errors in ΔH_d are comparable to ΔH_f, inhibiting accurate predictions of material stability. As new ML models for formation energy are developed, it is imperative to assess their viability as inputs for stability predictions and, most critically, for problems that resemble how the models would be implemented to address emerging materials science problems. In this work, we present a set of

tests that facilitate this assessment and allow for direct comparison to existing ML models. All data and code required to repeat this set of stability analyses is available at **[public github link]**.

Methods

Materials Project data

All entries in the Materials Project[9] database were queried on July 26, 2019 using the Materials Project API.[42] This produced 85,014 unique non-elemental chemical formulas. For each chemical formula, we obtained the formation energy per atom, ΔH_f, for all structure having that formula, and used the most negative (ground-state) ΔH_f for training the models and obtaining ΔH_d by the convex hull construction.

General training approach

5-fold cross validation was used to produce the model-predicted $\Delta H_{f,pred}$ shown in **Figure 2**. Each predicted value corresponds with the prediction made on that compound when it was in the validation set (i.e., not used for training). These $\Delta H_{f,pred}$ were used directly in the convex hull analysis to generate $\Delta H_{d,pred}$ shown in **Figure 3** and stability classifications in **Figure 4**. For the Li-Mn-TM-O examples (**Figure 5** and **Table 1**), each model was trained on all Materials Project entries except those 267 quaternary compounds belonging to the Li-Mn-TM-O chemical spaces. An analogous approach was used when training on ΔH_d instead of ΔH_f to generate the results shown in **Figure 6**, **Figure S3-S4**, and **Table S1**.

Compositional model training

Four of the compositional models—elfrac, prb14,[19] prb16,[20] and npj16[21]—were implemented using matminer[43] and trained using gradient boosting as implemented in XGBoost[44] with 200 trees and a maximum depth of 5. Preliminary tests showed XGBoost and these hyperparameters lead to the highest accuracy of tested algorithms. auto[22] was used as implemented in Ref. [22]. arXiv19[23] was trained for 500 epochs using an Adam optimizer with an initial learning rate of 5×10^{-4} and an L_1 loss function.

CGCNN training

We used a modified 5-fold cross-validation to train the CGCNN[26] model for the MP ΔH_f dataset. As a general procedure for cross-validation, the dataset is split to 5 groups and each group is iteratively taken as a hold-out test set. For each fold, we then split the training set to 75% training and 25% validation, thus the overall size ratio of training, validation, and test is 60%, 20%, and 20%, respectively. The CGCNN model is iteratively updated by minimizing the loss (mean squared error, MSE) on the training set, and the validation score (mean absolute error, MAE) is monitored after each epoch. After 1000 epochs, the model with the best validation score is selected and then evaluated on the hold-out test set. Results of the 5-fold hold-out test sets are accumulated as the final predictions of the dataset.

For the Li-Mn-TM-O case in which the test set is defined, we split the remaining compounds into 5 groups and iteratively took each group as the validation set (20%) and the remaining as the training set (80%). The best CGCNN model of each fold was selected as the one with the best validation score (MAE). We then applied the 5 CGCNN models to the 267 Li-Mn-TM-O test compounds and used the average of the predicted ΔH_f for each model.

Acknowledgments

[savio], [lawrencium], [eagle], [funding]

Author contributions

CJB, AT, and GC conceived the project. CJB, AT, QW, and AD designed the project. AT, QW, and AD implemented the machine learning models. CJB performed the stability analysis, processed the results, and drafted the manuscript. AJ and GC supervised the project. All authors reviewed and edited the manuscript.

Data availability

All results are available at **[public github link]** along with Python code for repeating the stability analysis for newly developed models and reproducing all figures shown in the main text and **Supporting Information**.

