Can machine learning find extraordinary materials?

Steven K. Kauwe, Jake Graser, Ryan Murdock, and Taylor D. Sparks

August 2019
Materials Science & Engineering, University of Utah, Salt Lake City, UT, 84112.
email: sparks@eng.utah.edu

One of the most common criticisms of machine learning is an assumed inability for models to extrapolate, *i.e.* to identify extraordinary materials with properties beyond those present in the training data set. To investigate whether this is indeed the case, this work takes advantage of density functional theory calculated properties (bulk modulus, shear modulus, thermal conductivity, thermal expansion, band gap, and Debye temperature) to investigate whether machine learning is truly capable of predicting materials with properties that extend beyond previously seen values. We refer to these materials as extraordinary, meaning they represent the top 1% of values in the available data set. Interestingly, we show that even when machine learning is trained on a fraction of the bottom 99% we can consistently identify $\frac{3}{4}$ of the highest performing compositions for all considered properties with a precision that is typically above 0.5. We explore model performance as the extrapolation distance is increased in various ways including, introduction of a gap, removal of certain elements, and removal of certain structure types. Moreover, we investigate a few different modeling choices and demonstrate how a classification approach can identify an equivalent amount of extraordinary compounds but with significantly fewer false positives than a regression approach. Finally, we discuss cautions and potential limitations in implementing such an approach to discover new record-breaking materials.

1 Introduction

Materials science has embraced the idea of data-driven research with consistent success [1, 2, 3, 4, 5, 6]. Although some research has focused on using experimentally derived properties from literature [7, 8, 9, 10], many researchers have also focused on predicting density functional theory (DFT) computed properties using a variety of different featurization schemes and learning approaches [11, 12, 13, 14, 15, 16, 17, 18].

A large portion of these publications cite the benefits of machine learning for screening purposes, as experts have long recognized that it is impossible to manually traverse chemical space [19, 20]. This direction of query has, however, invoked a very appropriate concern. Despite examples of success [4, 21, 22, 5], there is skepticism in the field [23] as to whether machine learning can truly find extraordinary materials from the overwhelming combinatorial complexity that arises from chemical space. This concern is one that lies in a fundamental assumption of many machine learning techniques, *i.e.*, that data is independent and identically distributed (i.i.d.) [24].

For a materials scientist, the i.i.d. assumption implies the training data fairly represents the full diversity of reality. This is clearly not the case due to data set bias. For example, some compounds are easier to synthesize and simulate or may be of more interest to researchers due to cost, performance in applications, or novelty. For this reason one must ask: do we have the information necessary, on a physical level, to even determine whether a material is extraordinary given highly clustered and largely "run-of-the-mill" materials in the data set? With this in mind, questions of the efficacy of machine learning for screening purposes are valid.

Disregarding the bias implicit in materials selection, this work seeks to establish whether machine learning systems have the potential to predict the most exceptional materials from existing data. As a best case scenario, we consider identifying extraordinary chemical compositions from a list of well-established DFT computed properties. The ability to complete this simplified screening task represents a minimum requirement if we are to seriously consider machine learning tools when screening for extraordinary materials. In this work we build upon recent publication by Xiong et al. exploring the degree to which extrapolation may or may not be possible for materials informatics [25]. We explore the performance of various machine learning algorithms for the task of predicting extraordinary materials. Using the Automatic-

FLOW for Materials Discovery (AFLOW) database [26], we are able to select compounds from a library which contains structures obtained from the Inorganic Crystal Structure Database (ICSD) [27]. With this data, we quantify the potential performance one might expect when seeking to identify extraordinary materials outside of the training set using a machine learning system. Furthermore, we show the results of applying a trained model to the Pearson Crystal Database (PCD) [28] and discuss the implications of the resulting predictions.

2 Methods

2.1 Featurizing the Data

The composition-based feature vector (CBFV) is a simple way of featurizing chemical compositions by performing mathematical operations on the element properties. Typically, the result of these operations are a set of unique vectors for each unique chemical composition. Using this approach, the CBFV can then be mapped to a target material property using various machine learning algorithms. This technique has been popular and fits well into a materials screening narrative [29, 30, 2]. In this work we use a featurization scheme (see supplemental code) that considers the average, range and variance of the element properties. The train and test data is also scaled via `scikit-learn`'s `StandardScaler` and `Normalizer` using the training data statistics.

2.2 Defining Extrapolation

The traditional notion of extrapolation loses meaning in the context of composition-space since elements are both discrete and finite. In the case of fractional compositions, all materials can be thought of as an interpolation between the pure elements. This is analogous to a ternary phase diagram with many regions which are not thermodynamically stable. A true example of extrapolation would be the ability to predict compounds featuring elements which are completely absent in the training data. Another definition would be the identification of the material compositions which have properties that extend beyond all values in the training data. To capture this notion of extrapolation, we operationally define extraordinary materials as any material in the top 1% of our data set in terms of their properties (see Figure 1).

