npj computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

A critical examination of robustness and generalizability of machine learning prediction of materials properties

Kangming Li[1]✉, Brian DeCost[2], Kamal Choudhary[2,3], Michael Greenwood[4] and Jason Hattrick-Simpers[1]

Recent advances in machine learning (ML) have led to substantial performance improvement in material database benchmarks, but an excellent benchmark score may not imply good generalization performance. Here we show that ML models trained on Materials Project 2018 can have severely degraded performance on new compounds in Materials Project 2021 due to the distribution shift. We discuss how to foresee the issue with a few simple tools. Firstly, the uniform manifold approximation and projection (UMAP) can be used to investigate the relation between the training and test data within the feature space. Secondly, the disagreement between multiple ML models on the test data can illuminate out-of-distribution samples. We demonstrate that the UMAP-guided and query by committee acquisition strategies can greatly improve prediction accuracy by adding only 1% of the test data. We believe this work provides valuable insights for building databases and models that enable better robustness and generalizability.

npj Computational Materials (2023)9:55 ; https://doi.org/10.1038/s41524-023-01012-9

INTRODUCTION

The use of machine learning (ML) has been increasingly popular in the materials science community[1–11]. Central to the training of machine learning models is the need for findable, accessible, interoperable, and reusable (F.A.I.R.)[12] materials science datasets. High-throughput density functional theory (DFT) calculations have proven to be an efficient and reliable way to generate materials property data, screen the target materials space and accelerate materials discovery[13–17]. Concentrated community efforts have led to the curation of large DFT databases for various materials properties, e.g., Materials Project[18], Automatic FLOW for Materials Discovery[19], Open Quantum Materials Database[14], and JARVIS-DFT[20]. The availability of large materials databases has fueled the development and application of machine learning methods based on a chemical formula or atomic structures, including traditional ML models with preselected feature sets[21–29] and neural networks with automatic feature extraction[30–41].

The continuously improved performance of ML models in the DFT database benchmarks shows the great potential of using these models as the surrogate of computationally expensive DFT calculations to explore unknown materials[37]. However, there are reasons to remain cautious, particularly for the generalization performance of the trained ML models[42]. First, the current DFT databases still cover only a very limited region of the potential materials space[43,44]. Some databases may be the results of mission-driven calculations and therefore be more focused on certain types of materials or structural archetypes, leading to biased distributions[45–48]. In addition, data distributions may shift even between different versions of an actively expanding database, due to a change in their focus with time. While the common practice is to train and validate ML models on the latest databases, we are unaware of any systematic study examining whether these models can predict reasonably (or at least qualitatively) well the properties of new materials added in the future database versions. Such an examination is critical for assessing the maturity of a database (namely, whether it is sufficiently representative of the materials space) and the robustness of the resulting ML prediction, both of which are essential for building trust in the use of these ML models.

Since the current databases may not yet offer an unbiased and sufficiently rich representation of the potential materials space, the performance scores of an ML model evaluated from a random train-validation-test split may be an optimistic estimate of the true generalization performance[49–51]. While the latter may be estimated more properly from grouped cross-validations (CV)[52–54], finding a well-defined method for grouping data is not always trivial and depends on the preselected input features, which may not be the optimal way to find the most physically relevant grouping[37]. On the other hand, one may consider it safer to limit the use of an ML model to its applicability domain, or the interpolation region[48]. However, in high-dimensional compositional-structural feature space, as is encountered in materials science, it is challenging to properly define an interpolation region and to determine when the model is extrapolating.

In this work, we highlight the limitations of the current ML methods in materials science for predicting out-of-distribution samples, by showing that ML models pretrained on the Materials Project[18] 2018 database have unexpectedly acute performance degradation on the latest database. Such performance degradation can occur in the deployment stage of any ML model and degrades community trust in their validity. Therefore, we also provide solutions for diagnosing, foreseeing, and addressing the issue, and discuss ways to improve prediction robustness and generalizability.

The paper is organized as follows. First, we examine the performance of a state-of-the-art neural network, with a comparison to traditional ML models. Next, we analyze the observed performance degradation in terms of the dataset's feature space. We then discuss different methods based on the dataset's representation and model predictions to foresee the generalization issue. Finally, we propose ways to improve prediction robustness for materials exploration.

[1]Department of Materials Science and Engineering, University of Toronto, 27 King's College Cir, Toronto, ON, Canada. [2]Material Measurement Laboratory, National Institute of Standards and Technology, 100 Bureau Dr, Gaithersburg, MD, USA. [3]Theiss Research, La Jolla, CA 92037, USA. [4]Canmet MATERIALS, Natural Resources Canada, 183 Longwood Road south, Hamilton, ON, Canada. ✉email: kangming.li@utoronto.ca

RESULTS

Failure to generalize in new regions of materials space

Formation energy (E_f) is a fundamental property that dictates the phase stability of a material. Formation energy prediction is a basic task for ML models used in materials science, including traditional descriptor-based models[26–28,55] and neural networks[31–34,36]. Among them, graph neural network (GNN) models with atomistic structures as inputs are currently considered to have state-of-the-art performance[7]. Here we consider the Atomistic LIne Graph Neural Network (ALIGNN), an architecture that performs message passing on both the interatomic bond graph and its line graph corresponding to bond angles[34]. The ALIGNN model shows the best performance in predicting the Materials Project[18] formation energy according to the Matbench[37] leader-board; we, therefore, choose it as the representative GNN model for the subsequent performance evaluation.

We use the ALIGNN model pretrained on the Materials Project 2018.06.01 version (denoted as MP18), which contains 69239 materials and has been used for benchmarking GNN models in the recent papers[32–34]. In the original ALIGNN paper, a 60000-5000-4239 train-validation-test split of the MP18 dataset was used, achieving a mean absolute error (MAE) of 0.022 eV/atom for the test set[34].