Supporting Information

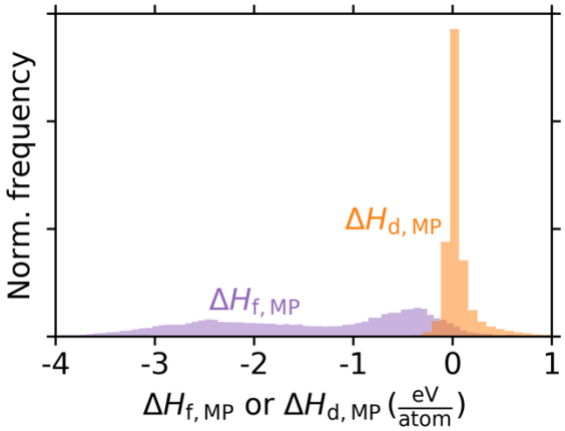

Figure S1. Distribution of ΔH_f and ΔH_d in Materials Project, with 85 bins and normalized such that the integral of the distribution is equal to 1.

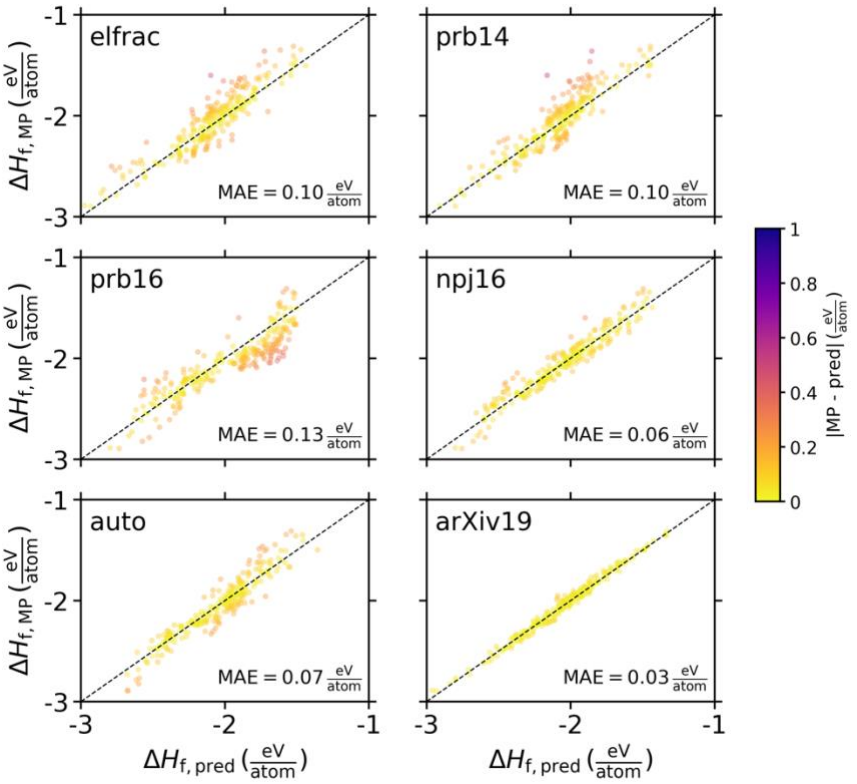

Figure S2. Re-training each model on all of MP minus 267 quaternary compounds in the Li-Mn-TM-O chemical space (TM ∈ {Ti, V, Cr, Fe, Co, Ni, Cu}) and predicting ΔH_f for each of the excluded compounds ($\Delta H_{f,pred}$) and comparing to MP, $\Delta H_{f,MP}$. All annotations are the same as in **Figure 2**.

