A Combined DFT/Machine Learning Framework for Materials Discovery: Application to Spinels and Assessment of Search Completeness and Efficiency

Joshua A. Schiller and Elif Ertekin*

Department of Mechanical Science & Engineering, 1206 W Green Street, University of Illinois at Urbana-Champaign, Urbana IL 61801

E-mail: ertekin@illinois.edu

Abstract

It is challenging to evaluate machine learning approaches developed for accelerating materials search and discovery in a realistic way. Machine learning approaches to materials stability prediction are typically assessed by their ability to reproduce results from direct physical modeling, whereas ideally both machine learning and direct physical modeling should be assessed by their ability to reproduce reality. Additionally, traditional evaluation metrics do not directly reflect the experience of an experimental search for unknown compounds in a large candidate phase space, and often result in overly optimistic assessments. Here, we (i) present a framework that combines density functional theory and traditional supervised machine learning methods (ML/DFT), and (ii) introduce the concepts of *search completeness* – the fraction of discoverable compounds found relative to the fraction of search space explored – and *search efficiency* – the rate of discovery relative to the fraction of search space explored –

to evaluate it. The ML/DFT framework is an iterative approach to predict stable chemistries of a fixed crystal structure (here, spinels) that uses DFT to generate a training set of unstable compounds. The training set of stable compounds is given by experimentally known spinels. The method is carried out using random forest, LASSO, and ridge regression to predict as-of-yet undiscovered spinel chemistries. TreeSHAP analysis is used to determine features that most contribute to stability/instability classification. While no single feature dominates, several emerge that align with chemical intuition. To estimate the efficacy of ML/DFT compared to pure DFT, we introduce a Bayesian description of DFT distribution of energies for stable and unstable spinels. The Bayesian model enables quantifying the search completeness and search efficiency of DFT, which is then compared to that of ML/DFT. ML/DFT achieves search completeness and efficiency on par with pure DFT, despite requiring fewer DFT simulations (∼300 vs. 14,200). More importantly, by quantitatively assessing ML approaches in ways that better reflect how they would be used in materials discovery experiments, we obtain key insights into the challenges that need to be overcome by such methods: that the small number of stable compounds to be found in a search space orders of magnitude larger places stringent demands on model accuracy to achieve good search efficiency. Finally, we report the top candidates of our spinel search, which may be of interest for synthesis experiments.

Introduction

Prediction and discovery of new materials using computation remains a longstanding challenge.[1–7] The challenge arises in part from the vast compositional and structural phase space in which materials live.[1,8,9] The large phase space, combined with the complex way the materials energy landscape varies with chemistry and crystal structure,[8] makes discovering a new stable compound akin to finding a needle in a haystack. To date, there is no robust and scalable method that can achieve the needed accuracy and precision for efficient materials

discovery.[1]

For a material to be stable, it should be the lowest energy compound in the chemical phase space of all its competing compounds. Assessing stability requires knowing the decomposition reaction energy to all other possible compounds, which in turn requires distinguishing relative formation energies. Decomposition reaction energies are often shown on a plot of formation energy *vs.* composition, as in Figure 1(a) which shows a hypothetical AB binary composition space. On this plot, the envelope of compounds with lowest formation energy forms the convex hull. All compounds that appear on the hull are thermodynamically stable. As Figure 1(a) shows, formation energies can differ from each other by small amounts. Small uncertainties in formation energies can translate to greater uncertainties in decomposition reaction energies, and ultimately in the determination of stability. To illustrate, in Figure 1(b) a random shift in formation energy has been drawn from a gaussian distribution centered at zero with width 0.2 eV/atom and applied to all compounds. For compounds estimated to be on or close to the convex hull, the uncertainty gives rise to both false negatives and false positives. By contrast, the uncertainty is less detrimental for compounds estimated to be far enough above the hull – they are likely to be unstable in spite of the uncertainty. Improving our ability to predict stability necessitates reducing uncertainties in particular for compounds that are believed to be on or in proximity to the hull.

The current workhorse for computational prediction of stability is density functional theory (DFT).[10,11] Many materials databases[12–19] that provide large datasets of DFT-computed material properties are now available. Depending on the DFT functional and material chemistry, estimates of the mean absolute error (MAE) in DFT-computed formation energies lie between 0.15 – 0.25 eV/atom.[14,20] Even when the systematic component to the error is accounted for, the residual MAE is around 0.05 eV/atom,[15] which is better but still leaves the possibility of false positives and false negatives. These limitations of DFT propagate to materials discovery methods that rely on DFT energies, such as genetic/evolutionary algorithms,[21–23] particle swarm optimization[24] and simulated annealing.[25]

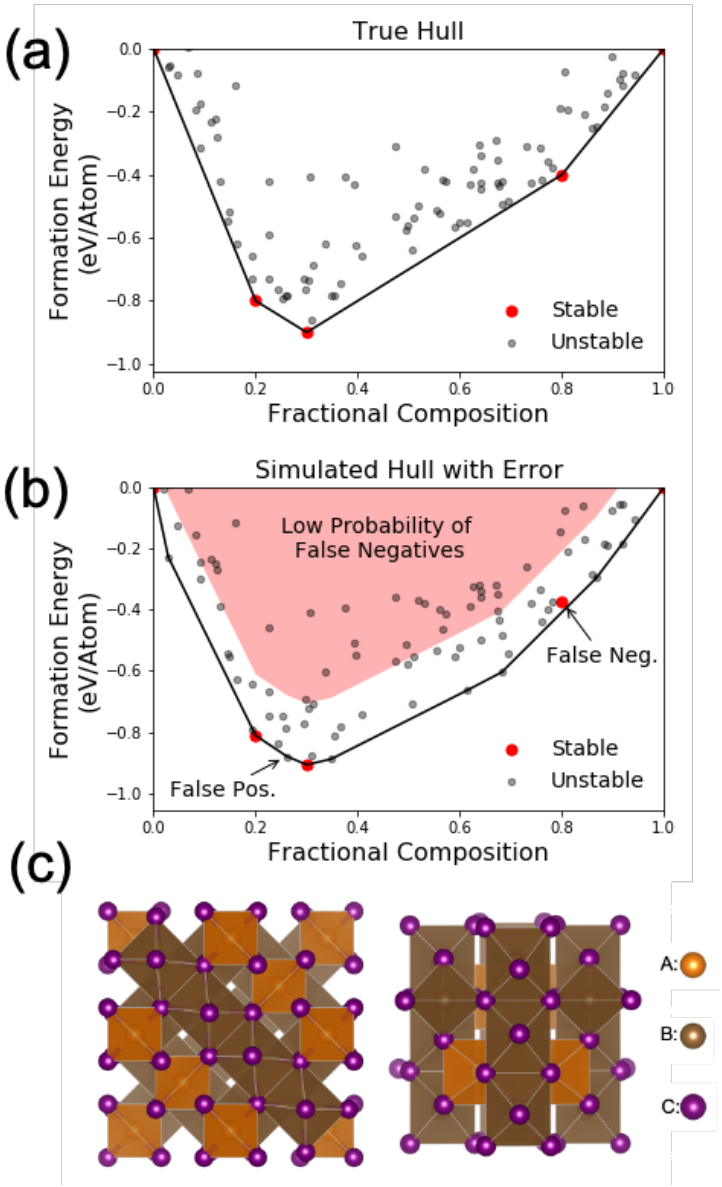

Figure 1: (a) Schematic illustration of convex hull in a hypothetical AB binary system. (b) The same hull, now with an 0.2 eV/atom uncertainty applied to formation energies. The uncertainty not only reshapes the convex hull itself, but also leads to false positives and negatives. Compounds that appear more than 0.2 eV/atom above the hull (red shading) have a low probability of producing false negatives. (c) The structure of the normal ($Fd\bar{3}m$, left) and proxy inverse (Imma, right) spinel with stoichiometry AB_2C_4. Gold, brown, and purple atoms correspond to species A, B, C respectively.

Distinct from physics-based models, machine learning and data-driven approaches have recently emerged as avenues for materials discovery.[2,26,27] One challenge is that acquiring large sets of both stable and unstable compounds for model training and evaluation can be difficult. For stable compounds the Inorganic Crystal Structure Database (ICSD)[28–31] contains $\sim 10^5$ known compounds, orders of magnitude smaller than the phase space of possible compounds. Knowledge of the space of unstable compounds, orders of magnitude larger in reality, is even more limited. Another challenge lies in evaluating the performance of ML methods. Since current methods typically seek to reproduce DFT energies, their performance is evaluated based on their ability to reproduce DFT.[2–7] Given DFT's own uncertainties, this makes it impossible to compare whether an ML approach can outperform DFT. It is currently unclear if ML errors relative to DFT are larger or smaller than the errors of DFT itself. Ideally, both ML and DFT should be assessed based on their ability to reproduce reality. Additionally, traditional metrics used to evaluate data-based approaches do not directly represent how completely or efficiently a model can discover the small number of stable compounds in a search space that is orders of magnitude larger. It is difficult to generate data sets that properly reflect the search space, and as such data sets may be biased and result in an overly optimistic assessment of model efficacy.[32,33]

A recent analysis[1] of several supervised learning approaches to materials stability[2–7] showed that, despite being able to learn DFT formation energies with reasonable accuracy, the methods struggled to reproduce DFT decomposition energies. The difference was attributed to the observation that DFT-computed energies benefit from a systematic cancellation of errors not present in ML, that helps DFT better distinguish relative formation energies. Only a crystal graph convolutional neural network[7] (and also Ref.[34]) that includes structural and compositional information could reasonably reproduce DFT stability predictions. Since structure contains information about bonding that is critical to distinguish between compounds with otherwise similar compositions, it is not surprising that models that include structure in the material representation outperform those that don't.

These findings invite two questions: what will it take for ML to perform as well as or better than pure DFT, as measured under conditions that represent an true experimental search for new materials?, and how can the determination even be made? Based on the observations above, there are several aspects to achieve accelerated materials discovery using machine learning. (i) Although challenging, it would be most beneficial to improve predictive capability for the stability of compounds that according to DFT are on or close to the convex hull. (ii) Demonstrating improvements over DFT is challenging, since DFT benefits from a cancellation of errors in formation energies not present in ML. (iii) Demonstrating improvements over DFT will likely require structural as well as chemical representations of materials. (iv) An approach to evaluating performance relative to reality, rather than relative to DFT, is needed.

To address these considerations, in this work we design a framework that combines DFT and ML for materials stability prediction (called ML/DFT), and critically assess whether it is possible for this framework to outperform pure DFT in a hypothetical materials search. We limit our search space to the prediction of only one crystal structure, spinel compounds, to obviate the need to provide structural information. Spinels serve as a good test case due to the large number of known stable compounds for model training. They are known to exhibit a range of properties, including magnetism,[35] superconductivity,[36] ion transport[37–42] and transparent conduction.[43,44] In our approach, DFT is only used where its uncertainties are least detrimental: to generate the dataset of unstable compounds. We label candidate compounds as unstable when they are found in DFT to lie > 0.2 eV/atom above the hull (as in Figure 1(b)), where the likelihood of false negatives is small.