We use the MP18-pretrained ALIGNN model (ALIGNN-MP18) to predict the formation energies of the new structures in the latest (2021.11.10 version) Materials Project database (denoted as MP21). Instead of testing on the whole MP21 dataset, we consider the scenario where we want to apply ML models to explore a particular material subspace of interest. In this work, we define the alloys of interest (AoI) as the space formed by the first 34 metallic elements (from Li to Ba) and the alloys formed exclusively by these elements. This AoI materials space is defined to include the most common components for high-entropy alloys, a class of alloys that has recently drawn much attention thanks to its superior performance compared to traditional alloys[56]. In the MP21 dataset, there are 7800 AoI, 2261 (or 29%) of which already appear in the MP18 dataset, while the rest are not contained within MP18. Therefore, we consider those 2261 alloys as the AoI in the training set, and the rest that appear only in the MP21 as the AoI in the test set.

A list of important acronyms used in this work is given in Table 1. A description of the MP18 dataset, and the AoI data is given in Table 2. We note that the mean absolute deviation (MAD) and the standard deviation (STD) of the data correspond to the mean absolute error (MAE) and the root mean square error (RMSE) of a baseline model whose prediction for every structure is equal to the mean of the training data.

Figure 1 shows the ALIGNN-MP18 performance on the formation energy predictions of the AoI. For the AoI in the training set, the ALIGNN-MP18 predictions agree well with the DFT values, with an MAE of 0.013 eV/atom. For the AoI test data, while there is still a reasonable agreement for the structures with E_f^{DFT} below 0.5 eV/atom, the ALIGNN-MP18 model strongly underestimates the formation energies for a significant portion of the structures whose E_f^{DFT} are above 0.5 eV/atom. In the latter case, the prediction errors range from 0.5 eV/atom to up to 3.5 eV/atom, which is 23 to 160 times larger than the MP18 test MAE of 0.022 eV/atom. Indeed, the prediction errors are nearly as large as E_f^{DFT} for those alloys, indicating that the ALIGNN-MP18 predictions fail to even qualitatively match the DFT formation energies. For reference, the MAE and the coefficient of determination (R^2 score) for the AoI test set are 0.297 eV/atom and 0.194, respectively (Table 3).

It can be seen from Fig. 1 that the ALIGNN-MP18 predictions are largely restricted to the value range below 1 eV/atom. Indeed, despite a large formation energy range (from −4.3 to 4.4 eV/atom)

Table 1. List of important acronyms used in this work.

Acronyms	Description
AoI	Alloys of interest are formed by one or more of the first 34 metallic elements.
MP18 (MP21) AoI	AoI in Materials Project 2018 (2021) database.
AoI in train set	AoI in MP18.
AoI in test set	AoI in MP21 but not in MP18.
XGB, RF, LF	XGBoost, random forest, linear-forest.
SG-X	Space group X, e.g., SG-71 for space group 71.
UMAP	Uniform Manifold Approximation and Projection.
QBC	Query by committee.

Table 2. Description of the MP18 data and the AoI data in the MP18 and MP21 datasets.

	Entries	Min	Max	MAD	STD
MP18	69239	−4.522	4.389	0.926	1.072
MP18 AoI	2261	−1.090	1.575	0.230	0.313
MP21 AoI	7800	−1.090	4.416	0.440	0.751

The number of entries, the minimum, maximum, mean absolute deviation and standard deviation of formation energies (in eV/atom) are given.

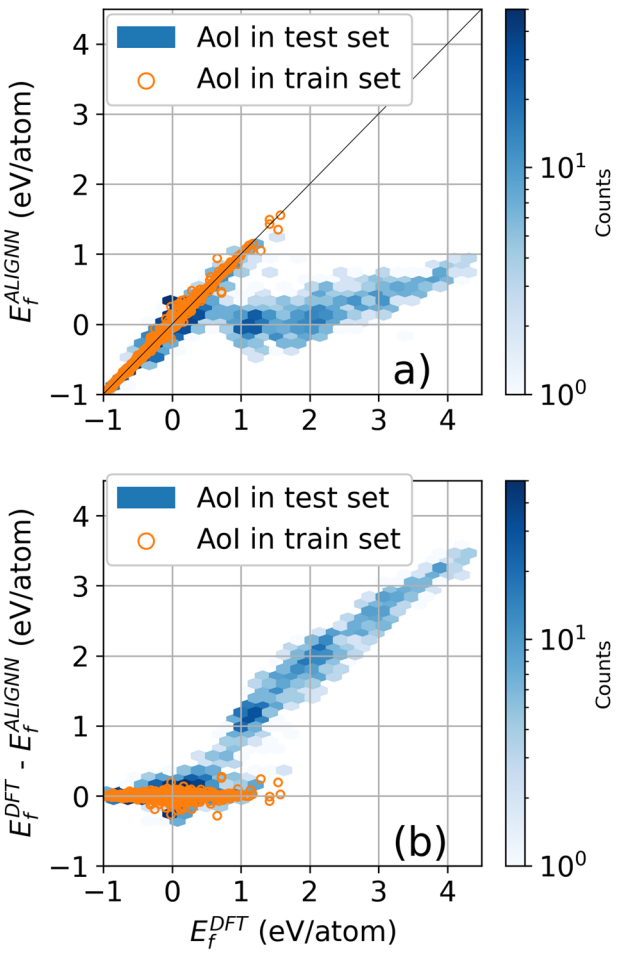

Fig. 1 Performance of the ALIGNN-MP18 model. a Parity plot and **b** prediction errors of the ALIGNN-MP18 model.

of the whole MP18 dataset, most of the formation energies of the AoI in the MP18 lie between −1 to 1 eV/atom. Therefore, it is not surprising that the ALIGNN-MP18 predictions are limited by the range of the formation energies of the AoI training set. However, it is unexpected to observe that the strong underestimation by the ALIGNN-MP18 model already occurs in the formation energy range of 0.5 to 1 eV/atom. For alloys with formation energies above 1 eV/atom, the ALIGNN-MP18 model predicts values that are well below the upper bound of formation energies in the training set, some of which are even negative. Consequently, the test set performance issue of the ALIGNN-MP18 model cannot be explained by the bounded energy range of the AoI in the training set. The origin of the issue will be discussed in the next section.