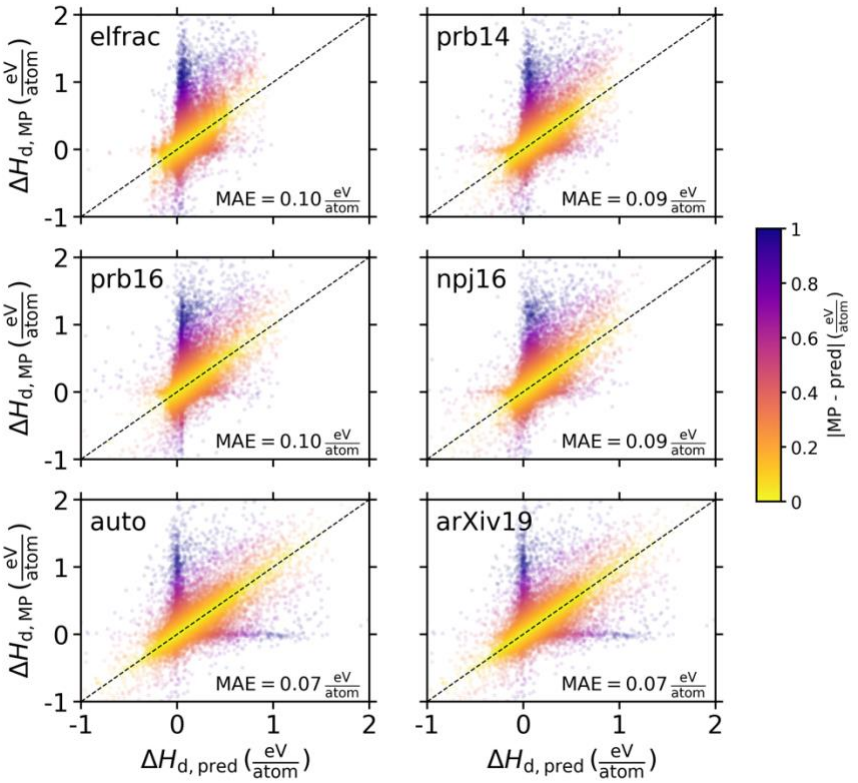

Figure S3. Reproducing **Figure 3** but training on ΔH_d instead of ΔH_f. All annotations are the same as in **Figure 3**. *** NOTE: "auto" data is not yet generated – placeholder with arXiv19 data ***