To assess the performance of our approach in a realistic way, we introduce the concepts of *search completeness* and *search efficiency*. Search completeness measures the proportion of discoverable compounds found as a function of the fraction of search space explored. Search efficiency measures the discovery rate, also as a function of the fraction of search space explored. We develop a Bayesian model that describes the distribution of stable and

unstable spinels' distance to the convex hull in DFT. This model allows us to infer true distributions, compared to DFT, and assess the search completeness and search efficiency of a hypothetical materials search for undiscovered spinels using pure DFT and ML/DFT. We find that within the reduced structural space, the search completeness and efficiency of ML/DFT is on par with that of pure DFT, due to its ability to identify the few undiscovered stable compounds within a large data set of mostly unstable compounds. However, ML/DFT has the advantage of achieving this efficiency with substantially fewer DFT simulations. Therefore, the approach may be an effective way to prioritize synthesis experiments with a limited amount of DFT data. By highlighting the importance of quantitatively assessing ML approaches in ways that better reflect how they would be used in materials discovery, we obtain key insights into the challenges that will need to be overcome by such methods. Namely, the small number of discoverable stable compounds in a search space that is orders of magnitude larger places stringent demands on accuracy to achieve a high search efficiency.

Methods

Crystal Structure

Spinels have chemical formula AB_2C_4, and crystallize in a cubic structure, illustrated in Figure 1(c). The anions C are arranged in a cubic closed-packed arrangement and the cations A and B occupy some or all of the octahedral or tetrahedral sites. In the prototype spinels (the '2-3' spinels considered here), the formal charges of cations A and B are +2 and +3 respectively, although other valences are possible. For 2-3 normal spinels, the A^{2+} cations occupy tetrahedral sites and the B^{3+} cations occupy octahedral sites. In the 2-3 inverse spinel, all A^{2+} and half of the B^{3+} occupy octahedral sites, while the other half of the B^{3+} occupy tetrahedral sites. In this case, the octahedral sites exhibit a disordered arrangement of two elements. Intermediate structures between normal and inverse are also possible. We consider the possibility of both normal and inverse spinels. When simulating

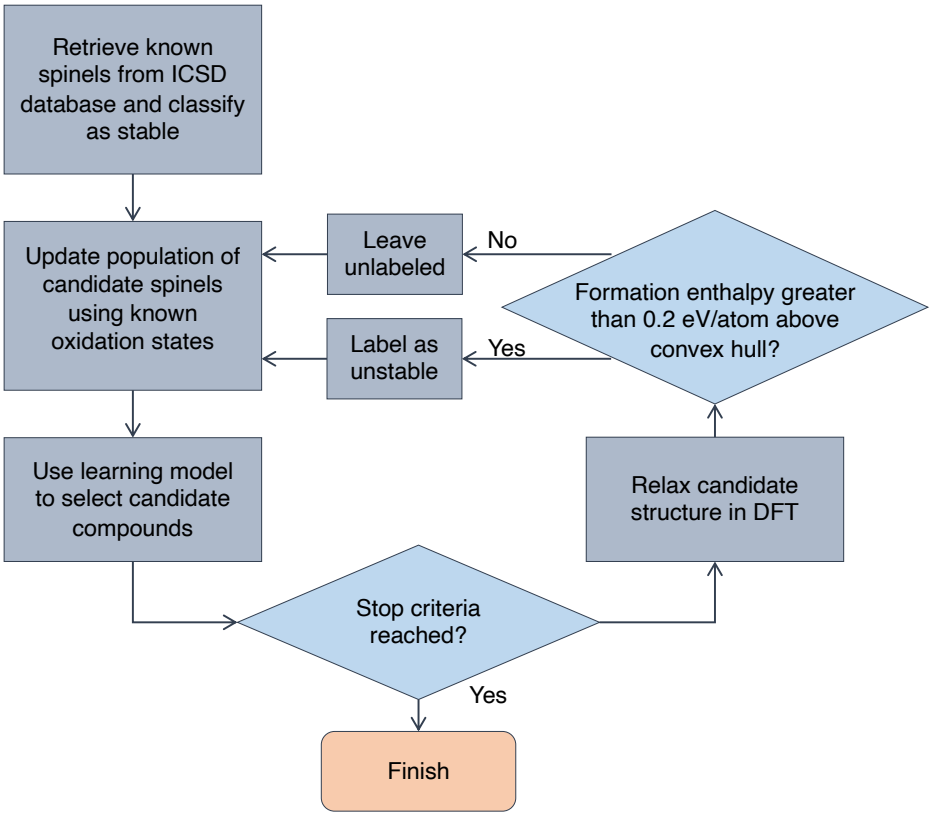

Figure 2: The sequential learning procedure for exploring the spinel phase space. Machine learning is used to predict undiscovered spinels using a training set consisting of experimentally known stable compounds and DFT-labeled unstable compounds. These candidates are relaxed in DFT and their distance from the convex hull calculated. Compounds with a distance greater than 0.2 eV/atom are labeled unstable and added to the training set. The procedure is repeated until performance metrics stabilize and the predicted top candidates become fixed.

the inverse spinel in DFT, we utilized a universal ordering of the cations that exhibits the the lowest electrostatic energy.[45]

Energetic Stability Calculations

Stability is governed by the Gibbs formation energy ΔG_f and Gibbs decomposition energy ΔG_d. When DFT is used, we compute the formation enthalpy ΔH_f and the decomposition enthalpy ΔH_d, respectively, and therefore neglect temperature/entropy effects. This does not impact the ML/DFT approach, since the stable set of compounds used for testing and training is given by the experimentally known spinels and is independent of DFT. DFT is used to generate the unstable set, but our criterion $\Delta H_d > 0.2$ eV/atom to label a compound unstable is sufficient to ensure actual instability. For instance, according to our data set only four known stable spinels out of 200 have a DFT-computed $\Delta H_d > 0.2$ eV/atom. If a candidate compound's DFT-computed distance from the hull is ≤ 0.2 eV/atom, it is left unlabeled. Here onwards, in our notation, the symbol ΔH_d always refers to DFT-computed distance to hull (in contrast to actual or experimental values).

Combined DFT/ML Iterative Approach

The combined DFT/ML framework operates cyclically and is described in Figure 2. The initial training set consisted of both stable and unstable compounds. The set of stable spinels consisted of the $\sim$ 200 known stable spinels reported in the ICSD. The initial training set also contained 40 DFT-labeled unstable compounds. These were generated by elemental substitution into the AB_2C_4 spinel normal and inverse structures. The C anions are restricted to the chalcogens {O, S, Se, Te}, and A and B cations were selected from elements that have known oxidation states of 2+ and 3+, respectively. This results in a search space of $\sim$ 14,200 candidate spinels (200 of which are the known stable spinels). Utilizing archetype normal and inverse spinel structures, we populate the A, B and C sites with the new potential elements selected at random, and use DFT to compute ΔH_d. The first forty candidates found to have

$\Delta H_d > 0.2$ eV/atom were included in the unstable training set.

Using this initial training set, three traditional supervised machine learning algorithms – LASSO, ridge regression, and random forest – were used to predict new candidate spinels, ranked according to their likelihood of stability. If any of the top twenty predicted candidates were found to exist in the literature (not all reported spinels are present in the ICSD), they were added to the dataset as stable compounds, and the step was repeated. The union of the top fifty new candidates from each learning approach were then simulated in DFT and classified according to the $\Delta H_d > 0.2$ eV/atom cutoff criterion. Any predicted unstable spinels were added to the dataset, and the procedure was repeated until both model performance and the new compounds predicted become stabilized. With each cycle, the training set therefore grows. It is updated with new unstable compounds, which are the ones that the model had assigned a high probability of stability to in the previous round. In other words, for each subsequent round the training set contains new information on compounds that the model mis-classified in the prior round. We also note that according our criterion, compounds with $0 < \Delta H_d < 0.2$ eV/atom are not used for training or testing in our ML/DFT approach, since they remain unlabeled.

Machine Learning Approach and Features

All three machine learning approaches (LASSO, ridge regression, and random forest) utilized the scikit-learn python library.[46,47]

The random forest classifier uses a set of decision trees (1000 tree estimators in our case) to determine the classification of a sample input. Each decision tree divides a training set into two groups at each node, based on one of the features and a splitting threshold. Splitting features are chosen from a random subset of sample features and splitting thresholds and optimized to produce the best class split utilizing gini impurity.

The LASSO and ridge regression classifiers are logistic regressions regularized, respectively, with the L1-norm and L2-norm of their constituent weighting parameters. Using an

appropriate regularization parameter λ, one can reduce overfitting, while maximizing model applicability. We use k-fold cross validation, and one of the folds is held out as a validation set and the rest are optimized for a set of penalty constants λ. The error is then determined by predicting classifications for each sample in the validation set and comparing the model prediction to the true classification. This is repeated with a different fold withheld as the validation set, for all folds, and the accuracy of the model is evaluated by averaging over all folds. The optimal value of λ is the one that yields the highest accuracy.

We used 10-fold cross validation for all regression models, and fit and tested each one hundred times utilizing a 90/10 train/test split to determine statistical metrics for each. To account for the asymmetry in the number of unstable and stable labels, in each iteration we oversampled the dataset so that there were equal amounts of stable and unstable labels. The training set was mean–centered and variance–normalized. After cross-validation, the final model was refit with all of the training data for making predictions. An example of predictions for part of a test set for the random forest model is shown in Table 1.

Table 1: Example predictions on a test set from the random forest model.

	Predicted Class	True Class
FeCr2O4	Stable	Stable
RhCo2O4	Stable	Stable
SnAl2Te4	Unstable	Unstable
MgHo2Se4	Unstable	Stable
TiV2Te4	Unstable	Unstable
ClTc2Te4	Unstable	Unstable
CuCr2O4	Stable	Stable
MgEr2Se4	Unstable	Stable
MgTm2Se4	Unstable	Stable
MnCo2O4	Stable	Stable

The features used were the experimental atomic mass, atomic radius, Pauling electronegativity, elemental row and elemental group for each species present. Products and quotients of the atomic mass, atomic radius and Pauling electronegativity for all combinations of the constituent atoms were added to account for interactions between attributes. Feature engi-

neering of existing attributes in this way has been previously used when applying machine learning to materials.[48,49]

Density Functional Theory Simulation Parameters

DFT calculations were carried out using VASP.[50–53] We used the PBE+U approximation,[54] with U applied to transition metal elements, and PAW pseudopotentials.[55,56] The parameter U was chosen to match settings used in Materials Project[12] and the resulting energies were adjusted in accordance with Material Project's settings[12,57–59] to make total energies comparable. An energy cutoff of 500 eV and a 4x4x4 Monkhorst-Pack mesh was used for all calculations, with ferromagnetic starting spin configurations. We did not check antiferromagnetic configurations. Typically antiferromagnetic orderings exhibit energy differences of around 0.04–0.08 eV/atom relative to the ferromagnetic configuration.[60,61] Since we only use DFT to classify compounds as unstable, neglecting antiferromagnetic configurations has the effect of only slightly blurring the threshold cutoff of $\Delta H_d > 0.2$ eV cutoff, and is not expected to substantially alter our results. All candidate spinels were fully relaxed (lattice constants and internal degrees of freedom) for both the normal and the proxy inverse configuration.