To verify whether the performance issue is common to other ML models, we perform the same training and test procedures with traditional descriptor-based ML models. To do so, we first use Matminer[26–28] to extract 273 features based on compositions and structures for the whole MP18 dataset and the alloys in MP21. Then, we down select features by sequentially dropping highly correlated features using a Pearson's R of 0.7 as the threshold, reducing the final number of features to 90. These 90 features are used for subsequent traditional ML model training and other analysis throughout this work.

Here we consider three traditional regression models: the gradient-boosted trees as implemented in `XGBoost` (XGB)[57], random forests (RF) as implemented in `scikit-learn`[58], and linear forests (LF) as implemented in `linear-forest` (https://github.com/cerlymarco/linear-tree)[59].

Table 3. Comparison of MAE (in eV/atom), RMSE (in eV/atom), and coefficient of determination (R^2) between different ML models.

	MP18			New AoI in MP21			Ratio of metrics	
	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE
ALIGNN-MP18	0.022	0.052	0.999	0.297	0.747	0.194	13.5	14.4
XGB	0.075	0.137	0.984	0.239	0.537	0.582	3.2	3.9
RF	0.088	0.165	0.977	0.382	0.879	−0.119	4.3	5.3
LF	0.108	0.179	0.972	0.327	0.606	0.469	3.0	3.4

The metrics for the MP18 dataset are obtained for the test set following the same 60000-5000-4239 train-validation-test split as in the ALIGNN-MP18 paper[34]. The metrics for the new alloys in the MP21 are obtained with the predictions of the MP18-pretrained models. The last two columns show the ratio of the prediction error of the new MP21 alloys compared to that of the MP18.

XGB builds sequentially a number of decision trees in a way such that each subsequent tree tries to reduce the residuals of the previous one. RF is an ensemble learning technique that combines multiple independently built decision trees to improve accuracy and minimize variance. LF combines the strengths of the linear and RF models, by first fitting a linear model (in this work, a Ridge model) and then building an RF on the residuals of the linear model.

The motivation for using the traditional models for understanding the ALIGNN-MP18 performance issue is three-fold. First, do traditional ML models fail to generalize as well, or is this failure unique to neural networks? Second, traditional models can provide more interpretability than neural networks and can be used as surrogate models of the ALIGNN in the subsequent analysis. Finally, traditional models are computationally much easier to train than large neural networks, allowing us to perform more detailed statistical examinations. In fact, the reference implementation of ALIGNN-MP18[34] required a total compute cost of 28 GPU hours plus 224 CPU hours for training on the MP18 dataset. For comparison, the same training with traditional ML models takes 0.02 CPU hours (four orders of magnitude less compute than the ALIGNN). First, the XGB, RF, and LF models are hypertuned, trained, and tested with the same train-validation-test split of the MP18 as for the ALIGNN model. Then, the models are trained on the MP18 and tested on the new AoI in the MP21. Comparisons of their performance metrics and predictions are shown in Table 3 and Fig. 2, respectively.

Table 3 shows that the MP18 test set MAE of the traditional models are three to five times larger than that of the ALIGNN-MP18. This is consistent with literature findings that neural networks usually outperform traditional models in various benchmarks of large materials databases[7,37]. However, evaluating model performance from the random train-validation-test split is based on the assumption that data distributions are identical for the training and test sets, which may not hold when exploring new materials. Therefore, such performance scores are not good estimates of the model's true generalizability[37,53]. Indeed, when the models are applied to the new AoI in MP21, the large performance difference between traditional and ALIGNN models disappears. More strikingly, XGB outperforms ALIGNN in terms of the MAE, RMSE, and R^2 scores, whereas LF outperforms ALIGNN in terms of the RMSE and R^2 scores. An equal footing comparison of the extrapolation performance should also take into account the complexity and capacity of the models. A more consistent comparison may be to compute the ratio of the performance metrics obtained with the training set and the test set, which are shown as the last two columns in Table 3. The performance degradation of the traditional models is less severe than that of the ALIGNN model.

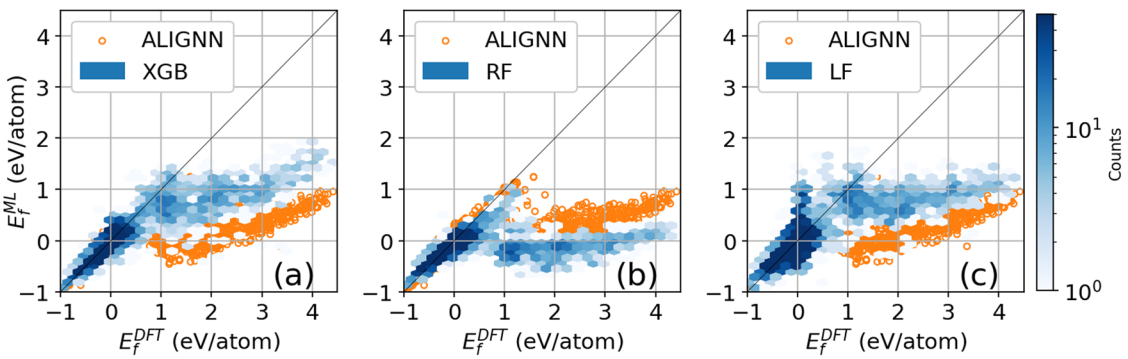

Fig. 2 Comparison of MP18-pretrained model performance on the AoI test set. The ALIGNN model performance is compared to that of the (**a**) XGB, **b** RF, and **c** LF models.

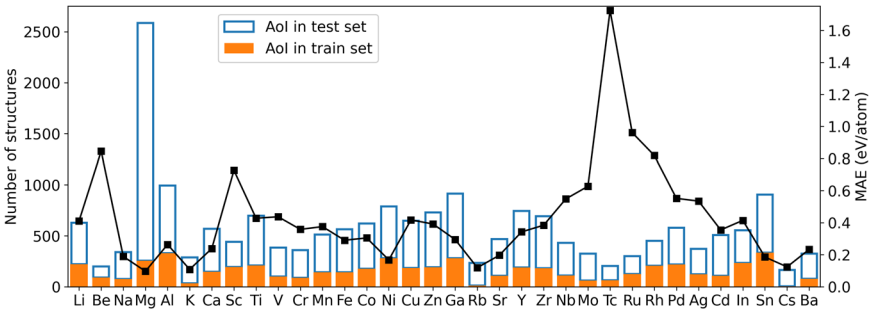

Fig. 3 Number of AoI containing a given element. The line plot (with respect to the right Y axis) indicates the MAE by the ALIGNN-MP18 on the AoI test set.