In order to evaluate the extrapolation task it becomes necessary to quantify the *degree of extrapolation*, *i.e.* the distance between the training data and the materials labeled extraordinary. This can be done in three different ways:

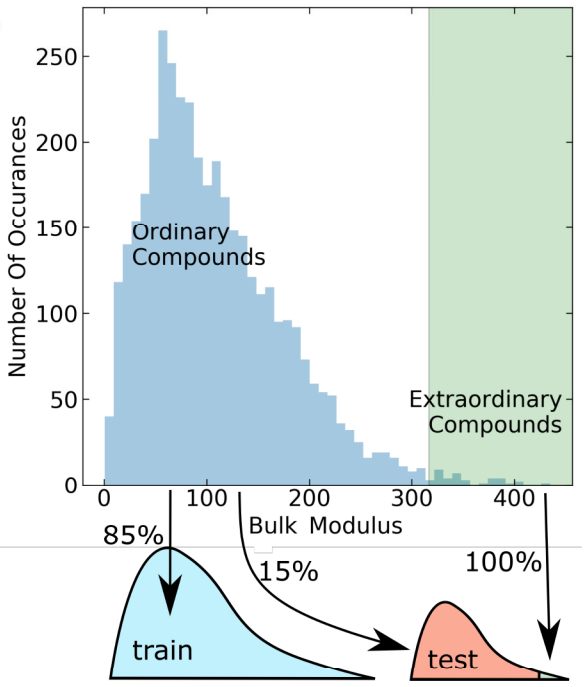

Figure 1: The distribution of bulk modulus data is separated into ordinary and extraordinary compounds. The train test split ensures that the top 1% is never trained on. 15% of the non-extraordinary (ordinary) data are included in the test set so we can see whether our model can successfully distinguish between extraordinary and ordinary values.

1. All of the training data is present except for those within an artificially defined "gap" area. This is done by withholding top 4%, 8%, and 12% of the data below the extraordinary threshold (See the gray cross-hatched region to the left of the vertical dashed line in Figure 2).

2. In addition to the gap area, we can also exclude from the training data compounds containing the most prevalent element from the data labeled "extraordinary" (See the yellow data points representing boron-containing structures when predicting shear modulus in Figure 2).

3. Alternatively, rather than excluding the most prevalent element, we can exclude the most prevalent "structure type" in the extraordinary data (See red data points representing space group 221 with 2 atoms per unit cell in Figure 2). Structure type is defined by taking the space group plus the number of atoms in the unit cell. This allows flexibility to find closely related entry prototypes within structure space.

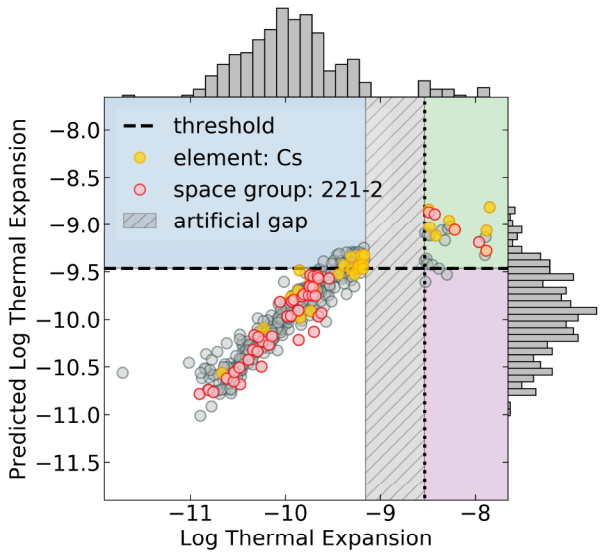

Figure 2: Actual versus predicted for logarithmic thermal expansion showing three ways of extrapolations: gap, withheld element, and withheld structure.

2.3 Data and Property Procurement

In order to evaluate the performance of our machine learning-based extrapolation task, we use data available from `AFLOWlib.org` [31] whose properties are calculated using structures from the ICSD library. The following properties were available and deemed appropriate in preparation for our study: bulk modulus, shear modulus, thermal conductivity, thermal expansion, band gap, and Debye temperature. In the case of duplicates, we sorted by ICSD number and kept the last entry. The data for these properties are available in a GitHub repository [32]. In the case of Debye temperature, thermal conductivity, shear modulus, and thermal expansion properties we scale the target by applying a base 10 logarithm to more closely match a normal distribution for learning purposes. To test the efficacy of our learning models, we use the train-test scheme shown in Figure 1. First, we isolate the top 1% of properties, label them as extraordinary, and add them to the test set. Next, we randomly sample 15% of the bottom 99%; these 15% are labeled as 'ordinary' and added to complete the test set. The training set is represented by the remaining data, and the highest 6% of this data is then assigned 'extraordinary' labels to match the same ratio of ordinary and extraordinary labels in the test set. For clarity's sake, the code to generate the ordinary/extraordinary data is available on GitHub.

We also obtained compositions corresponding to 156 421 unique measured structures in the PCD and a list of 10 590 computed elpasolite compounds [33]. After extracting the chemical compositions from these data sets, we featurize and scale them using the same steps as above. The resulting data is then input into our best screening tool, a trained classification model, to obtain probabilities which are then ranked to generate a list of interesting compositions to investigate.