Results and Discussion

Model Performance – Traditional Metrics

Traditional metrics – accuracy, precision, and recall – were recorded for each cycle. These metrics are defined as

$$\text{Accuracy} = \frac{TP+TN}{P+N} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3}$$

where TP is the number of correctly labeled stable compounds, TN is the number of correctly-labeled unstable compounds, and P and N are respectively the total number of stable and unstable compounds in the entire population. The accuracy is the overall fraction of spinels correctly classified, precision is the likelihood that a compound predicted to be stable is in fact stable, and recall is the fraction of truly stable spinels predicted to be stable.

The results for each metric over seven cycles are plotted in Figure 3(a-c), with final values displayed in Table 2. The random forest classifier outperforms both LASSO and ridge regression across all model metrics. Additionally, all three metrics drop somewhat as the number of cycles increases. This is likely due to the increasing complexity of the data set, since with each cycle a larger number of unstable compounds are are present. For instance, initially the training/testing set contains only 40 unstable compounds, but by the end of the seven rounds there are $\sim$160 compounds labeled unstable using our cutoff criterion.

Figure 3(d) shows the precision–recall curves for all three classifiers at the end of the cycles. A precision–recall curve is an effective way to measure success of prediction when the classes are imbalanced. It highlights the tradeoff between high precision (low false positive rate, i.e. that materials predicted to be stable are in fact stable) and high recall (low false negative rate, i.e. ability to identify all discoverable compounds) as a threshold (likelihood of stability) is varied. It illustrates how well-separated the stable and unstable classes are with respect to a predicted score, here the likliehood of stability. For RF/DFT the plots are made by adjusting the threshold score required to classify an outcome as stable and plotting the resulting precisions and recalls. A high area under the curve represents both high recall and high precision, indicating confidence that both a compound predicted to be stable is actually stable and that actually stable compounds are not missed. All three classifiers show good precision–recall here, but we will later show how this changes when searching a more realistic space.

Accuracy is often used as a primary means of quantifying model performance, but it may not give a good representation of a model when working with asymmetric test sets

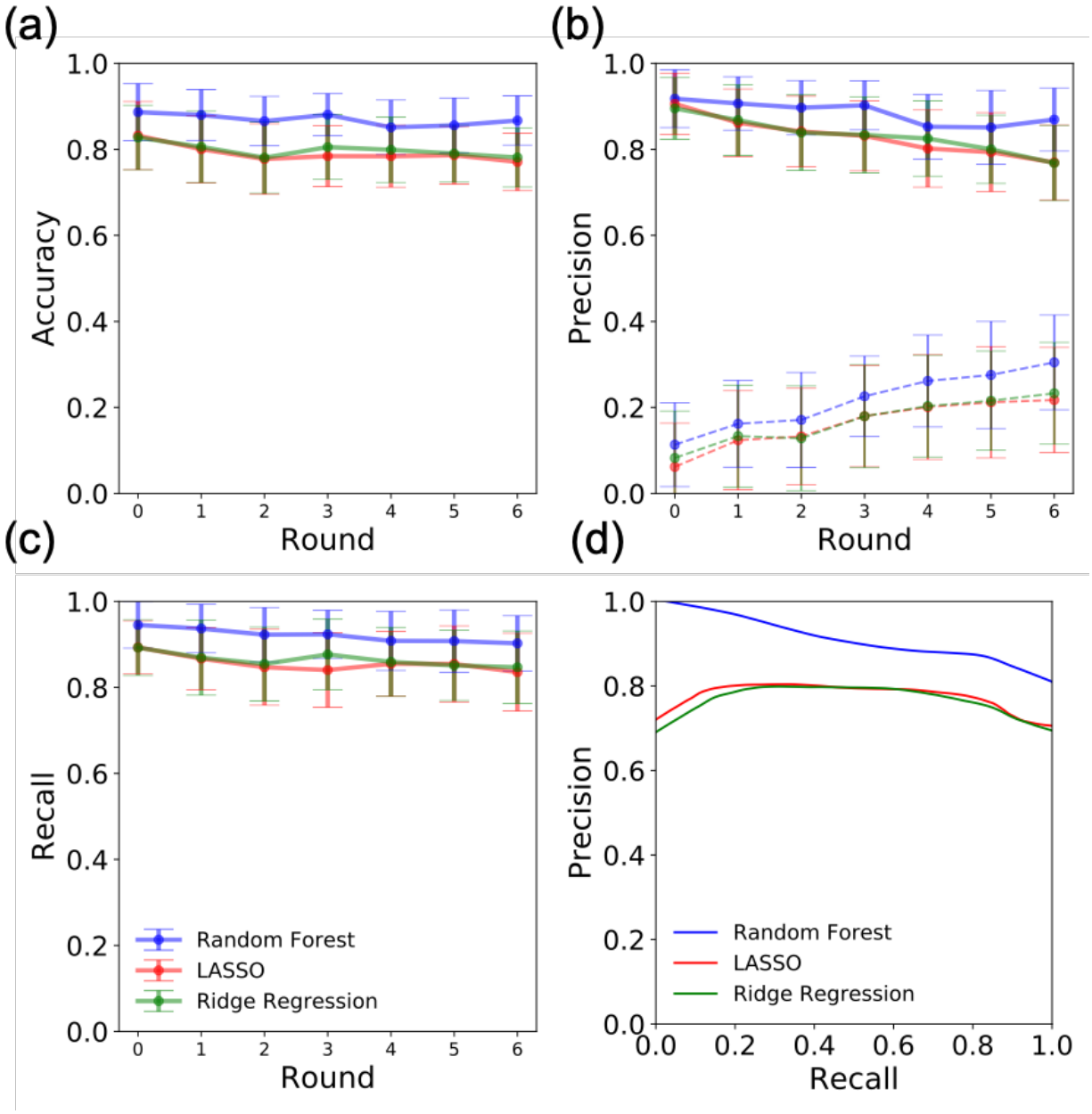

Figure 3: (a) Accuracy, (b) precision, (c) recall and (d) precision-recall for random forest (blue), LASSO (red) and ridge regression (green) across seven rounds of the sequential learning procedure. The solid lines refer to the scores for each method. In (b), the dashed lines indicate the improvement of the precision over that of a random classifier. This shows that although the precision drops slightly over the cycles, the improvement grows. (d) Precision–recall curve.

Table 2: Results for the final round of the sequential learning process for accuracy, recall, precision and improvement over a random classifier for precision as defined in Equations (1–3). The random forest classifier outperforms both LASSO and ridge regression.

	Random Forest	LASSO	Ridge Regression
Accuracy	0.87 ± 0.06	0.77 ± 0.07	0.78 ± 0.07
Recall	0.90 ± 0.06	0.84 ± 0.09	0.85 ± 0.09
Precision	0.87 ± 0.07	0.77 ± 0.09	0.77 ± 0.09
Improvement	0.31 ± 0.11	0.22 ± 0.12	0.23 ± 0.12

(the large difference in the number of stable *vs.* unstable compounds) since a model that favors classification of the majority class will show high accuracy but low precision. When recommending experimental synthesis of a candidate new compound, precision is perhaps more important as certainty that a predicted stable compound is truly stable is desirable. Recall is also valuable since it reflects the capability to exhaustively search a phase space and be sure that a stable compound is not missed. It is useful to compare the precision of the model to that of a random classifier, since a random classifier that labels candidates as stable or unstable 50% the of the time would show a precision equal to the fraction of stable samples in the test data. For asymmetric test sets like ours, comparison to random classification can be used as a benchmark for efficacy. The dashed lines in Figure 3(b) indicates the improvement in precision between the model scores and those of a random classifier. It shows that while the model precision drops somewhat over the seven cycles, the improvement of the model over random selection grows.

Before comparing the efficacy of ML/DFT to pure DFT, some aspects of the results of the model and an analysis of the feature space is presented.

Comparison of Random Forest, Ridge, and LASSO

Some of the reasons for the improved metrics for random forest compared to both LASSO and ridge regression can be understood from Figure 4. Figure 4 shows histograms of the scores for candidate spinels for each method, further categorized by the anion (O,S,Se,Te) present. A score near one indicates that the method attributes a greater likelihood of stability. All

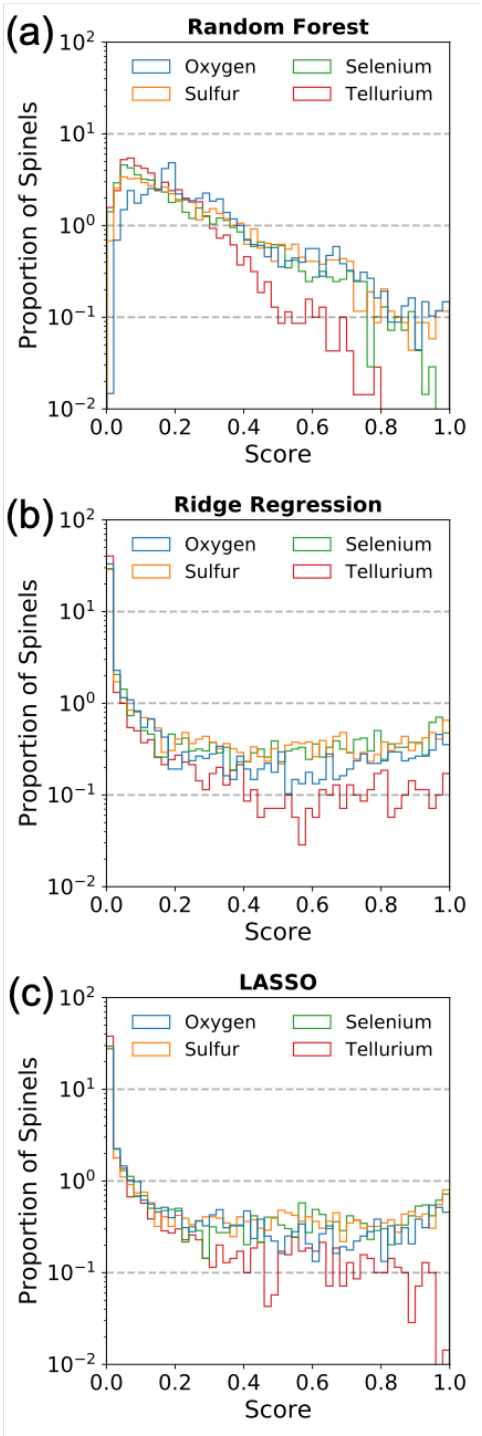

Figure 4: Histogram of predicted stability probabilities by C-anion chalcogen for (a) random forest, (b) ridge regression and (c) LASSO. Random forest shows more selectivity, with fewer high scoring candidates and disfavoring compounds with tellurium and selenium.

three methods show some degree of selectivity with a smaller proportion of compounds with high scores than with low scores. However, random forest prediction frequencies appear to exponentially decay with score (the histograms appear linear on a logarithmic plot). On the other hand ridge regression and LASSO show a spike towards the lowest scores, a rapid drop, and then a relatively flat profile for scores $>\sim 0.3$. The comparatively large proportion of high scores for these latter methods suggests that they may be overestimating the likelihood of stability. A possible reason may be that the sets of stable and unstable spinels cannot be well separated in feature space for linear methods (there may not be a hyperplane that can easily delineate the two classes).