Figure 2 gives a more detailed comparison of the prediction performance of the MP21 new alloys. Compared to the ALIGNN model, the XGB model leads to larger errors in the E_f^{DFT} range below 0.5 eV/atom, but performs considerably better for predicting high-energy alloys, of which there are fewer structures that are misclassified as having negative formation energies. On the other hand, the RF model performs similarly to the ALIGNN model in the E_f^{DFT} range below 0.5 eV/atom but worse than the latter for high-energy structures. Interestingly, the LF model, in which the linear model is first fitted before training the RF model, improves the predictions for high-energy structures to an extent similar to the XGB model. The better RMSE scores for the XGB and LF models are attributed to the less degraded predictions for those high-energy structures.

The above discussion of Table 3 and Fig. 2 shows that the performance degradation issue observed in the ALIGNN model also occurs in other traditional descriptor-based models, but the performance degradation can be quite different, with the XGB and LF models demonstrating less performance degradation. In the following sections, we will reveal the origin of the performance degradation and the reasons behind the better generalizability of the XGB and LF models.

Diagnosing generalization performance degradation

In the previous section, we have shown that the performance issue on the AoI test set is common to different ML models, indicating that it is likely related to the distribution shift between the training and the test sets. For instance, the test set may cover compositions or structures that lie far away from the training set. Here we show how to diagnose this issue in a holistic and detailed manner, and discuss some important insights resulting from this analysis.

We start by comparing the distributions of some basic compositional and structural features between the MP18 and MP21 datasets. In Fig. 3, we count for each element *X* of 34 metallic elements the number of *X*-containing AoI in the training and the test set (see Supplementary Fig. 1 for the distribution of all materials). We also plot the MAE of ALIGNN-MP18 for the corresponding *X*-containing AoI in the test set to investigate potential correlations between large MAE and elements that are underrepresented in the AoI training set. We find that although there are few AoI that contain elements such as K, Rb, and Cs in the training set, the corresponding test MAE are actually rather small. Indeed, we find a Spearman's rank correlation coefficient (r_S) of 0.06, i.e., negligible correlation, between the test MAE of *X*-containing AoI and the number of *X*-containing AoI in the training set. Meanwhile, we find a weak anti-correlation (r_S equal to −0.42) between the test MAE of *X*-containing AoI and the number of *allX*-containing structures (i.e., AoI and Non-AoI) in the training set, although such a correlation vanishes above a threshold of 1000 *X*-containing structures (Supplementary Fig. 2). This suggests that chemically less relevant data can still inform ML models and may reduce generalization errors in a target subspace, though to a limited extent.

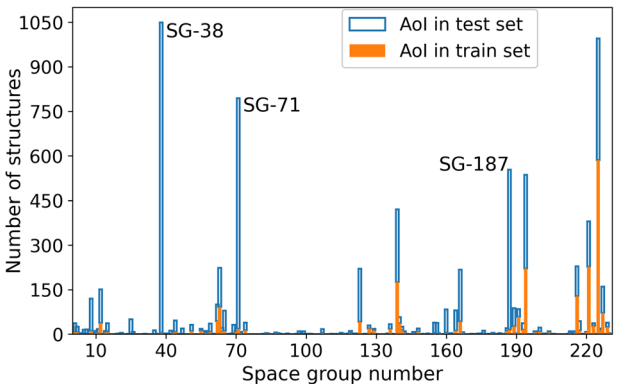

Fig. 4 Number of structures as a function of space group number. For reference, the lattice type for a given interval of SG numbers is as follows: [1,2] triclinic, [3,15] monoclinic, [16,74] orthorhombic, [75,142] tetragonal, [143,167] trigonal, [168,194] hexagonal, and [195,230] cubic.

Another basic composition-related feature is the number of elements contained in a structure. We find that the majority of the AoI in the training and the test sets are binary and ternary systems. The poorly predicted structures are the ternary alloys and some binary ones that have a E_f^{DFT} larger than 0.5 eV/atom (Supplementary Fig. 2).

To study the data distribution in the structural space, we consider the crystallographic space group (SG), which describes the symmetry of a crystal. There are, in total, 230 SG for three-dimensional crystals, and the numbers of AoI belonging to these SG are shown in Fig. 4 (see also Supplementary Fig. 3 for the error distribution). It can be seen that there are few training data but much more test data for the SG-38, SG-71, and SG-187 structures. The parity plots for these structures are shown in Fig. 5. The formation energies for the 538 SG-187 structures in the test set are well predicted, although there are only 15 training AoI with this SG. For the SG-38 AoI, the 1045 test samples that lie well beyond the small formation energy range of the 4 training data are also reasonably well predicted. By contrast, while the formation energies of the test SG-71 AoI in the E_f^{DFT} range covered by the training data are well predicted, those with E_f^{DFT} higher than 0.5 eV/atom are considerably underestimated by the ALIGNN-MP18 model. The different generalization behavior among these three SG suggests that failure to generalize is not strictly explained by the underrepresentation of a given SG in the training data, nor by the range of target values.