2.4 Models and Performance Metrics

To investigate whether extrapolation is possible, we apply two linear and non-linear models. These are compared to nearest neighbor classification and regression models. For the task of predicting compound values we use a ridge regression and a support vector regression with a radial basis function (rbf) kernel. We also approach the problem as a classification task using a logistic regression and a support vector classification (rbf). For simplicity we employ the `scikit-learn` [34] implementation of these models and optimize parameters using grid search techniques (see GitHub for details of the implementation). To assess performance we use the classification metrics of precision and recall, which are defined in equations 1 and 2 where tp, tn, fp, fn are true positive, true negative, false positive, and false negative, respectively. Precision is an effective metric to determine how often our predictions result in extraordinary compounds. Recall is a metric used to determine what fraction of extraordinary compounds are correctly identified by those predictions.

$$precision = \frac{tp}{tp + fp} \tag{1}$$

$$recall = \frac{tp}{tp + fn} \tag{2}$$

The use of these metrics requires that some threshold value be established from which a label "extraordinary" or "ordinary" can be assigned. It is natural to use the default threshold of 0.5 for a classifier. For the regression models, we select these thresholds by optimizing the F1 score on the training data. The F1 score represents the harmonic mean of precision and recall:

$$F1 = 2\left(\frac{precision \cdot recall}{precision + recall}\right) \tag{3}$$

Although optimization on the F1 represents a good compromise between the two metrics, one could favor precision or recall when choosing the threshold in practice. In addition to these performance metrics, it is also possible to define threshold independent metrics such as the area under the precision-recall curve. This has the added benefit of being robust to imbalanced data.

3 Results & Discussion

The capability of machine learning to identify extraordinary materials is tested for the following properties: bulk modulus, shear modulus, thermal expansion, thermal conductivity, band gap and Debye temperature. The regression and classification of these properties seeks to optimize different loss functions. Because of this, they each have their place for use as screening tools. In this work, we are particularly interested in materials discovery, or a model's ability to extrapolate to extraordinary materials and consider the ability of both methods for this application.

3.1 Screening as a Regression Task

The ability to effectively train a regression model is dependent on diverse data and a reasonable distribution of values. We demonstrate the ability to extrapolate to a majority of extraordinary materials for each property. Figure 3a clearly shows the ability of a ridge regressor to identify extraordinary compounds for the property of bulk modulus. (Remaining property figures are available with the supplementary code.) The threshold values in this figure are obtained from the training data and used for generating classification metrics. This performance is representative of all properties tested.

3.2 Screening as a Classification Task

Classification is a great alternative to regression if one is only interested in identifying extraordinary materials. The trade off when using classification is that the property value of extraordinary materials will not be predicted. Additionally, as a real task, the process of choosing how many true and false labels to use is unclear; a balance must be struck between precision and recall. The performance of the logistic regression is visualized in Figure 3b for the same bulk modulus data. In our implementation, classification is almost always superior to the regression task, as demonstrated by a consistently higher precision and a nearly equivalent recall.

Moreover, classification does not depend on the distribution of the data. However, the choice of threshold for the training set will drastically affect how aggressive the classifier is. If too few data are labeled as extraordinary the model will fail to generate sufficient labels to represent the test set. If too many are labeled as extraordinary the model is susceptible to very high false positive rates, and therefore low precision scores.

3.3 Quantifying Extrapolation

The results of this work show that in principle, extrapolation via both classification and regression-based machine learning approaches should be able to identify and predict extraordinary materials with properties beyond those present in the training data. Although there are no previously defined values that might constitute "adequate extrapolation performance," it is clear that the non-naïve models demonstrate the ability to identify a large fraction of the compositions labeled as extraordinary.

This outcome provides a positive response to the persistent question of whether or not materials informatics approaches will be able to identify new compounds with exceptional properties. Does this research suggest that extrapolative studies for materials discovery will be useful or even transformative? To answer this question we need to consider four scenarios.

Scenario 1: Identifying extraordinary compounds by random guessing. In our original data set we set aside the top 1% as extraordinary and all of these compounds were added to a mixture of 15% of the lower 99% of ordinary compounds. Therefore, random guessing would yield a $1/15 = 6\%$ success rate in classifying extraordinary compounds.

Scenario 2: Relying on chemical intuition. Researchers almost never rely on random guessing when screening for candidate materials. A researcher will have a degree of domain expertise resulting from training as well as their ability to uptake information from the published literature. However, a researcher will also be impacted and biased by their "chemical intuition" drawn from prior experience. Despite *Scenario 2* being the *modus operandi*, the Materials Genome Initiative [35] has urged the research community to reconsider this approach. Not only is the outcome of *Scenario 2* highly variable and subjective, but history has shown that this approach is slow and leads to local optimization. Rather than exploring the breadth of chemical whitespace, researchers have focused on clusters of known chemistries and structures with primarily minor elemental substitutions.