Also from Figure 4 it can be seen that random forest disfavors both selenium and tellurium based spinels, whereas ridge regression and LASSO predict all four anions with high probability. This again may arise if the anions cannot be separated in feature-space using the linear methods. Ultimately, it is possible that random forest is capturing the reality that there are fewer stable (Se,Te) spinel compounds than (O,S). Alternatively, random forest may be responding to a bias in the dataset since there are at present fewer known stable Te and Se based spinels.

The relative selectivity of random forest is further illustrated in Figure 5, where the machine learning scores of each candidate compound are plotted. Each data point represents a candidate, and the x and y axis the candidate's score according to random forest and LASSO, respectively. The color bar shows the score according to ridge regression. There is clear correlation between the two linear techniques, indicating that the choice of regularization penalty did not have much of an effect. Additionally, it is notable that compounds that have high probability of stability according to random forest also have high probability of stability in both LASSO and Ridge, whereas the converse does not hold. In the Supporting Information, we provide a list of the top candidates according to each of the three methods at the end of the seven iterative cycles, in case they are of interest for potential synthesis experiments.

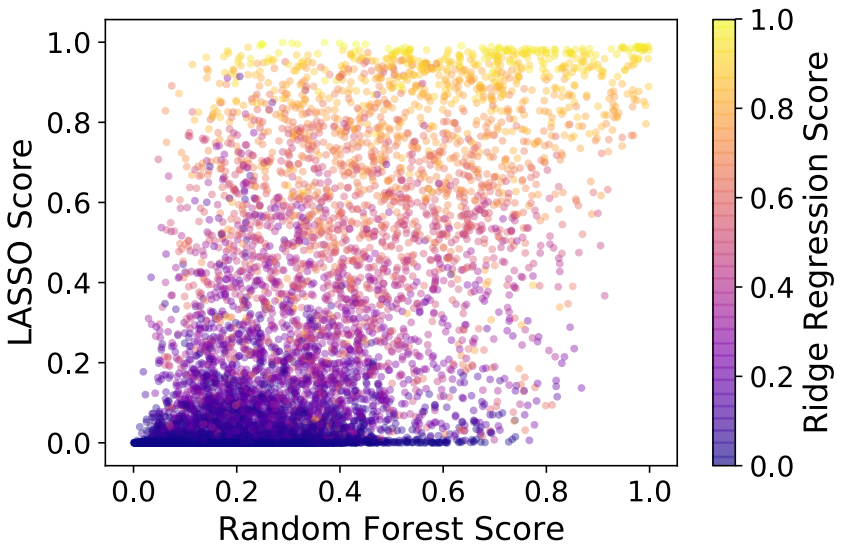

Figure 5: Scatter plot of the results of the sequential learning procedure after seven iterations. The model scores for each candidate compound are plotted for LASSO vs. random forest and colored according to the ridge regression score. The linear models are highly correlated suggesting the choice of penalty term did not have an appreciable effect on the outcome. However, there are some disparities between the linear models and random forest.

Given its superior performance, our results suggest random forest is the best classifier for predicting new compounds. Consequently, we choose to focus our attention on this method for further analysis and refer to it as RF/DFT.

Feature Importance

A challenge when applying machine learning to materials discovery is to determine which features are most important, and whether any physical significance can be ascribed to them. To determine the contribution of the different features to the model output, we conducted SHAP (SHapley Additive exPlanations) analysis on the random forest model, using the TreeSHAP code.[62,63] SHAP provides a way to determine how important each feature is (its contribution) within the full feature set to the predicted outcome. For each feature x_i in full set x, its SHAP value ϕ_i is determined from the difference in the value $v(S)$ of a subset of features S that does not contain x_i and the corresponding value $v(S \cup x_i)$ of the subset

S with feature x_i added. The value function v maps subsets S to the real numbers, and describes how effective a model based on the combination of features in S is. Formally the Shapley value for feature x_i given value function v is

$$\phi_i(v) = \sum_{S \subseteq \{x_1, \ldots x_p\} \setminus \{x_i\}} \frac{|S|!\,(p - |S| - 1)!}{p!} (v(S \cup x_i) - v(S)) \tag{4}$$

where p is the total number of features. The sum extends over all subsets S of the full feature set not containing x_i. The summation can be understood by imagining that the complete feature set is formed one feature at a time, with feature x_i assigned a contribution $v(S \cup x_i) - v(S)$, and then for each feature we average over the possible different permutations in which the feature set can be formed.

Here, for value function v we use the probability of stability itself. Then the SHAP values define an additive attribution model

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{5}$$

where ϕ_0 is an intercept, M is the number of features in the full feature set, and vector z' is a possible 'coalition' (subset) of features. In coalition z', element $z'_j = 1$ when feature j is present and $z'_j = 0$ when it is not. For the full set of features where $z' = \{1\}$, then $g(z' = \{1\})$ is simply the probability of stability, and SHAP value ϕ_i describes the extent to which feature i contributed to that probability. Large positive (negative) values of ϕ_i indicate that feature i contributed to the determination of stability (instability). Defined this way, SHAP has properties of local accuracy, consistency, and missingness. Local accuracy allows for a clear explanation of how features combine to form the output, which makes interpretation of attributions easier. Consistency ensures that if one model relies on a feature more than another model, then the feature's attribution will be greater. Missingness ensures that missing features have no attribution.

Using SHAP, for each candidate spinel and within each cycle, we can compare the attribu-

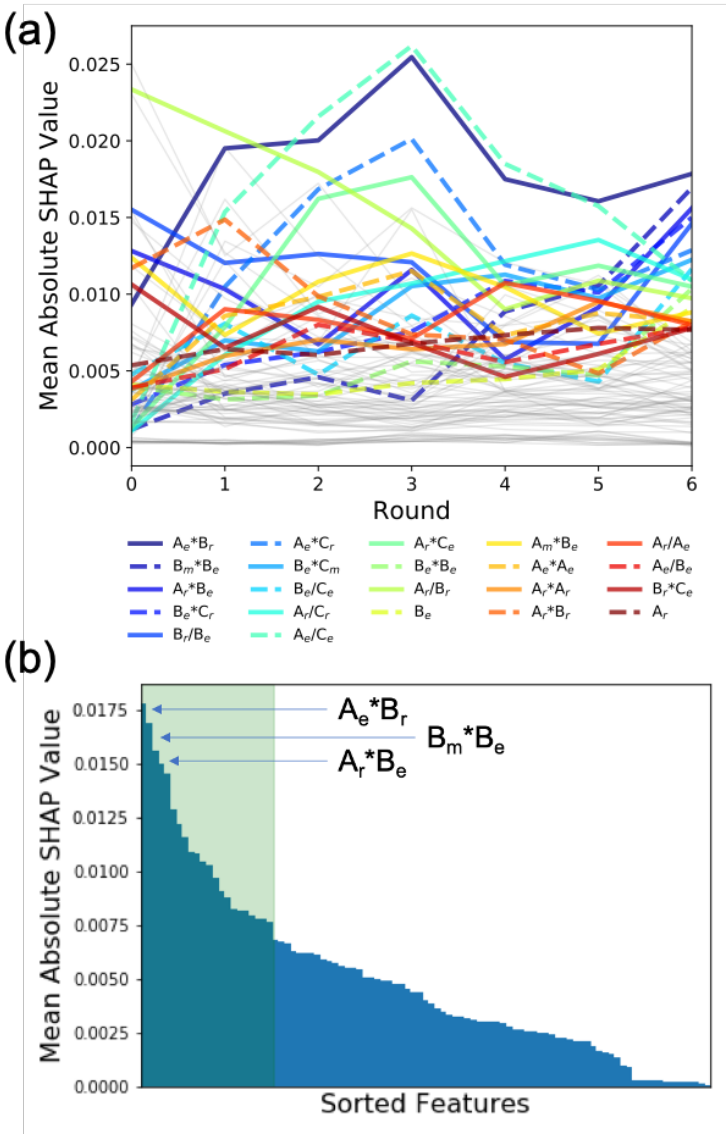

Figure 6: (a) The mean absolute SHAP calculated from the unlabeled portion of the dataset on the random forest model, for each iteration. The top 22 features in the final round are labeled. (b) Histogram of the mean absolute SHAP values for the unlabeled portion of the dataset with the top 22 features highlighted in green and the top three features labeled. (X_r: row, X_m: mass, X_e: electronegativity where X is an element in AB_2C_4)

tion of each feature to the determination of stability. These feature attributions are plotted in Figure 6(a), where each line represents the mean absolute SHAP value for a particular feature. It is clear from the plot that no single feature dominates the attribution. Rather, the output is dependent on a number of input features, rather than a few key inputs. Moreover, the SHAP values and their relative ordering vary from round to round. The addition of new data to the training set after each round causes a reorganization. For instance, the addition of new data causes the feature $A_e{*}B_r$ (electronegativity of atom A times the radius of atom B) to be the most important by the final round, while feature A_r/B_r (radius of atom A divided by radius of atom B) declines in importance. In spite of the reorganization from one round to the next, the set of 15-20 features with the largest attribution become reasonably consistent by the end of the cycles, suggesting that these features are important factors rather than the consequence of variance error from different training sets.

Figure 6(b) contains a histogram of the mean absolute SHAP values in the final round. We examined the top 22 features (highlighted in green) for further analysis, which are also colored and labeled in Figure 6(a). These top attributions are almost all from interacting features, which supports the utility of using feature engineering in this instance.

In order to determine how these features interact with the outcome, Figure 7 plots the SHAP values for the predictions of the random forest classifier in the final round of calculations for the top 22 features. Each point represents the SHAP value for a feature corresponding to a particular candidate spinel. The points are colorized in accordance with their feature value and jittering is added to better demonstrate the density of instances. While it can be difficult to interpret the significance of some of these interacting features, some values stand out. For instance, it can be seen that lower values of B_e/C_e, *i.e.* lower electronegativity of the B-site cation relative to the anion is preferable for stability, which is expected chemically. By contrast, larger values of the B-site cation to anion electronegativity result in predictions of instability. The effect of A_r/B_r is also interesting as it appears that very low and very high values signify instability, but intermediate values contribute to stability.