While it is found that the poorly predicted data are primarily associated with ternary SG-71 structures, it is unclear why it is

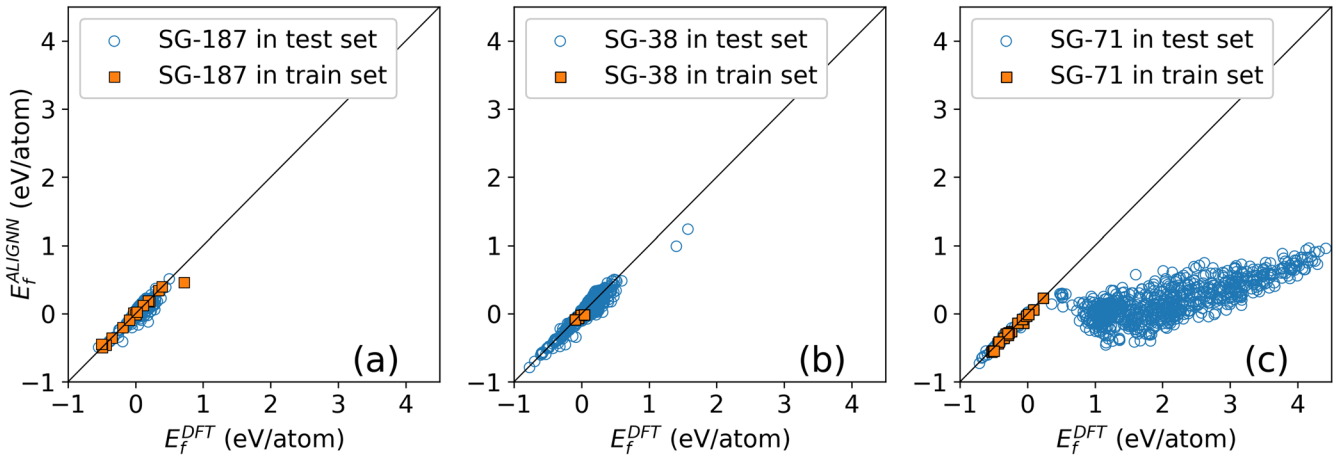

Fig. 5 Parity plot for the AoI data in different space groups (SG). a SG-187, **b** SG-38, **c** SG-71.

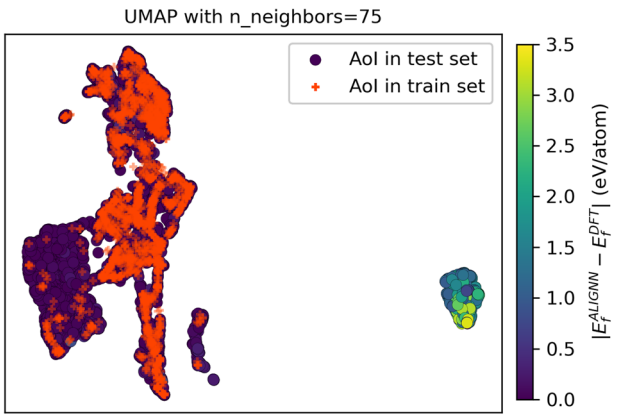

Fig. 6 UMAP projection of the 90-dimensional feature space for the AoI training and test. The X and Y axis are not shown because the two dimensions in UMAP have no particular meanings. For the UMAP projection with only the test data, the reader is referred to Supplementary Fig. 4.

these structures that are particularly hard to predict for the ALIGNN model. It would be difficult to interrogate the ALIGNN model for a physical understanding of the problem. On the other hand, we find that there is a relatively strong correlation in the test set predictions between the ALIGNN and traditional ML models (Pearson's r for ALIGNN versus RF: 0.83, ALIGNN versus XGB: 0.77, ALIGNN versus LF: 0.68) and we can therefore use these models as surrogates for the ALIGNN to study the feature space in place of the neural network's representation.

As mentioned in the previous section, there are 90 features after dropping the highly correlated ones from the initial set of 273 Matminer-extracted features. A typical way to understand high-dimensional data is to project them on a two-dimensional plane by applying dimension reduction. Here we use Uniform Manifold Approximation and Projection (UMAP), a stochastic and non-linear dimensionality reduction algorithm that preserves the data's local and global structure[60]. One of the key hyperparameters in UMAP is `n_neighbor` which constrains the size of the local neighborhood for learning the data's manifold structure. Lower values of `n_neighbor` force UMAP to concentrate on the local structure of the data, whereas higher values push UMAP to provide a broader picture by neglecting finer details[60]. By varying this hyperparameter, one can therefore obtain an idea of the data's structure at different scales. In Fig. 6, we show a UMAP visualization of the feature space of the AoI training and test data. The test samples with low prediction errors are those clusters covered by the training data, whereas the majority of the poorly predicted alloys (which are largely SG-71 structures) form an isolated cluster away from the rest of the data. Supplementary Figure 4 provides additional UMAP visualizations with smaller `n_neighbor`, where there are smaller and more dispersed clusters.

It is worth noting that we have also attempted the commonly used principal component analysis (PCA) but found no clear clustering trend (Supplementary Fig. 5). This can be related to the fact that PCA is a linear algorithm and is not good at decoding the potentially non-linear relationships between features. Another reason may be that PCA looks for new dimensions that maximize the data's variance but does not preserve the local topology of the data as UMAP does in Fig. 6.

Figure 6 is a clear demonstration in the feature space that the poorly predicted test samples lie in an area well beyond that of the training AoI data. A complementary and more detailed understanding can be obtained by comparing the feature value ranges between the AoI training and test data. Figure 7 shows the features whose ranges in all AoI data are larger by more than 5% than the ranges in training AoI data. Substantial changes in the value ranges can be noted for some features. In particular, only the lower 1/3 portion of `mean neighbor distance variation` and the higher 4/5 of `mean CN_VoronoiNN` feature values are covered in the AoI training data. The feature `mean CN_VoronoiNN` corresponds to the average number of nearest neighbors, while the feature `mean neighbor distance variation` is the mean of the nearest neighbor distance variation, which measures the extent of atom displacement from high-symmetry sites and the extent of the lattice distortion against high-symmetry structures[27,61]. The right panel in Fig. 7 clearly reveals that the test data with large prediction errors have high `mean neighbor distance variation` and low `mean CN_VoronoiNN` values, namely the poorly predicted structures are the ones with strong lattice distortion and a small number of nearest neighbors.