Scenario 3: A nearest-neighbor approach. Humans approaching the task of extrapolation will naturally identify extraordinary materials based on composition and operate under the assumption that similar chemistries might have similar properties. For example, if main group borides and carbides represent superhard materials, perhaps borocarbides mixed with transition metals would not be unreasonable as superhard candidates. If scenario 2 relies on human intuition, scenario 3 is our attempt to automate this intuition. For any unknown compound, we can assign it the same property value as its closest neighbor. This is done by encoding composition into vectors and

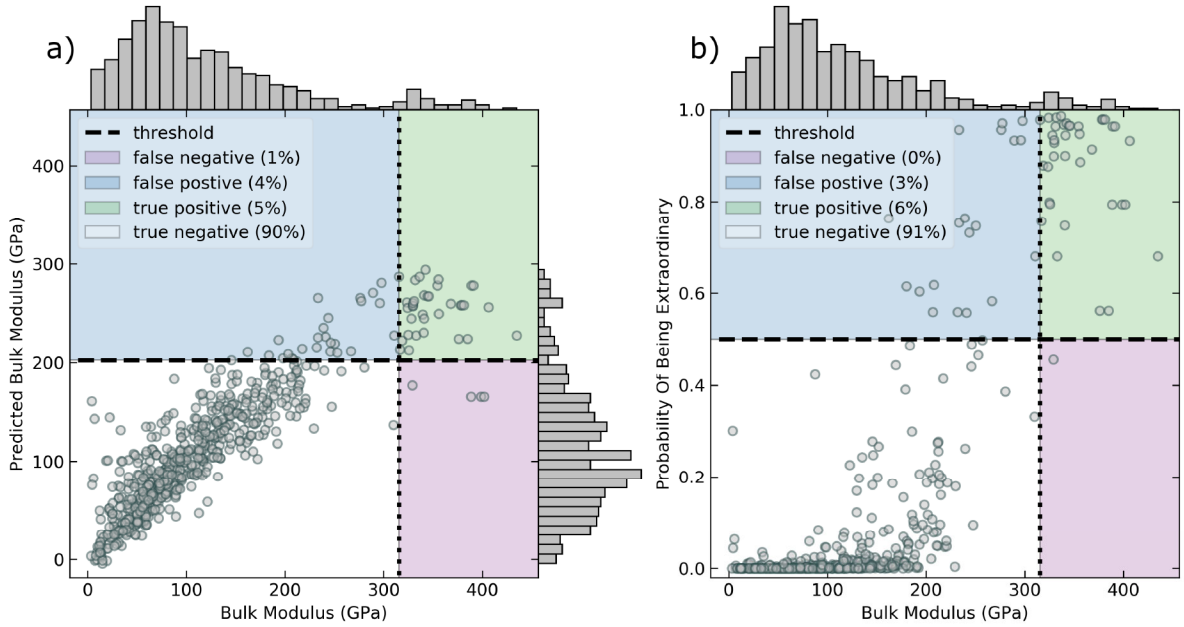

Figure 3: *(a)* The actual vs. predicted values for a ridge regression trained on bulk modulus. *(b)* The actual values vs. the probability, obtained from a logistic regression, of being extraordinary. All data in the training set has values lower than 300 GPa. Data to the right of the vertical line represent extraordinary compositions.

finding pairs with the smallest euclidean distance.

Scenario 4: Machine learning-based predictions of benchmark materials combined with domain knowledge and chemical intuition. Researchers seeking to identify global optima in extraordinary materials will probably benefit from this approach. Our research shows that, depending on the property being predicted, models typically exhibit a precision of ∼0.5. In other words, every other compound suggested would be extraordinary! In the subsequent section, we explain why this will likely have reduced efficacy. However, if true, this represents a dramatic breakthrough for materials discovery far superseding even the most optimistic outcomes of *Scenario 1, 2, or 3*. Moreover, researchers could look at the list of candidate extraordinary materials and then focus specifically on examples that fall outside of typically studied chemistries or crystal structures in an attempt to identify globally optimal extraordinary compounds. Finally, the recall values are typically 0.75 or greater for all of the properties we predicted. Therefore, in a best case scenario (where i.i.d. holds true), this work suggests an ability to identify a new benchmark material on average every other attempt while capturing the vast majority of possible extraordinary compounds. Though there are few corroborative examples of this degree of success.

In comparing scenario 3 and 4, we can look at the average performance across all properties. For these predictions we compare regression, classification, and nearest neighbor approaches. All approaches outperform random guessing, nevertheless, a naïve nearest neighbor approach is typically inferior to our ridge regression and logistic regression. The discrepancy between the nearest-neighbor and other approaches only increases as we extend the extrapolation task by increasing gap distance or excluding prevalent elements. Removing structure types has a negligible effect likely because the space group plus number of elements in unit cell approach is inadequate to describe structures. Indeed, structure featurization is an ongoing research area in materials informatics. Surprisingly, introduction of a gap has minor impact on the classification when all elements are present, but *improves* classification performance when prevalent elements are removed. This is likely because those compounds near the ordinary-extraordinary boundary have the highest probability of being mislabeled.

3.4 Limitations and Cautions

The precision and recall of the models we report here suggest exciting avenues for discovering new record-breaking materials. However, we caution that this approach will have fundamental limitations which suggest a more wary optimism.