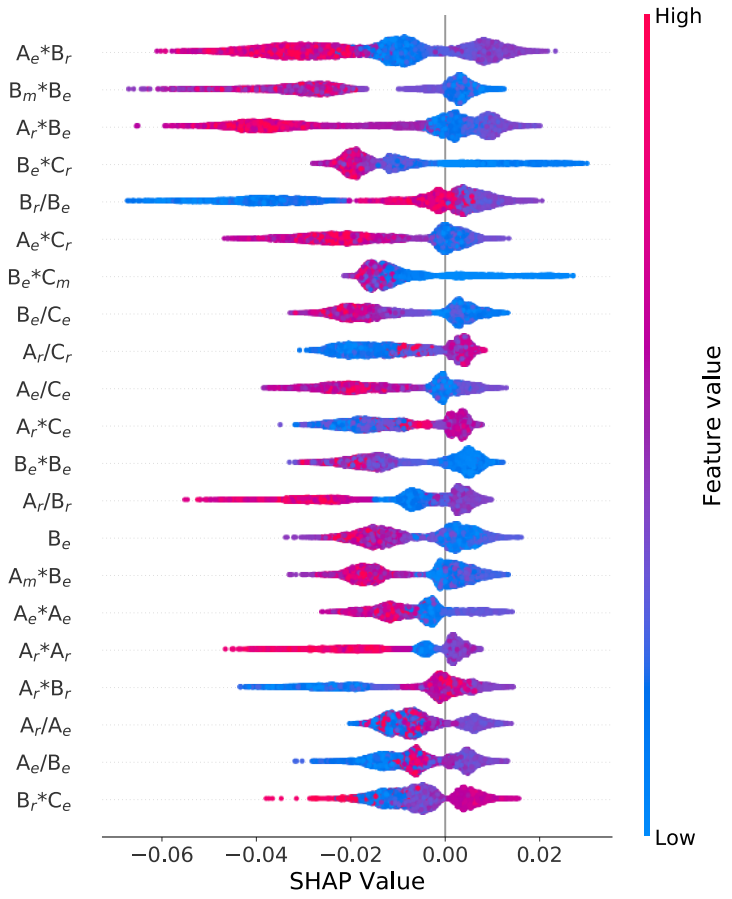

Figure 7: The SHAP value for each predicted compound derived from random forest from the final round. Only the top 22 features are shown. The location along the x-axis corresponds to the feature's SHAP value for a particular candidate. The color bar corresponds to the feature value and jittering is added along the y-axis to illustrate candidate density. (X_r: row, X_m: mass, X_e: electronegativity where X is an element in AB_2C_4)

Ratios of atomic radii have long been used as a descriptor for structural stability, such as the Goldschmidt factor for ABO_3 chemistries.[64] The atomic radius of the A cation and the electronegativity of the B cation appear to play a big factor in the predictions as both the individual attribute and its square appear in the top attributions. In this case, low values of the electronegativity and average values of the atomic radius of the respective cations correlates to greater probability of stability.

Search Completeness and Search Efficiency of Materials Discovery via RF/DFT and DFT

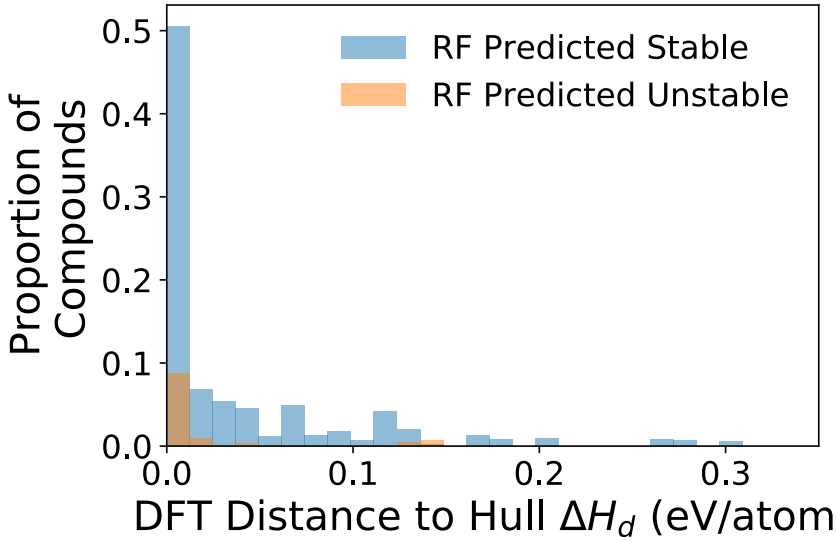

Figure 8: Histogram of the distance to convex hull and their random forest predicted stability for known stable compounds across all cross-validated test sets, for a bin size of 0.01 eV. Random forest performs well throughout and identifies stable compounds even when their DFT distance to hull is large; the compounds predicted as unstable by random forest are distributed evenly amongst all data points and appear not to be correlated to DFT distance to hull.

We turn to the question of how to compare the effectiveness of RF/DFT to pure DFT when searching for new materials. The question of interest is "which approach should an experimentalist use to guide their search for new materials?". The RF/DFT approach presented here is distinct from prior ML frameworks, most of which[2–7] are predicated on learn-

ing DFT formation energies across a variety of materials. While it has been demonstrated that DFT formation energies can be learned with reasonable accuracy, DFT decomposition energies typically are not as well–reproduced.[1] Also since these methods are assessed by comparison to DFT, there is no way to determine if they can outperform DFT.

Here, rather than learning a formation enthalpy we trained an RF model to make a binary classification of stable or unstable. The assessment of RF/DFT in Figure 3 shows good performance on traditional metrics. However, the training and testing sets used are different from what would be encountered when actually searching through chemical phase space for new spinels. First, they contain a roughly equal number of stable and unstable compounds. When trying to discover a new composition, reality presents a much more challenging search problem: there are a relatively small number of stable compounds contained within a large, mostly unstable search space. An effective approach should select the few stable compounds from a set that is several orders of magnitude larger. Second, based on how our approach is formulated, candidate compounds for which $0 \leq \Delta H_d \leq 0.2$ eV/atom (perhaps the more uncertain compounds) are not included in testing or training. When assessing both ML/DFT and DFT, it is important to assess the models based on an unbiased sample. In the following discussion, we consider a hypothetical search for new spinels based on a phase space that more closely reflects this reality.

Additionally, model efficacy should be measured in a way that reflects the outcomes of a series of synthesis experiments on candidate compounds. In a realistic scenario, candidate materials would be rank ordered by their likelihood of stability, and synthesis experiments would be carried out starting from the highest ranked candidate downwards. To an experimentalist, the important quantities may be (i) how many new compounds will be discovered, if 10, 100, or 1000 synthesis experiments are carried out? and (ii) how many experiments need to be done to ensure that, say 50%, 75%, or 99% of the discoverable compounds in the search space are found? These questions can be answered by assessing *search completeness* and *search efficiency*. Search completeness measures the proportion of discoverable

compounds found as a function of the fraction of search space explored. It reflects how exhaustively a search space needs to be explored to identify specified proportions of all missing compounds. Search efficiency measures the discovery rate. It indicates the number of compounds found per experiment carried out, also as a function of the fraction of search space explored. Because search efficiency and search completeness reflect the actual experience and outcomes of a hypothetical search, they are useful measures of model performance. However, existing analyses of DFT or ML have rarely (if ever) attempted to evaluate these.

Bayesian Model of DFT Distribution of Distance to Hull for Stable and Unstable Compounds

Comparing the search completeness and efficiency of DFT and RF/DFT is complicated by the fact that knowledge of the 'ground truth', *i.e.* the true sets of stable and unstable AB_2C_4 spinels, is unknown. The lack of this knowledge makes it impossible to strictly assess either method. To carry out the analysis, a model for the distribution of DFT ΔH_d for both true stable and unstable spinels is needed. We present a Bayesian model that estimates these distributions. A DFT-based classification scheme can be tested on the distributions, and its search completeness and efficiency compared to that of ML/DFT. The Bayesian model is based on that of Narayan *et al.* developed while investigating transition metal sulfides and selenides.[65] The main quantity of interest is the probability that a spinel is stable given that its DFT-computed distance to hull ΔH_d lies within a designated cutoff ϵ. It is given by

$$P(S|\Delta H_d \leq \epsilon) = \frac{P(\Delta H_d \leq \epsilon|S)P(S)}{P(\Delta H_d \leq \epsilon)} \quad . \tag{6}$$

This quantity can also be interpreted as the precision of a DFT-based classification scheme at a particular value of ϵ, if all compounds with $\Delta H_d \leq \epsilon$ are classified as stable.

To get an idea of the distribution of ΔH_d for stable spinels, Figure 8 shows the set of experimentally known spinels for which we have DFT calculations (138 out of 200 total)

plotted according to their DFT distance to hull ΔH_d. The distribution peaks near zero, and then decays with increasing distance to hull. Figure 8 shows that if DFT results were interpreted 'as-is', only the compounds on the hull would be classified as stable and many stable compounds would be missed (false negatives). Of the 138 spinels, 83 are on the hull – yielding an estimated recall of ≈ 0.60. Additionally, there are 32 additional spinels predicted to be on the hull, that are not experimentally known. Assuming these are false positives, this yields DFT a precision of ≈ 0.72 for compounds on the hull. The estimated precision and recall here will be used as boundary conditions for our Bayesian model for $\epsilon \to 0$.

We also note that only 4 of the 138 compounds exhibit DFT $\Delta H_d > 0.2$ eV/atom, supporting the criterion adopted in our ML/DFT approach that compounds that have DFT $\Delta H_d > 0.2$ be labeled as unstable. The data points in Figure 8 have also been colored based on their classification according to random forest. Random forest consistently labels the stable compounds as stable, even for those compounds that according to DFT lie farther from the convex hull. One observation is that large DFT ΔH_d do not necessarily result in a classification of unstable within random forest. Rather, of the stable compounds that are (incorrectly) labeled as unstable, they appear to be distributed largely in proportion to the density of compounds at each given ΔH_d.

To obtain $P(S|\Delta H_d \leq \epsilon)$ in Equation (6), we approximate the quantities on the right hand side using the dataset generated by our study. This model and the resulting distributions are therefore applicable to the space of compounds considered: candidate 2–3 spinels AB_2C_4 where A, B are elements with known oxidation states of $+2, +3$ respectively and C is restricted to O, S, Se, and Te. From pure elemental substitutions, this generates a set of $\sim$14,200 total possible compounds of which 200 are the known stable spinels.