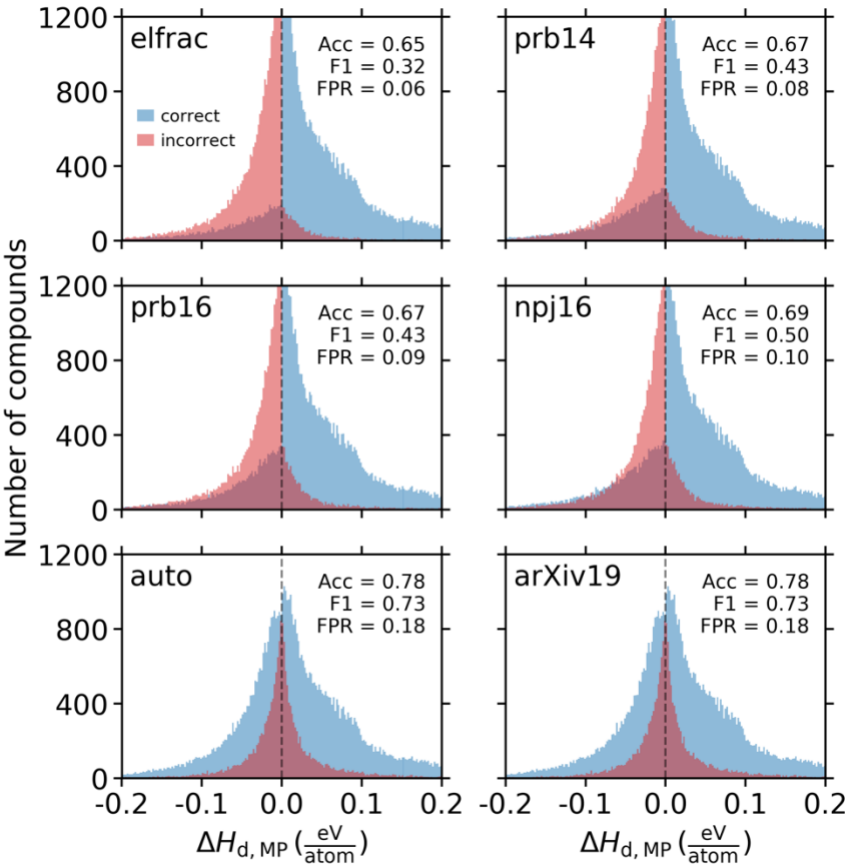

Figure S4. Reproducing **Figure 4**, but training on ΔH_d instead of ΔH_f. All annotations are the same as in **Figure 4**. *** NOTE: "auto" data is not yet generated – placeholder with arXiv19 data ***

Table S1. Reproducing **Table 1** but training on ΔH_d instead of ΔH_f

	elfrac	**prb14**	**prb16**	**npj16**	**auto**	**arXiv19**
candidate compounds	13,659	13,659	13,659	13,659	13,659	13,659
stable compounds in MP	9	9	9	9	9	9
compounds predicted stable	0	0	0	0	0	299
% predicted stable	0	0	0	0	0	2.2
pred. stable and stable in MP	0	0	0	0	0	0

References

1. Himanen, L., Geurts, A., Foster, A. S. & Rinke, P. Data-Driven Materials Science: Status, Challenges, and Perspectives. *Advanced Science* **0**, 1900808 (2019).
2. Schleder, G. R., Padilha, A. C. M., Acosta, C. M., Costa, M. & Fazzio, A. From DFT to machine learning: recent approaches to materials science–a review. *Journal of Physics: Materials* **2**, 032001 (2019).
3. Goldsmith, B. R., Esterhuizen, J., Liu, J.-X., Bartel, C. J. & Sutton, C. A. Machine learning for heterogeneous catalyst design and discovery. *AIChE-Journal* **64**, 2311–2323 (2018).
4. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
5. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials* **3**, 54 (2017).
6. Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials* **5**, 83 (2019).
7. Pizzi, G., Cepellotti, A., Sabatini, R., Marzari, N. & Kozinsky, B. AiiDA: automated interactive infrastructure and database for computational science. *Computational Materials Science* **111**, 218–230 (2016).
8. Kirklin, S. *et al.* The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials* **1**, 15010 (2015).
9. Jain, A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **1**, 011002 (2013).
10. Draxl, C. & Scheffler, M. NOMAD: The FAIR concept for big data-driven materials science. *MRS Bulletin* **43**, 676–682 (2018).
11. Curtarolo, S. *et al.* AFLOW: An automatic framework for high-throughput materials discovery. *Computational Materials Science* **58**, 218–226 (2012).
12. Hellenbrandt, M. The Inorganic Crystal Structure Database (ICSD)—Present and Future. *Crystallography Reviews* **10**, 17–22 (2004).
13. Davies, D. W. *et al.* Computational Screening of All Stoichiometric Inorganic Materials. *Chem* **1**, 617–627 (2016).