The first limitation has to do with the mechanisms associated with different properties. If a researcher is looking to discover a new composition which achieves a record-breaking property and does so with the same

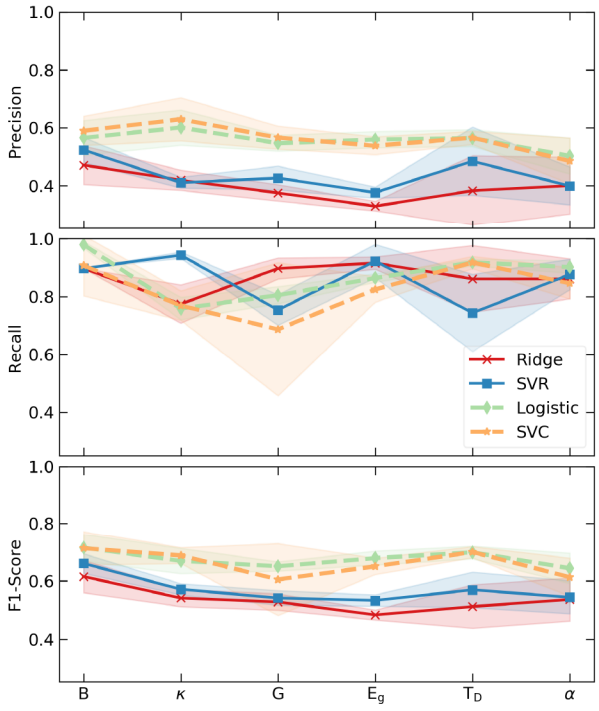

Figure 4: All machine learning models are able to successfully identify extraordinary compositions from the test set. However, the classification models, shown with dotted lines, also have the advantage of significantly fewer false positives. Standard deviations (shown via shading) were generated by assigning the bottom 99% train-test split using 5 different random seeds. The properties bulk modulus, thermal conductivity, shear modulus, band gap, Debye temperature and thermal expansion are represented by their symbols B, κ, G, E_g, T_D, α.

fundamental mechanism present in the training data compositions, they will likely be successful. On the other hand, if a researcher is seeking to discover a record-breaking material which achieves its extraordinary properties by leveraging a new mechanism, not common in the training data, this will be unlikely. Without numerous examples to train from, this extrapolation approach is unlikely to yield new physics and mechanistic insight. For example, given many compounds exhibiting Bardeen-Cooper-Schrieffer superconductivity to train from, it is very unlikely that cuprate oxides would have been identified as extraordinary candidates for high-T_c superconductivity since these operate on a completely different (and yet unexplained) mechanism of superconductivity[36]. At the same time, careful examination of poor predictions in the training data set could lead to physical insight into new mechanisms. For example, if a specific chemistry or class of material are consistently poorly predicted despite sufficient training data then a researcher could postulate that their descriptors are simply not capturing the unique physics operating in that chemistry and could therefore be investigated in more detail [37].

A second limitation centers around the critical i.i.d. assumption in materials data used for materials informatics. Despite emerging efforts from Citrine Informatics [38], the Materials Data Facility (MDF) [39], and others, materials science as a discipline is still lacking robust data repositories for many properties of interest. Additionally, even where data is available there exist challenges with data heterogeneity, inherent error in measurement or calculation of materials properties, imbalanced classes, sparsity, bias towards high performing materials, and more [40, 41, 42]. Models are truly only as good as the data available to train from. Or, as Charles Babbage [43] put it:

> "On two occasions I have been asked, 'Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?'...I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question."

Inherent to the issue of data sampling is the way in which a researcher chooses to generate a list of potential candidates to screen over. In our setup, this is trivial. However, an honest attempt to screen for extraordinary properties must confront this in a reasonable way. As a most simple approach, researchers could use a database of known compounds. We demonstrate this method by utilizing the Pearson Crystal Database as a list of potential candidates to ensure that predicted compositions are viable on a physical level. An alternative approach may be to generate a list of possible compositions and to screen these fictional compositions similarly to the elpasolite work of Faber et al [33]. One such list could then be taken and an attempt to simulate or synthesize materials from the given composition could be made in order to validate the results. However, regardless of selection, one cannot be sure that the resulting data is amenable to learning from the trained model.

We see evidence of this limitation when we compare the compositions of the extraordinary bulk modulus materials predictions generated from the PCD vs those top 1% that were assigned from the AFLOW repository data. The ranked elemental prevalence of identified extraordinary compounds shows that while nearly all the same elements are present, the proportion and ratio of these is highly variable suggesting a lack of parity between training data and those compounds which have been synthesized. This is further highlighted by the fact less than 0.1% of compositions from the PCD are predicted as extraordinary