In Equation (6), the distribution $P(S)/P(\Delta H_d \leq \epsilon)$ represents the total number of stable spinels divided by the total number of spinels with $\Delta H_d \leq \epsilon$. It is approximated by

$$\frac{P(S)}{P(\Delta H_d \leq \epsilon)} \approx \frac{N_S}{N(\Delta H_d \leq \epsilon)} \tag{7}$$

where N_S is the total number of stable compounds in the dataset and $N(\Delta H_d \leq \epsilon)$ is the total number of compounds within ϵ of the hull. Moreover, $N(\Delta H_d \leq \epsilon) = N_S * P(\Delta H_d \leq \epsilon|S) + N_U * P(\Delta H_d \leq \epsilon|U)$, where N_U is the number of unstable compounds. Assuming the number of undiscovered experimentally stable spinels in our dataset is small, we can approximate N_S as the number of stable compounds in our dataset (200) and N_U as all spinels in our dataset not labeled stable (∼14,000). $P(\Delta H_d < \epsilon|S)$ and $P(\Delta H_d < \epsilon|U)$ represent, respectively, the probability of finding a stable or unstable compound with hull distance $\Delta H_d < \epsilon$.

These two distributions are estimated from histograms of our DFT computed data, the blue and red bars shown in Figure 9. The logspline package in the statistical computing language R[66] was employed to estimate their respective distributions, which fits a spline function to the log-density of the inputs. This method is effective for bounded data (since $\Delta H_d \not< 0$) and utilizes a Bayesian information criterion to determine the number of knots in the spline. The resulting distributions generated by the logspline estimator are shown in black. The splines were fitted to our DFT data, ensuring that boundary conditions at $\Delta H_d = 0$ are matched (precision and recall are fixed to ≈ 0.72 and ≈ 0.60). The stable distribution is clustered around $\Delta H_d = 0$ and rapidly decaying, showing that given a stable compound, its most probable ΔH_d is zero. The unstable distribution resembles a lognormal distribution. It initially increases and peaks around $\Delta H_d \approx 0.2$ eV/atom, and then decreases (the dip around $\Delta H_d = 0.3$ eV/atom is likely an artifact). The initial increase reflects that, given that a compound is unstable, the most probable ΔH_d is not on the hull but lies some distance above it. The subsequent decrease might reflect our selection criteria for the compounds under consideration (*i.e.* AB_2C_4 compounds where A, B element have known oxidation states of +2,+3), as such compounds are not so unreasonable so their ΔH_d lies within a finite distance of the hull.

Figure 10(a) plots $P(S|\Delta H_d \leq \epsilon)$ for DFT from Equation (6), equivalent to the precision of the DFT classifier. As expected, the precision is highest close to or on the convex hull,

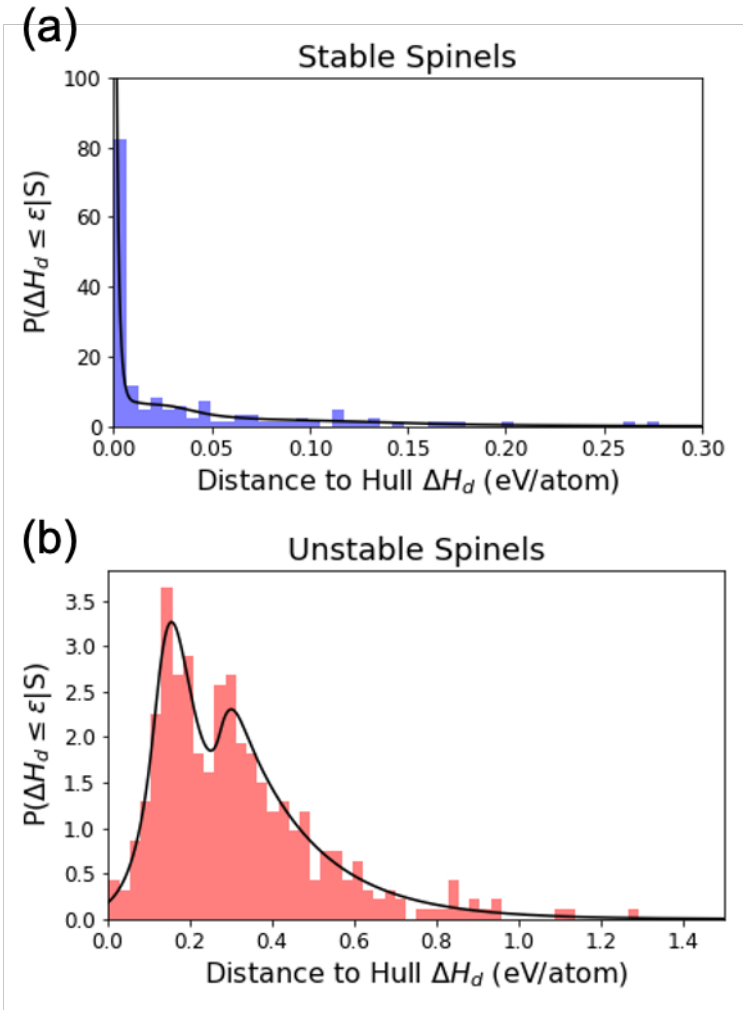

Figure 9: Distribution of DFT distance to hull ΔH_d for (a) stable and (b) unstable spinels, with log-spline smoothing superimposed. Fitted curves are normalized to integrate to one. The stable spinel compounds primarily group around the convex hull, whereas unstable compounds are offset and exhibit a greater variance.

≈ 0.72 at $\epsilon = 0$. As cutoff ϵ increases (blue line in Figure 10(a)), DFT precision drops and lies below 0.1 for cutoff $\epsilon = 0.2$ eV/atom.

Results for RF/DFT and DFT

We can now evaluate the search completeness and efficiency of ML/DFT and DFT based on the Bayesian prediction of the DFT distributions of ΔH_d for stable and unstable compounds. We use the Bayesian model for $P(S|\Delta H_d \leq \epsilon)$, and define a DFT classification scheme in which a compound is classified as stable if $\Delta H_d \leq \epsilon$ and unstable if $\Delta H_d > \epsilon$.

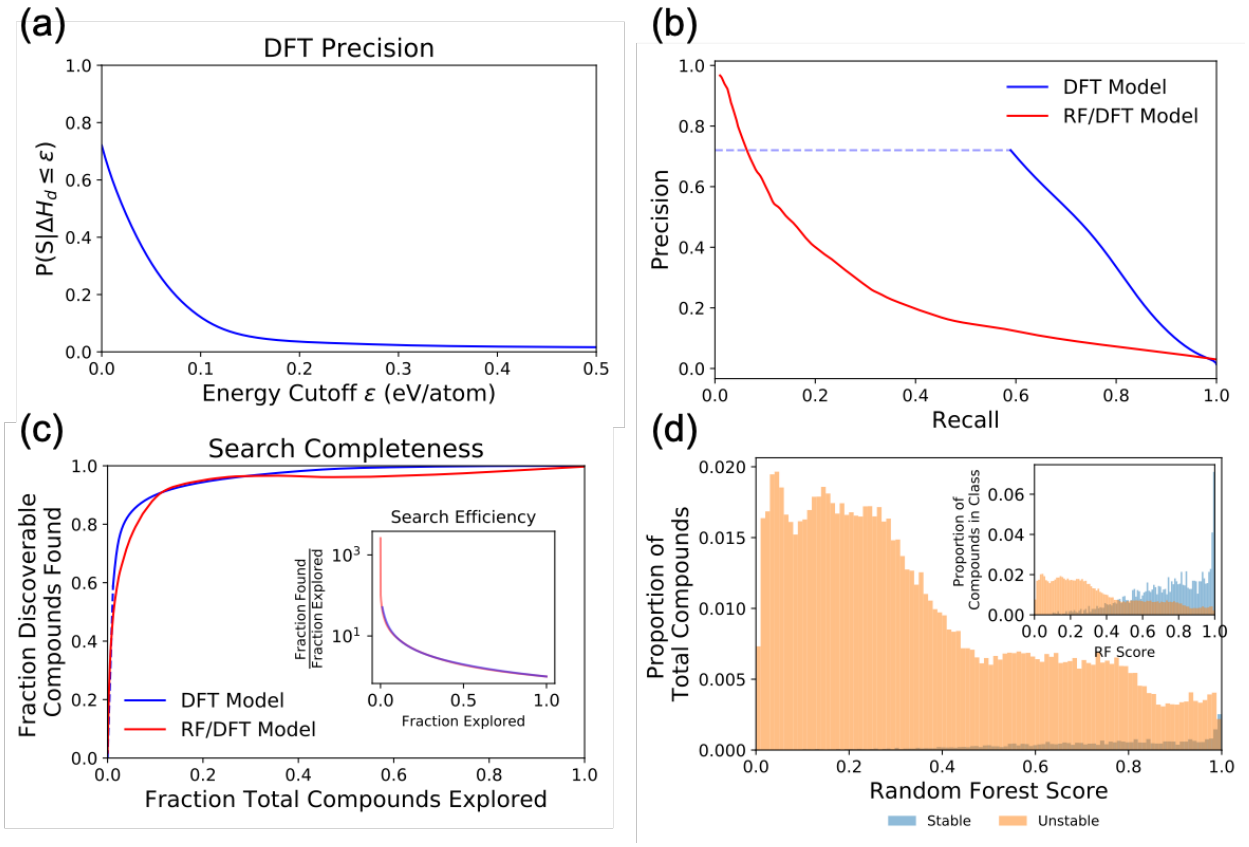

Figure 10: (a) Precision of DFT model as a function of energy cutoff ϵ. (b) Experiment efficiency for each approach is shown by plotting the fraction of discoverable compounds found (recall) against the fraction of the phase space explored. (c) The precision-recall curves for DFT and the RF/DFT models. The random forest model in (b) and (c) have been averaged across 100 tests using LOESS regression. (d) The distribution of stable and unstable compounds with respect to their random forest score (bin size 0.01 eV).

In order to also assess ML/DFT on an unbiased dataset that more closely matches the

Bayesian model that DFT is evaluated on, we fitted another random forest model. The training set is similar to before and is comprised of 75% of the stable spinels and all ∼160 unstable labeled spinels as classified by the $\Delta H_d > 0.2$ eV/atom energy cutoff. The remaining 25% of the stable spinels (50 in total) were withheld for testing. The testing set also included all other compounds not found in the training set, which were labeled unstable based on the assumption that there are few stable spinels that remain to be discovered. This results in a test set that contains 50 stable compounds to be discovered mixed in with a set of ∼ 13900 unstable compounds. This better reflects the inherent challenge for any approach to materials prediction: there are typically a few undiscovered compounds in a large search space (the 'needle in a haystack'). An effective method should cleanly separate the few stable compounds from the unstable ones.