It should be noted that the generalization performance degradation discussed in this work is likely to be a widespread issue across materials datasets. Indeed, human bias is known to present in materials data and have adverse effects on ML performance[45]. Its presence has recently been documented for other computational databases such as OQMD and JARVIS-DFT[62]. This could be related to the fact that these databases are the results of mission-driven calculations which are focused on specific materials domains and applications instead of a general, diversified, and unbiased representation of materials. Therefore, as

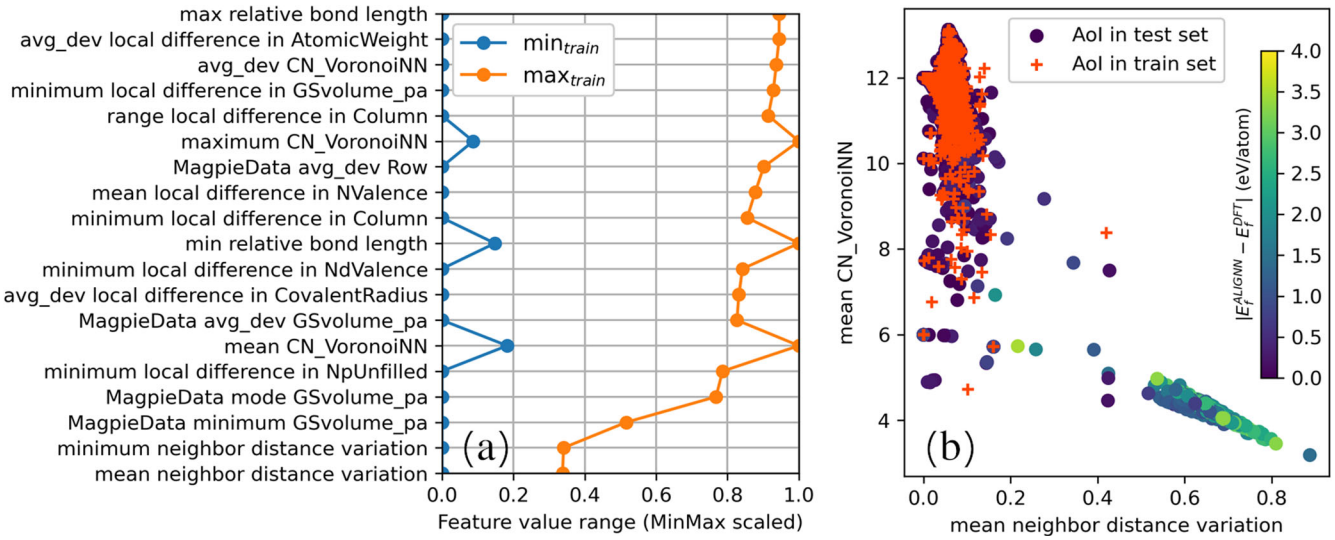

Fig. 7 Distribution of the AoI training and test data. a Feature value range of the training AoI rescaled with respect to that of all the AoI data. **b** Scatter plot of the AoI training and test data.

the funding and mission of the database builders change with time, so do the materials representation and the distribution bias in the datasets, leading to the degraded performance on out-of-distribution data.

Foreseeing performance issue

Our analysis in the previous section shows that ML models fail to generalize for compounds with large DFT formation energies relative to the range of formation energies in the training data. However, in a materials discovery setting, we must foresee this generalization risk without prior knowledge. In other words, it is important to identify the applicability domain and know whether ML models may be extrapolating and unreliable when used to explore unknown materials.

The natural idea is to define an applicability domain based on the training data density and coverage in the feature space, or equivalently estimate the similarity and distance between the training and test sets. However, this is not trivial in practice. While estimating data density based on basic compositional and structural features, as shown in Figs. 3, 4 could provide some indications of a potential distribution shift; our discussion, such as the one for Fig. 5, also shows that fewer data for some SGs do not necessarily lead to poor predictions. Perhaps a more robust and comprehensive picture of the data can be obtained by extracting meaningful and predictive features and visualizing them with the aid of dimension-reduction techniques such as UMAP. The distribution and clustering of the training and test data, as shown in Fig. 6, can clearly help identify the test samples for which the ML predictions would be problematic. In addition, comparing the range of feature values in the training and test data (Fig. 7) is a simple yet effective way to find out whether ML models are extrapolating when used to explore new regions of materials space. Various techniques, including the above-mentioned ones, should be used to inspect the training and the target space during the deployment of ML models, in order to reduce the risk of extrapolation in materials exploration.

Apart from carefully examining the feature space of datasets, one can also train multiple ML models and be more skeptical of the predictions of the test data with significant disagreement. For instance, our results in Fig. 2 show that different ML models show considerable disagreement for those out-of-distribution samples. Therefore, the degree of disagreement between the ML models can also be used to identify out-of-distribution samples. To better

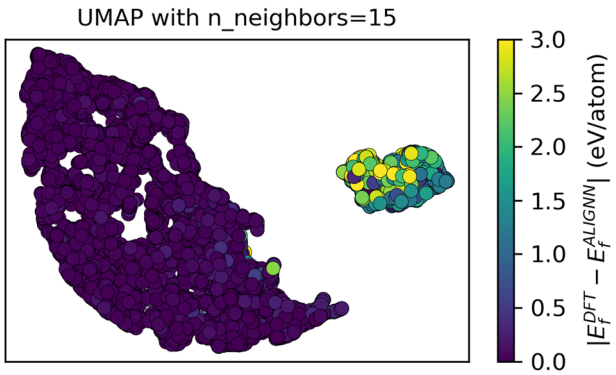

Fig. 8 UMAP projection of the AoI test data represented by the ML model disagreement. The disagreements between the ALIGNN and other models are used to represent the data points.

illustrate this point, we compute the prediction difference between the ALIGNN and other models, namely $|E_f^{\mathrm{ALIGNN}} - E_f^{\mathrm{XGB}}|$, $|E_f^{\mathrm{ALIGNN}} - E_f^{\mathrm{RF}}|$, and $|E_f^{\mathrm{ALIGNN}} - E_f^{\mathrm{LF}}|$ for each of the test data. We then use UMAP to project the test data represented by the model disagreement in Fig. 8, where the data are separated into two clusters. The cluster located on the left is associated with test data having, on average, a much larger disagreement compared to the cluster on the right. Specifically, the mean value of $|E_f^{\mathrm{ALIGNN}} - E_f^{\mathrm{XGB}}|$ is 0.69 eV/atom (0.07 eV/atom) for the cluster located on the left (right).