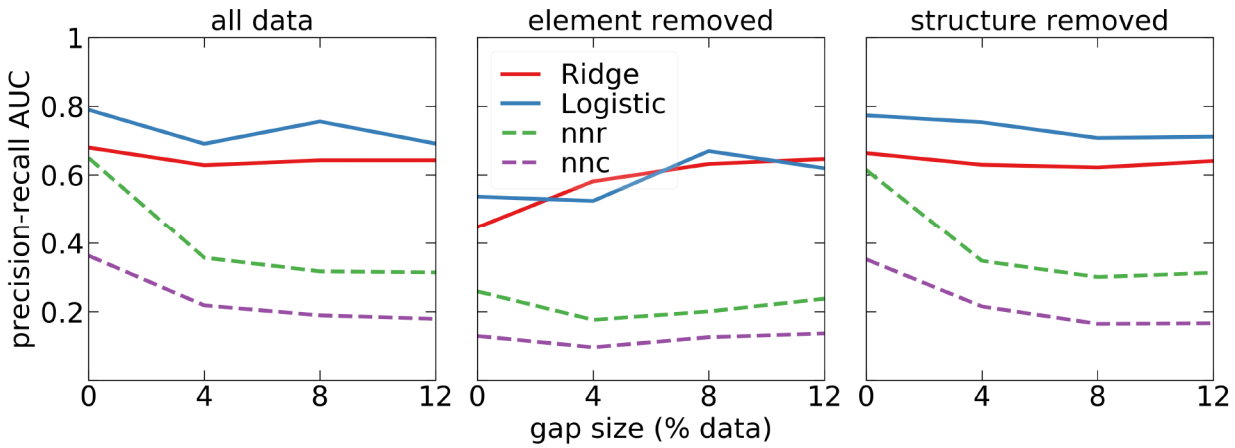

Figure 5: The precision-recall Area-under-curve (AUC) is computed as the average across all properties. The three extrapolation cases are considered for all gap sizes using four different model types, ridge regression (Ridge), logistic regression (Logistic), nearest-neighbor regression (nnr), and nearest-neighbor classification (nnc). *(left)* The entire training data is used (except the gap), *(center)* the most prevalent element is removed from training, and *(right)* the most prevalent structure is removed from training.

while the the top 1% of training data were labeled as extraordinary for model generation. The disparity between the AFLOW data and the elpasolite data could not even be evaluated as the model failed to label a single composition as extraordinary.

Additional shortcomings in this approach are that the input for the prediction is the chemical formula for an individual compound, but many of the most important materials are actually composite mixtures of phases which synergistically produce a desired outcome. For example, ductile ferrite and brittle cementite in steels or precipitate hardened aluminum alloys. To our knowledge, there are not yet examples in the materials informatics literature where authors make predictions of a composite property by training a machine learning model on each individual phase using a structure or composition-based feature vector.

A related problem is associated with rare events such as doping where a few percent elemental substitution can lead to drastic changes in properties due to complicated defect chemistry. For instance, doping silicon with phosphorus from 10^{12} cm^{-3} ($\sim 0\%$) up to 10^{21} cm^{-3} ($\sim 2\%$) is accompanied by a change of electrical conductivity approximately eight orders of magnitude [44]!

It is not that machine learning is fundamentally incapable of modeling composite materials or the effect of rare events like doping. However, in order to capture these effects it will be necessary to have training data which includes these phase mixtures and dopant compositions with sufficient granularity to train models on their effect. Put plainly, a database of stoichiometric compounds will not be able to predict the influence of doping, but rather, it will require a database where many slight dopant compositions are reported with an associated material property.

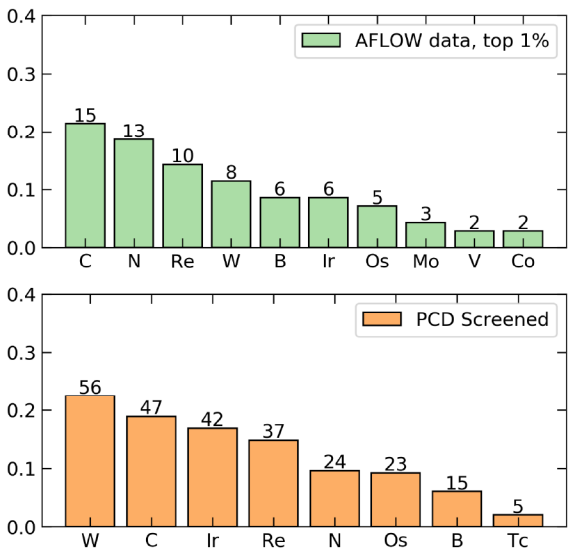

Figure 6: The ranked elemental prevalence of compounds labeled extraordinary in the original AFLOW data versus those screened from the PCD.

4 Conclusion

The predictive power of machine learning is established in the field of materials science. Researchers have demonstrated many different models that can effectively map chemical compositions to material properties. One of the most promising aspects of this is

the use of machine learning predictions for screening materials. Unfortunately, many of these models start with fundamental assumptions that would imply an inability to extrapolate into unique and interesting chemical species. For this reason it is natural to question whether we should be using machine learning to screen for high performance materials.

We show that, in a best case scenario—DFT data with a large variety of compounds and with well-distributed properties—machine learning is consistently able to identify a significant portion (average recall$\sim$0.6) of the top 1% of materials using both classification and regression approaches. This approach is successful across six diverse materials properties. Moreover, we show that the classification-based approach identifies a near equivalent amount of extraordinary compositions while returning fewer false positives than both nearest-neighbor and regression approaches.