To begin we compare the precision–recall curves of DFT and RF/DFT in Figure 10(b). The threshold varied is the the probability of stability for RF/DFT as before, and cutoff parameter ϵ for DFT. This was done on each of the 100 test sets and the resulting combined data was smoothed using LOESS regression. To perform the LOESS regression, we utilized the R package fANCOVA, implementing generalized cross validation in order to determine the bandwidth.[67] For the DFT model, ϵ was varied and the resulting precisions and recalls plotted. Here, RF/DFT and DFT show qualitatively different behaviors. In DFT (area under curve: 0.57), the horizontal dotted line is associated with the cluster of stable compounds with $\Delta H_d = 0$, where precision and recall are fixed to ≈ 0.72 and ≈ 0.60. In DFT as threshold ϵ increases, recall further increases (fewer false negatives) as more of the stable instances are classified correctly, but the precision decreases too as more unstable instances are incorrectly classified as stable (more false positives). For RF/DFT (AUC: 0.25), initially the precision exceeds that of the DFT-model, demonstrating a high selection capability and low false positive rate amongst the highest scoring candidates. However, the precision very rapidly decreases as the threshold score is decreased – showing that unstable compounds quickly become mixed in to the stable classification. The difference between the precision–

recall curves in Figure 3(d) and Figure 10(b) shows the sensitivity of the model performance to the data set, and highlights the importance of using data sets that better reflect reality.

Figure 10(c) shows the search completeness and the inset shows the search efficiency for DFT and RF/DFT. The vertical axis – the fraction of discoverable compounds discovered – is equivalent to the model recall. Ideally, the recall would rise rapidly with increased testing, limiting the need for excess experimentation. For the DFT model, we assume that experiments are carried out in order of ΔH_d, smallest to largest. For RF/DFT, as before, the tests were done 100 times with random splitting of the stable data to gather statistics. The ML/DFT model is LOESS–regressed across the 100 tests. The search completeness of RF/DFT is on par with DFT with the two curves close to each other. For instance, the RF/DFT approach is able to discover half of the discoverable compounds by scanning only $\approx 1.25\%$ of the search space, similar to DFT. The inset shows the search efficiency, which is the derivative of the search completeness. For both DFT and RF/DFT the search efficiency is highest at the beginning, but decreases quickly as more of the search space is explored. As more of the search space is explored, more unstable compounds become mixed in to the high scoring candidates.

While the search completeness and efficiency of RF/DFT and DFT are comparable, the DFT results assume that ΔH_d has been computed and is available for all $\sim$14200 compounds. In contrast, the RF/DFT approach required only $\sim$280 simulations throughout the seven iterative cycles to generate the $\sim$160 unstable labeled compounds. Rather than being provided the complete unstable set of compounds, *i.e.* all compounds in the full set for which $\Delta H_d > 0.2$, RF/DFT was fitted to only a small subset of the full data set, and then tested on the full data set. So, ML/DFT shows a nearly indistinguishable search efficiency, despite requiring substantially fewer DFT calculations. This demonstrates the possible utility of such an approach, which could be used to prioritize experiments using a limited amount of DFT calculations in order to improve success rate.

To better understand the comparable search efficiencies in spite of the qualitatively dif-

ferent precision/recall curves for RF/DFT and DFT, Figure 10(d) shows the distribution of RF scores for stable and unstable compounds according to RF/DFT. The challenge becomes evident here. As desired most of the stable compounds score high and most of the unstable compounds score low, indicating that there is reasonable separation in the two classes. This can be more clearly seen in the inset of Figure 10(d) which shows the proportion of compounds relative to each separate classification (stable or unstable). Although the classes are reasonably well separated, the sheer number of unstable compounds causes the tail of the unstable distribution to overwhelm the higher scoring stable distribution. This points to possible directions for improving the model's selectivity, such as including more unstable training data. The challenge revealed by this analysis is inherent to any data-based approach to materials discovery, and will be faced by all materials discovery frameworks. We hope that by clearly elucidating the nature of the challenge, we can provide some guidelines for identifying effective routes to overcome it.

Conclusion

In this work we present a combined ML/DFT framework to predict stable chemistries of a given crystal structure, and introduce an approach to assessing the efficacy of the approach in comparison to that of direct physical modeling (DFT) alone. In ML/DFT, DFT is used as a generator of unstable compounds to be used for ML training and testing, while the stable set of compounds is given by those that are experimentally reported. The method is applied to spinel compounds, and traditional supervised ML methods (random forest, ridge regression, and LASSO) are used to predict new compounds with high probability of stability. Rather than comparing the performance of ML/DFT to the predictions of DFT itself, the performance of both DFT and ML/DFT is assessed by comparison to reality. We approximate reality by introducing a Bayesian model that allows us to infer the distribution of the DFT-predicted distance to hull for stable and unstable compounds. The concepts

of search completeness – the ability to exhaustively search a phase space and identify all missing compounds – and search efficiency – the success rate for finding a new compound – are defined and proposed to be an appropriate measure of ML efficacy. On one hand, we find that the ML/DFT approach described here obtains search completeness and efficiency on par with that of DFT when searching for undiscovered spinels. On the other hand, the use of realistic data sets highlights the inherent challenge to be overcome by all materials discovery approaches: namely that being able to cleanly separate the few discoverable stable compounds from the large phase space of unstable ones places stringent demands on model accuracy to achieve high search completeness and efficiency.

Acknowledgements

We would like to thank the National Science Foundation for their support of this work through a Graduate Research Fellowship to J.A.S. Computational resources were provided by the Illinois Campus Cluster. We would also like to thank the advice provided by Lucas K. Wagner, Harley Johnson, Daniel Shoemaker and Sameh Tawfick.

Supporting Information

Tabulated list of the top 50 stable spinel candidates predicted according to LASSO, ridge regression, and random forest.

References

(1) Bartel, C. J.; Trewartha, A.; Wang, Q.; Dunn, A.; Jain, A.; Ceder, G. A critical examination of compound stability predictions from machine-learned formation energies. *npg Computational Materials* **2020**, *97*.

(2) Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Physical Review B* **2014**, *89*, 094104.

(3) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials* **2016**, *2*, 16028.

(4) Dunn, A.; Wang, Q.; Ganose, A.; Dopp, D.; Jain, A. Benchmarking Materials Property Prediction Methods: The Matbench Test Set and Automatminer Reference Algorithm. *arXiv preprint arXiv:2005.00707* **2020**,

(5) Jha, D.; Ward, L.; Paul, A.; Liao, W.-k.; Choudhary, A.; Wolverton, C.; Agrawal, A. Elemnet: Deep learning the chemistry of materials from only elemental composition. *Scientific reports* **2018**, *8*, 1–13.

(6) Goodall, R. E.; Lee, A. A. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *arXiv preprint arXiv:1910.00617* **2019**,

(7) Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters* **2018**, *120*, 145301.

(8) Oganov, A. R.; Valle, M. How to quantify energy landscapes of solids. *The Journal of chemical physics* **2009**, *130*, 104504.

(9) Zunger, A. Inverse design in search of materials with target functionalities. *Nat Rev Chem* **2018**, *2*, 0121.

(10) Hohenberg, P.; Kohn, W. Inhomogeneous electron gas. *Physical review* **1964**, *136*, B864.

(11) Kohn, W.; Sham, L. J. Self-consistent equations including exchange and correlation effects. *Physical review* **1965**, *140*, A1133.

(12) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. a. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **2013**, *1*, 011002.

(13) Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *Jom* **2013**, *65*, 1501–1509.

(14) Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials* **2015**, *1*, 1–15.

(15) 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. *Physical Review B* **2012**, *85*, 115104.

(16) others,, et al. AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. *arXiv preprint arXiv:2003.12476* **2020**,

(17) Pizzi, G.; Cepellotti, A.; Sabatini, R.; Marzari, N.; Kozinsky, B. AiiDA: automated interactive infrastructure and database for computational science. *Computational Materials Science* **2016**, *111*, 218–230.

(18) Draxl, C.; Scheffler, M. The NOMAD laboratory: from data sharing to artificial intelligence. *Journal of Physics: Materials* **2019**, *2*, 036001.

(19) others,, et al. AFLOWLIB. ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Computational Materials Science* **2012**, *58*, 227–235.

(20) Lany, S. Semiconductor thermochemistry in density functional calculations. *Physical Review B* **2008**, *78*, 245207.

(21) Oganov, A. R.; Glass, C. W. Crystal structure prediction using ab initio evolutionary techniques: Principles and applications. *The Journal of chemical physics* **2006**, *124*, 244704.

(22) Lyakhov, A. O.; Oganov, A. R.; Stokes, H. T.; Zhu, Q. New developments in evolutionary structure prediction algorithm USPEX. *Computer Physics Communications* **2013**, *184*, 1172–1182.

(23) Oganov, A. R.; Lyakhov, A. O.; Valle, M. How Evolutionary Crystal Structure Prediction Works and Why. *Accounts of chemical research* **2011**, *44*, 227–237.

(24) Wang, Y.; Lv, J.; Zhu, L.; Ma, Y. Crystal structure prediction via particle-swarm optimization. *Physical Review B* **2010**, *82*, 094116.

(25) Pannetier, J.; Bassas-Alsina, J.; Rodriguez-Carvajal, J.; Caignaert, V. Prediction of crystal structures from crystal chemistry rules by simulated annealing. *Nature* **1990**, *346*, 343–345.

(26) Hautier, G.; Fischer, C. C.; Jain, A.; Mueller, T.; Ceder, G. Finding natures missing ternary oxide compounds using machine learning and density functional theory. *Chemistry of Materials* **2010**, *22*, 3762–3767.

(27) Fischer, C. C.; Tibbetts, K. J.; Morgan, D.; Ceder, G. Predicting crystal structure by merging data mining with quantum mechanics. *Nature materials* **2006**, *5*, 641–646.

(28) others,, et al. Crystallographic databases. *International Union of Crystallography, Chester* **1987**, *360*, 77–95.

(29) Allmann, R.; Hinek, R. The introduction of structure types into the Inorganic Crystal Structure Database ICSD. *Acta Crystallographica Section A: Foundations of Crystallography* **2007**, *63*, 412–417.

(30) Belsky, A.; Hellenbrandt, M.; Karen, V. L.; Luksch, P. New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. *Acta Crystallographica Section B: Structural Science* **2002**, *58*, 364–369.

(31) Zagorac, D.; Müller, H.; Ruehl, S.; Zagorac, J.; Rehme, S. Recent developments in the Inorganic Crystal Structure Database: theoretical crystal structure data and related features. *Journal of applied crystallography* **2019**, *52*.

(32) Meredig, B.; Antono, E.; Church, C.; Hutchinson, M.; Ling, J.; Paradiso, S.; Blaiszik, B.; Foster, I.; Gibbons, B.; Hattrick-Simpers, J.; Mehta, A.; Ward, L. Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. *Mol. Syst. Des. Eng.* **2018**, *3*, 819–825.

(33) Sutton, C.; Boley, M.; Ghiringhelli, L.; Rupp, M.; Vreeken, J.; Scheffler, M. Identifying domains of applicability of machine learning models for materials science. *Nat Commun* **2020**, *11*, 4428.

(34) 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.* **2019**, *31*, 3564–3572.

(35) A. Broese Van Groenou, P. F. Bongers, A. S. Magnetism , Microstrueture and Crystal Chemistry of Spinel Ferrites. *Mater. Sci. Eng.* **1968**, *3*, 317–392.