14. Bartel, C. J. *et al.* Physical descriptor for the Gibbs energy of inorganic crystalline solids and temperature-dependent materials chemistry. *Nature communications* **9**, 4168 (2018).
15. Bartel, C. J., Weimer, A. W., Lany, S., Musgrave, C. B. & Holder, A. M. The role of decomposition reactions in assessing first-principles predictions of solid stability. *npj Computational Materials* **5**, 4 (2019).
16. Hautier, G., Ong, S. P., Jain, A., Moore, C. J. & Ceder, G. Accuracy of density functional theory in predicting formation energies of ternary oxides from binary oxides and its implication on phase stability. *Phys. Rev. B* **85**, 155208 (2012).
17. Ong, S. P., Wang, L., Kang, B. & Ceder, G. Li−Fe−P−O2 Phase Diagram from First Principles Calculations. *Chem. Mater.* **20**, 1798–1807 (2008).
18. Zunger, A. Inverse design in search of materials with target functionalities. *Nature Reviews Chemistry* **2**, 0121 (2018).
19. Meredig, B. *et al.* Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **89**, 094104 (2014).
20. Deml, A. M., O'Hayre, R., Wolverton, C. & Stevanović, V. Predicting density functional theory total energies and enthalpies of formation of metal-nonmetal compounds by linear regression. *Phys. Rev. B* **93**, 085142 (2016).
21. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials* **2**, 16028 (2016).
22. AutoMatminer. *https://github.com/hackingmaterials/automatminer*.
23. Goodall, Rhys & Lee, Alpha. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *arXiv pre-print 1910.00617*.
24. Faber, F. A., Lindmaa, A., von Lilienfeld, O. A. & Armiento, R. Machine Learning Energies of 2 Million Elpasolite ABC2D6 Crystals. *Physical Review Letters* **117**, 135502 (9).
25. Ward, L. *et al.* Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Phys. Rev. B* **96**, 024104 (2017).
26. Xie, T. & Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters* **120**, 145301 (June 4).
27. Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* **31**, 3564–3572 (2019).

28. Stevanović, V., Lany, S., Zhang, X. & Zunger, A. Correcting density functional theory for accurate predictions of compound enthalpies of formation: Fitted elemental-phase reference energies. *Phys. Rev. B* **85**, 115104 (2012).
29. Pandey, M. & Jacobsen, K. W. Heats of formation of solids with error estimation: The mBEEF functional with and without fitted reference energies. *Phys. Rev. B* **91**, 235201 (2015).
30. Zhang, Y. *et al.* Efficient first-principles prediction of solid stability: Towards chemical accuracy. *npj Computational Materials* **4**, 9 (2018).
31. Miedema, A. R. Simple model for alloys. *Philips Tech. Rev., v. 33, no. 6, pp. 149-160* (1973).
32. Isaacs, E. B. & Wolverton, C. Performance of the strongly constrained and appropriately normed density functional for solid-state materials. *Phys. Rev. Materials* **2**, 063801 (2018).
33. Xu, K. *et al.* Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *preprint at https://arxiv.org/abs/1502.03044*.
34. Thackeray, M. M., David, W. I. F., Bruce, P. G. & Goodenough, J. B. Lithium insertion into manganese spinels. *Materials Research Bulletin* **18**, 461–472 (1983).
35. Armstrong, A. R. & Bruce, P. G. Synthesis of layered LiMnO2 as an electrode for rechargeable lithium batteries. *Nature* **381**, 499–500 (1996).
36. Thackeray, M. M. *et al.* Li2MnO3-stabilized LiMO2 (M = Mn, Ni, Co) electrodes for lithium-ion batteries. *J. Mater. Chem.* **17**, 3112–3125 (2007).
37. Lee, J. *et al.* Reversible Mn2+/Mn4+ double redox in lithium-excess cathode materials. *Nature* **556**, 185–190 (2018).
38. Oganov, A. R., Pickard, C. J., Zhu, Q. & Needs, R. J. Structure prediction drives materials discovery. *Nature Reviews Materials* **4**, 331–348 (2019).
39. Noh, J. *et al.* Inverse Design of Solid-State Materials via a Continuous Representation. *Matter* doi:10.1016/j.matt.2019.08.017.
40. Isayev, O. *et al.* Universal fragment descriptors for predicting properties of inorganic crystals. *Nature Communications* **8**, 15679 (2017).
41. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).

42. Ong, S. P. *et al.* The Materials Application Programming Interface (API): A simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST) principles. *Computational Materials Science* **97**, 209–215 (2015).
43. Ward, L. *et al.* Matminer: An open source toolkit for materials data mining. *Computational Materials Science* **152**, 60–69 (2018).
44. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* 785–794 (2016) doi:10.1145/2939672.2939785.

Suggested reviewers

Keith Butler

Jeff Grossman

Alpha Lee

Bryce Meredig

Chris Wolverton