Introducing a gap between the data available for training and the extraordinary compounds has a relatively minor impact on both precision and recall. In fact, classification can slightly improve as the gap is introduced since this removed compounds most subject to ambiguity. On the other hand, withholding the most prevalent element among those compounds labeled extraordinary leads to slight reductions in the classification metrics. That said, the values still strongly suggest extrapolation is possible in this dataset.

Overall, this work demonstrates promise in using machine learning models to facilitate the discovery of record-breaking materials. Unfortunately, we also show that predictions made on the PCD fail to break into new or unexpected compositions in the case of bulk modulus. Despite the inability to find new chemistries, the models do a great job of identifying the extraordinary compositions that are included in the dataset we used.

5 Acknowledgements

The authors gratefully acknowledge support from NSF CAREER Award DMR 1651668. Special thanks to Anthony Yu-Tung Wang for his assistance and insight when preparing the manuscript. The authors also would like to acknowledge and recognize the excellent reviewer comments we received which greatly improved the manuscript.

References

[1] A. O. Oliynyk, E. Antono, T. D. Sparks, L. Ghadbeigi, M. W. Gaultois, B. Meredig, and A. Mar, "High-throughput machine-learning-driven synthesis of full-heusler compounds," *Chemistry of Materials*, vol. 28, no. 20, pp. 7324–7331, 2016.

[2] M. W. Gaultois, A. O. Oliynyk, A. Mar, T. D. Sparks, G. J. Mulholland, and B. Meredig, "Perspective: Web-based machine learning models for real-time screening of thermoelectric materials properties," *APL Materials*, vol. 4, no. 5, p. 053213, 2016.

[3] R. Liu, A. Kumar, Z. Chen, A. Agrawal, V. Sundararaghavan, and A. Choudhary, "A predictive machine learning approach for microstructure optimization and materials design," *Scientific reports*, vol. 5, p. 11551, 2015.

[4] G. Hautier, C. C. Fischer, A. Jain, T. Mueller, and G. Ceder, "Finding nature's missing ternary oxide compounds using machine learning and density functional theory," *Chemistry of Materials*, vol. 22, no. 12, pp. 3762–3767, 2010.

[5] A. Mansouri Tehrani, A. O. Oliynyk, M. Parry, Z. Rizvi, S. Couper, F. Lin, L. Miyagi, T. D. Sparks, and J. Brgoch, "Machine learning directed search for ultraincompressible, superhard materials," *Journal of the American Chemical Society*, vol. 140, no. 31, pp. 9844–9853, 2018.

[6] T. D. Sparks, M. W. Gaultois, A. Oliynyk, J. Brgoch, and B. Meredig, "Data mining our way to the next generation of thermoelectrics," *Scripta Materialia*, vol. 111, pp. 10–15, 2016.

[7] Y. Zhuo, A. Mansouri Tehrani, and J. Brgoch, "Predicting the band gaps of inorganic solids by machine learning," *The journal of physical chemistry letters*, vol. 9, no. 7, pp. 1668–1673, 2018.

[8] P. Raccuglia, K. C. Elbert, P. D. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, and A. J. Norquist, "Machine-learning-assisted materials discovery using failed experiments," *Nature*, vol. 533, no. 7601, p. 73, 2016.

[9] S. K. Kauwe, J. Graser, A. Vazquez, and T. D. Sparks, "Machine learning prediction of heat capacity for solid inorganics," *Integrating Materials and Manufacturing Innovation*, vol. 7, no. 2, pp. 43–51, 2018.

[10] J. Graser, S. K. Kauwe, and T. D. Sparks, "Machine learning and energy minimization approaches for crystal structure predictions: A review and new horizons," *Chemistry of Materials*, vol. 30, no. 11, pp. 3601–3612, 2018.

[11] G. Pilania, A. Mannodi-Kanakkithodi, B. Uberuaga, R. Ramprasad, J. Gubernatis,

and T. Lookman, "Machine learning bandgaps of double perovskites," *Scientific reports*, vol. 6, p. 19375, 2016.

[12] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, and A. Tropsha, "Universal fragment descriptors for predicting properties of inorganic crystals," *Nature communications*, vol. 8, p. 15679, 2017.

[13] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, "Crystal structure representations for machine learning models of formation energies," *International Journal of Quantum Chemistry*, vol. 115, no. 16, pp. 1094–1101, 2015.

[14] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, and A. Jain, "Unsupervised word embeddings capture latent knowledge from materials science literature," *Nature*, vol. 571, no. 7763, p. 95, 2019.

[15] Q. Zhou, P. Tang, S. Liu, J. Pan, Q. Yan, and S.-C. Zhang, "Learning atoms for materials discovery," *Proceedings of the National Academy of Sciences*, vol. 115, no. 28, pp. E6411–E6417, 2018.

[16] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, "A general-purpose machine learning framework for predicting properties of inorganic materials," *npj Computational Materials*, vol. 2, p. 16028, 2016.

[17] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "Schnet–a deep learning architecture for molecules and materials," *The Journal of Chemical Physics*, vol. 148, no. 24, p. 241722, 2018.