Another commonly employed method to identify out-of-distribution test samples is to use of uncertainty quantification. However, quantifying the uncertainty associated with the neural network predictions is challenging[7,63] and is beyond the scope of this work. Instead, we consider the uncertainty associated with the RF model, based on the quantile regression forests[64]. The prediction uncertainty of the RF model is computed as the width of the 95 % confidence interval, namely the difference between the 2.5 and 97.5 percentiles of the trees' predictions. As shown in Fig. 9, the RF uncertainty is only moderately correlated with the true prediction error for the test data. Based on the uncertainty distribution of the AoI in the training set, one may consider an uncertainty threshold between 1.5 and 2.0 eV/atom for identifying

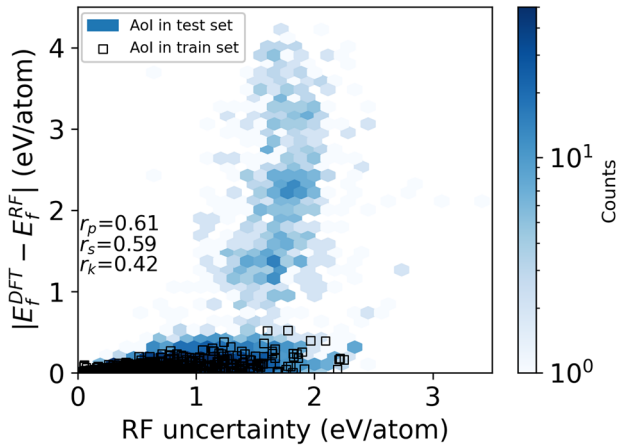

Fig. 9 Prediction uncertainty versus prediction error on the training and test AoI data for the MP18-pretrained RF model. The Pearson (r_p), Spearman (r_s), and Kendall (r_k) correlation coefficients for the prediction uncertainties and errors of the test AoI data are shown for reference.

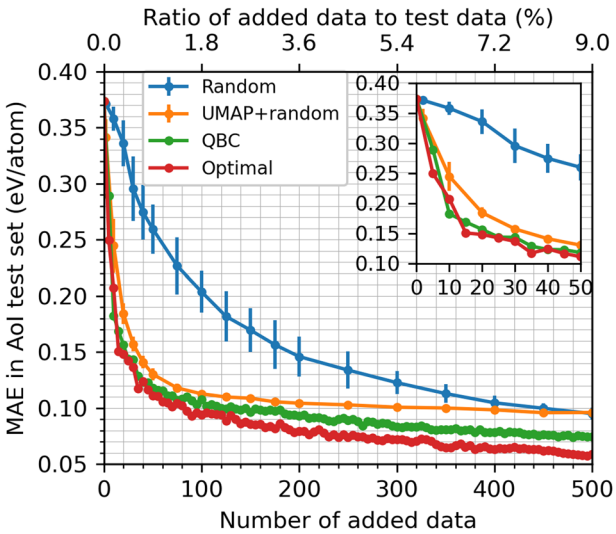

Fig. 10 Test MAE as a function of a number of selected test data added to the training set. The inset shows the enlarged region at the early stage of the active learning process. All the results are reported for the RF model. The random sampling and the UMAP-plus-random results are the averages of 10 runs with different random seeds, with the error bars indicating the standard deviation.

samples that cannot be reliably predicted. However, using these thresholds not only includes many structures that actually have low prediction errors, but also excludes the poorly predicted structures whose prediction uncertainties are between 1.0 to 1.5 eV/atom. Therefore, the RF uncertainty quantification does not allow for effectively discerning the out-of-distribution from the in-distribution samples.

Improving prediction robustness for materials exploration

Once we spot the gap between the training and test data, the next step is to improve prediction robustness by acquiring new data, ideally with a minimum additional cost. In the following discussion of different acquisition policies, we consider the RF model as a proxy for the ALIGNN model, because it is much faster to update than the ALIGNN model, and its predictions have the best correlation with the ALIGNN predictions compared to the LF and XGB models. Active learning with the ALIGNN model is beyond the scope of this work.

Our discussion in the previous sections can provide insights for establishing the acquisition policy. For instance, one can prioritize the UMAP space poorly covered by the training data. In Fig. 10, we demonstrate the effectiveness of this simple idea. We add a given number of samples randomly taken from the isolated cluster in the UMAP plot (Fig. 6) to the original MP18 training set to train the RF model. We find a significant decrease in the test MAE, compared with the baseline acquisition policy of randomly taking data from the whole test set. With only 50 data (out of 5539 test data) added, the UMAP-guided random sampling leads to a test MAE of 0.13 eV/atom, which is only half of the test MAE of 0.27 eV/atom resulting from the baseline policy (random sampling) with the same number of added samples. The latter needs five times the number of samples to arrive at the same MAE.

As discussed in Fig. 8, the level of disagreement between the ML models is also useful in finding the poorly predicted samples. We, therefore, consider the query by committee (QBC) acquisition, where we select the test data that have the strongest disagreement among the three committee members (RF, LF, and XGB). As shown in Fig. 10, the QBC strategy shows a slightly better performance than the UMAP-guided random sampling. Hoping to find an even better performance in the early acquisition stage, we further consider combining the QBC with the UMAP-guided sampling, but find the resulting performance is similar to using only the QBC strategy. To estimate whether this is because we are reaching the optimal strategy, we compute another acquisition curve, where we select samples that have the largest RF-DFT disagreement. As the DFT labels are assumed known, this curve is not regarded as a true active learning acquisition, but is only used to estimate the optimal performance that active learning can reach. It is clear from Fig. 10 that the QBC curve is quite close to the estimated optimal curve, so it is not surprising that combining it with UMAP does not bring further improvement.