(36) Johnston, D. Superconducting and normal state properties of Li 1+ x Ti 2- x O 4 spinel compounds. I. Preparation, crystallography, superconducting properties, electrical resistivity, dielectric behavior, and magnetic susceptibility. *Journal of Low Temperature Physics* **1976**, *25*, 145–175.

(37) Thackeray, M.; De Kock, A.; Rossouw, M.; Liles, D.; Bittihn, R.; Hoge, D. Spinel electrodes from the Li-Mn-O system for rechargeable lithium battery applications. *Journal of The Electrochemical Society* **1992**, *139*, 363–366.

(38) Howard Jr, W. F.; Sheargold, S. W.; Story, P. M.; Peterson, R. L. Stabilized spinel battery cathode material and methods. 2003; US Patent 6,558,844.

(39) Hosono, E.; Kudo, T.; Honma, I.; Matsuda, H.; Zhou, H. Synthesis of single crystalline spinel LiMn2O4 nanowires for a lithium ion battery with high power density. *Nano letters* **2009**, *9*, 1045–1051.

(40) Liu, M.; Jain, A.; Rong, Z.; Qu, X.; Canepa, P.; Malik, R.; Ceder, G.; Persson, K. A. Evaluation of sulfur spinel compounds for multivalent battery cathode applications. *Energy Environ. Sci.* **2016**, *9*, 3201–3209.

(41) Gao, X.-W.; Deng, Y.-F.; Wexler, D.; Chen, G.-H.; Chou, S.-L.; Liu, H.-K.; Shi, Z.-C.; Wang, J.-Z. Improving the electrochemical performance of the LiNi _{0.5} Mn _{1.5} O ₄ spinel by polypyrrole coating as a cathode material for the lithium-ion battery. *J. Mater. Chem. A* **2015**, *3*, 404–411.

(42) Dou, S. Review and prospects of Mn-based spinel compounds as cathode materials for lithium-ion batteries. *Ionics* **2015**, *21*, 3001–3030.

(43) Kawazoe, H.; Ueda, K. Transparent Conducting Oxides Based on the Spinel Structure. *Journal of the American Ceramic Society* **1999**, *36*, 3330–3336.

(44) Windisch Jr, C. F.; Ferris, K. F.; Exarhos, G. J. Synthesis and characterization of transparent conducting oxide cobalt–nickel spinel films. *Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films* **2001**, *19*, 1647–1651.

(45) Stevanović, V.; D'Avezac, M.; Zunger, A. Universal electrostatic origin of cation ordering in A2BO 4 spinel oxides. *Journal of the American Chemical Society* **2011**, *133*, 11649–11654.

(46) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(47) Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J.; Layton, R.; VanderPlas, J.; Joly, A.; Holt, B.; Varoquaux, G. API design for machine learning software: experiences from the scikit-learn project. ECML PKDD Workshop: Languages for Data Mining and Machine Learning. 2013; pp 108–122.

(48) Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M. Big data of materials science: critical role of the descriptor. *Physical review letters* **2015**, *114*, 105503.

(49) Ouyang, R.; Curtarolo, S.; Ahmetcik, E.; Scheffler, M.; Ghiringhelli, L. M. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Physical Review Materials* **2018**, *2*, 083802.

(50) Kresse, G.; Hafner, J. Ab initio molecular dynamics for liquid metals. *Physical Review B* **1993**, *47*, 558.

(51) Kresse, G.; Hafner, J. Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium. *Physical Review B* **1994**, *49*, 14251.

(52) Kresse, G.; Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational materials science* **1996**, *6*, 15–50.

(53) Kresse, G.; Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical review B* **1996**, *54*, 11169.

(54) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Physical review letters* **1996**, *77*, 3865.

(55) Blöchl, P. E. Projector augmented-wave method. *Physical review B* **1994**, *50*, 17953.

(56) Kresse, G.; Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical review b* **1999**, *59*, 1758.

(57) Ong, S. P.; Wang, L.; Kang, B.; Ceder, G. Li-Fe-P-O2 Phase Diagram from First Principles Calculations. *Chemistry of Materials* **2008**, *20*, 1798–1807.

(58) Jain, A.; Hautier, G.; Ong, S.; Moore, C.; Fischer, C.; Persson, K.; Ceder, G. Formation enthalpies by mixing GGA and GGA + U calculations. *Physical Review B* **2011**, *84*, 045115.

(59) Ong, S. P.; Cholia, S.; Jain, A.; Brafman, M.; Gunter, D.; Ceder, G.; Persson, K. A. 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* **2015**, *97*, 209–215.

(60) Perron, H.; Mellier, T.; Domain, C.; Roques, J.; Simoni, E.; Drot, R.; Catalette, H. Structural investigation and electronic properties of the nickel ferrite NiFe2O4: a peri-

odic density functional theory approach. *Journal of Physics: Condensed Matter* **2007**, *19*, 346219.

(61) Huang, J.-R.; Cheng, C. Cation and magnetic orders in MnFe2O4 from density functional calculations. *Journal of Applied Physics* **2013**, *113*, 033912.

(62) Lundberg, S. M.; Lee, S.-I. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. 2017; pp 4765–4774.

(63) Lundberg, S. M.; Erion, G. G.; Lee, S.-I. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888* **2018**,

(64) Goldschmidt, V. M. Die gesetze der krystallochemie. *Naturwissenschaften* **1926**, *14*, 477–485.

(65) Narayan, A.; Bhutani, A.; Rubeck, S.; Eckstein, J. N.; Shoemaker, D. P.; Wagner, L. K. Computational and experimental investigation for new transition metal selenides and sulfides: The importance of experimental verification for stability. *Physical Review B* **2016**, *94*, 045105.

(66) Kooperberg, C.; Stone, C. J. Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics* **1992**, *1*, 301–328.

(67) Golub, G. H.; Heath, M.; Wahba, G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **1979**, *21*, 215–223.

Supplementary Information

A Combined DFT/Machine Learning Framework for Materials Discovery: Application to Spinels and Assessment of Search Completeness and Efficiency

Joshua A. Schiller and Elif Ertekin*
Department of Mechanical Science & Engineering, 1206 W Green Street, University of Illinois at Urbana-Champaign, Urbana IL 61801

Highest ranking predicted candidate spinels at the end of the sequential learning procedure:

LASSO		Ridge Regression		Random Forest	
Compound	**Probability**	**Compound**	**Probability**	**Compound**	**Probability**
CuNi2O4	0.9933	SiCl2O4	0.9987	VCr2O4	1
CuCu2O4	0.9923	SiCl2Te4	0.9942	CuNi2S4	0.999
NiCu2O4	0.9922	NiCl2O4	0.9904	NiCr2S4	0.998
HgSc2O4	0.991	PCl2O4	0.9901	CoCo2S4	0.997
CoCu2O4	0.9899	InSc2O4	0.9899	CoFe2S4	0.996
FeCu2O4	0.9884	GaCl2O4	0.9894	CuFe2S4	0.996
ZnSc2O4	0.9884	SiCl2S4	0.9886	NiFe2S4	0.995
GaSc2O4	0.9882	CuCl2O4	0.9885	CuNi2O4	0.995
GaNi2O4	0.9873	GeCl2O4	0.9883	CrV2O4	0.994
NiCu2S4	0.9862	CdSc2O4	0.9874	MnMn2O4	0.992
CuSc2O4	0.986	SiBr2O4	0.9871	ZnMn2O4	0.988
CuCu2S4	0.9859	CoCl2O4	0.9868	FeFe2S4	0.987
GeCu2O4	0.9857	ZnNi2O4	0.9863	CoCu2O4	0.986
CuNi2S4	0.9856	ZnCl2O4	0.9859	CuMn2S4	0.986
GaCu2O4	0.985	FeCl2O4	0.9849	FeCu2O4	0.984
SnNi2O4	0.9839	AgSc2O4	0.9849	CuIn2S4	0.984
NiSc2O4	0.9834	ZnSc2O4	0.9846	NiCu2O4	0.979

SnLu2S4	0.9826	ZnCu2O4	0.9835	CuCu2O4	0.979
SnCu2O4	0.9824	SnSc2O4	0.9834	ZnNi2O4	0.978
ZnNi2O4	0.9824	CuNi2O4	0.983	CoMn2S4	0.976
CuFe2S4	0.9822	GaNi2O4	0.9828	VAl2O4	0.976
NiFe2S4	0.9822	SiS2O4	0.9822	FeCo2S4	0.976
GaCo2O4	0.9816	InFe2O4	0.9805	CuCu2S4	0.973
CuLu2S4	0.9815	ScSc2O4	0.9796	CoCu2S4	0.973
SnSc2O4	0.9814	CuCu2O4	0.9795	CuAl2S4	0.971
CoSc2O4	0.9814	GaCu2O4	0.9795	CrGa2O4	0.971
CoCu2S4	0.9809	FeCu2O4	0.9787	MnAl2O4	0.971
NiLu2S4	0.9804	NiCu2O4	0.9783	NiCu2S4	0.971
GeSc2O4	0.9801	NiBr2O4	0.9766	NiMn2S4	0.969
TeFe2O4	0.9798	SiCl2Se4	0.9752	TcNi2O4	0.968
CrCu2O4	0.9798	AsCl2O4	0.9752	CoAl2S4	0.964
ZnCu2O4	0.979	CrCl2O4	0.9748	MoCo2O4	0.964
CrSc2O4	0.9787	SnNi2O4	0.9747	FeAl2S4	0.964
CuNi2Se4	0.9787	GaCo2O4	0.9741	CoV2O4	0.963
FeCu2S4	0.9787	MgSc2O4	0.9741	RhNi2O4	0.961
CuCu2Se4	0.9782	AgFe2O4	0.9734	NiV2O4	0.961
NiNi2Se4	0.9775	CuBr2O4	0.9733	TiTi2O4	0.959
CdSc2O4	0.9772	PCl2Te4	0.9732	NiAl2S4	0.959
NiCu2Se4	0.9769	CuSc2O4	0.9725	VNi2O4	0.955
SnCu2S4	0.9762	VCl2O4	0.972	PdCo2O4	0.954
HgLu2S4	0.9761	GaSc2O4	0.9718	CuV2O4	0.952
CuFe2Se4	0.9758	CoCu2O4	0.9717	FeCu2S4	0.952
CoFe2S4	0.9758	InNi2O4	0.9717	MoCu2O4	0.951
TeNi2O4	0.9751	TeFe2O4	0.9707	CrCu2O4	0.95
TeLu2S4	0.975	SnCu2O4	0.9701	TcCu2O4	0.95
InFe2O4	0.9749	CrSc2O4	0.97	CoCr2Se4	0.949
SnNi2S4	0.9746	GaBr2O4	0.9697	MoNi2O4	0.948
CoLu2S4	0.9742	TeSc2O4	0.9692	TcFe2O4	0.942
NiFe2Se4	0.9738	NiSc2O4	0.9689	RhCu2O4	0.942
SnFe2S4	0.9735	FeBr2O4	0.9675	CuMn2Se4	0.941