[18] T. Xie and J. C. Grossman, "Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties," *Physical review letters*, vol. 120, no. 14, p. 145301, 2018.

[19] J.-L. Reymond and M. Awale, "Exploring chemical space for drug discovery using the chemical universe database," *ACS chemical neuroscience*, vol. 3, no. 9, pp. 649–657, 2012.

[20] P. Kirkpatrick and C. Ellis, "Chemical space," 2004.

[21] Y. Iwasaki, I. Takeuchi, V. Stanev, A. G. Kusne, M. Ishida, A. Kirihara, K. Ihara, R. Sawada, K. Terashima, H. Someya, *et al.*, "Machine-learning guided discovery of a new thermoelectric material," *Scientific reports*, vol. 9, no. 1, p. 2751, 2019.

[22] J. H. Martin, B. D. Yahata, J. M. Hundley, J. A. Mayer, T. A. Schaedler, and T. M. Pollock, "3d printing of high-strength aluminium alloys," *Nature*, vol. 549, no. 7672, p. 365, 2017.

[23] H. Bhadeshia, "Neural networks and information in materials science," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 1, no. 5, pp. 296–305, 2009.

[24] K. Jacobs, "Independent identically distributed (iid) random variables," in *Discrete Stochastics*, pp. 65–101, Springer, 1992.

[25] Z. Xiong, Y. Cui, Z. Liu, Y. Zhao, M. Hu, and J. Hu, "Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation," *Computational Materials Science*, vol. 171, p. 109203, 2020.

[26] Aflow, "Aflow - automatic - flow for materials discovery," 2018. [Online; accessed 14-July-2019].

[27] A. Belsky, M. Hellenbrandt, V. L. Karen, and P. Luksch, "New developments in the inorganic crystal structure database (icsd): accessibility in support of materials research and design," *Acta Crystallographica Section B: Structural Science*, vol. 58, no. 3, pp. 364–369, 2002.

[28] P. Villars and K. Cenzual, "Pearson's crystal data: Crystal structure database for inorganic compounds (on dvd) 2.0," 2018.

[29] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, "Combinatorial screening for new materials in unconstrained composition space with machine learning," *Physical Review B*, vol. 89, no. 9, p. 094104, 2014.

[30] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, "Machine learning in materials informatics: recent applications and prospects," *npj Computational Materials*, vol. 3, no. 1, p. 54, 2017.

[31] S. Curtarolo, W. Setyawan, G. L. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, *et al.*, "Aflow: an automatic framework for high-throughput materials discovery," *Computational Materials Science*, vol. 58, pp. 218–226, 2012.

[32] S. K. Kauwe and R. Murdock, "`https://github.com/kaaiian/can_machine_learning_find_extraordinary_materials`," Aug 2019.

[33] F. A. Faber, A. Lindmaa, O. A. Von Lilienfeld, and R. Armiento, "Machine learning energies of 2 million elpasolite (a b c 2 d 6) crystals," *Physical review letters*, vol. 117, no. 13, p. 135502, 2016.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[35] J. P. Holdren, "Materials genome initiative for global competitiveness," 2011-06-23.

[36] B. Meredig, E. Antono, C. Church, M. Hutchinson, J. Ling, S. Paradiso, B. Blaiszik, I. Foster, B. Gibbons, J. Hattrick-Simpers, *et al.*, "Can machine learning identify the next high-temperature superconductor? examining extrapolation performance for materials discovery," *Molecular Systems Design & Engineering*, vol. 3, no. 5, pp. 819–825, 2018.

[37] F. Belviso, V. E. Claerbout, A. Comas-Vives, N. S. Dalal, F.-R. Fan, A. Filippetti, V. Fiorentini, L. Foppa, C. Franchini, B. Geisler, *et al.*, "Atomic-scale design protocols toward energy, electronic, catalysis, and sensing applications," 2019.

[38] J. O'Mara, B. Meredig, and K. Michel, "Materials data infrastructure: a case study of the citrination platform to examine data import, storage, and access," *JOM*, vol. 68, no. 8, pp. 2031–2034, 2016.

[39] B. Blaiszik, K. Chard, J. Pruyne, R. Ananthakrishnan, S. Tuecke, and I. Foster, "The materials data facility: Data services to advance materials science research," *JOM*, vol. 68, no. 8, pp. 2045–2052, 2016.

[40] R. Seshadri and T. D. Sparks, "Perspective: interactive material property databases through aggregation of literature data," *APL Materials*, vol. 4, no. 5, p. 053206, 2016.

[41] S. K. Kauwe, T. D. Rhone, and T. D. Sparks, "Data-driven studies of li-ion-battery materials," *Crystals*, vol. 9, no. 1, p. 54, 2019.

[42] S. Kauwe and T. Welker, "Extracting knowledge from dft," 2018.

[43] C. Babbage, *Passages from the Life of a Philosopher*. Cambridge University Press, 2011.

[44] W. R. Thurber, *The relationship between resistivity and dopant density for phosphorus-and boron-doped silicon*, vol. 400. US Department of Commerce, National Bureau of Standards, 1981.