It is worth noting that with the UMAP-guided sampling or the QBC policy, adding only 1% of the data already results in a reasonable test MAE. Therefore, these two strategies are very effective in identifying the most diversified and informative samples. Though Fig. 10 shows that adding even more samples can further improve the model performance in the AoI subspace, such an improvement is rather incremental. The compute should be saved to explore the regions of materials space that could bring potentially drastic gain in the prediction robustness and accuracy.

We note that the active learning strategies proposed here focus on finding out-of-distribution data points rather than eliminating dataset bias, which is a plausible source of the observed generalization performance degradation. Indeed, human bias in datasets is known to have negative impacts on ML models. While simple random sampling might mitigate these impacts and result in better models than human selection when building a database from scratch[45], it is not necessarily the optimal strategy for expanding the existing database (see Fig. 10). In practice, mission-driven databases may already have biases and continue to expand with biases into the new material space defined by new funding projects. Therefore, focused on the scenario where an existing database and a pool of candidate materials to explore are proposed, our active learning strategies enable the identification of the gap between the existing and proposed datasets and the acquisition of only the data points that can best improve the model performance.

In case acquiring new data is not possible, prediction robustness for those out-of-distribution samples can still be improved by using more extrapolative models. For instance, tree-based models are usually considered to be interpolative, as is also found here for the RF model (Fig. 2). By simply adding a linear component to the RF model; however, our LF model gives a more robust estimation of the stability for those out-of-distribution samples. The better extrapolation performance is enabled by the features whose ranges in the test set far exceed those in the training set, as removing these

features from the LF model reduces the extrapolation performance to the same level as the RF model. On the other hand, the better extrapolation performance of the XGB and LF models also comes from the training data outside the space of interest (namely non-AoI), since training only on the MP18 AoI data leads to performance similar to that of the RF model. This indicates that ML models can learn from less relevant data outsides the target space for better generalization performance.

DISCUSSION

This work is focused on the prediction robustness of ML models, by examining the formation energy predictions of the MP18-pretrained models for the new alloys in the latest MP21 dataset. We considered the ALIGNN model, a graph neural network with state-of-the-art performance in the Matbench formation energy prediction task, as well as three traditional descriptor-based ML models (XGB, RF, and LF). Despite the excellent test performance in the MP18, the MP18-pretrained ALIGNN model strongly underestimated the DFT formation energies of some test data in the MP21. While this performance issue was also found in the traditional ML models, the XGB and LF models provided more robust phase stability estimation for the test data. We analyzed and discussed the origins of performance degradation from multiple perspectives. In particular, we used UMAP to perform dimension reduction on the high-dimensional Matminer-extracted features, revealing that the poorly predicted data lie far beyond the feature space occupied by the training set. With these insights, we then discussed possible methods, including the UMAP-aided clustering and the query of multiple ML models, to identify out-of-distribution data and foresee performance degradation. Finally, we provided suggestions to improve prediction robustness for materials exploration. We showed that the accuracy can be greatly improved by just adding a very small amount of new data as identified by UMAP clustering and querying different ML models. We believe that UMAP-guided active learning shows promising potential for future dataset expansion. In cases where data acquisition is not possible, we also propose to include extrapolative components such as linear models for a more robust prediction for out-of-distribution samples. We hope this work can raise awareness of the limitations of the current ML approaches in the materials science community and provide insights for building databases and ML models with better prediction robustness and generalizability. As a perspective, a similar but more extended and systematic analysis of ML generalization performance on other materials properties, including spectral ones such as density of states, and across multiple databases, will be an interesting and important future work.

METHODS

The 2018.06.01 snapshot of Materials Project is retrieved by using JARVIS-tools[20], while the latest 2021.11.10 version is retrieved by using the Materials Project API[18]. For each material, Materials Project uses the `material_id` field as its identifier and the `task_ids` field to store its past and current identifiers. The structures in the MP21 `task_id` field contains an MP18 identifier and are considered as the materials existing in the MP18, whereas the rest in the MP21 are considered to be the new materials unseen in the MP18.

We use the ALIGNN-MP18 model that was published with the original paper[34]. We use Matminer[28] to extract 273 compositional and structural features[27], and obtain 90 features after sequentially dropping highly correlated features (with a Pearson's r of 0.7 as the threshold). We use three traditional ML models: the gradient-boosted trees as implemented in `XGBoost` (XGB)[57], random forests (RF) as implemented in `scikit-learn`[58], and linear forests (LF) as implemented in `linear-forest` (https://github.com/cerlymarco/linear-tree)[59]. For the XGB model, we use 2000 estimators, a learning rate of 0.1, an L1 (L2) regularization strength of 0.1 (0.0), and the histogram tree grow method. For the RF model, we use 100 estimators, 30% of the features for the best splitting. We combine the same RF model with a Ridge model with a regularization strength of 1 for the LF model. We use the packages' default settings for other hyperparameters not mentioned here.

DATA AVAILABILITY

The data required and generated by our code can be downloaded from https://zenodo.org/record/7659267#.ZBi1-9l8rm6.

CODE AVAILABILITY

The code for ML training, analysis, and figure generation in this work can be found at https://github.com/mathsphy/paper-ml-robustness-material-property.

Received: 23 October 2022; Accepted: 28 March 2023;
Published online: 07 April 2023

REFERENCES

1. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
2. Vasudevan, R. K. et al. Materials science in the artificial intelligence age: high-throughput library generation, machine learning, and a pathway from correlations to the underpinning physics. *MRS Commun.* **9**, 821–838 (2019).
3. Morgan, D. & Jacobs, R. Opportunities and challenges for machine learning in materials science. *Annu. Rev. Mater. Res.* **50**, 71–103 (2020).
4. DeCost, B. L. et al. Scientific AI in materials science: a path to a sustainable and scalable paradigm. *Mach. Learn. Sci. Technol.* **1**, 033001 (2020).
5. Hart, G. L. W., Mueller, T., Toher, C. & Curtarolo, S. Machine learning for alloys. *Nat. Rev. Mater.* **6**, 730–755 (2021).
6. Stach, E. et al. Autonomous experimentation systems for materials development: a community perspective. *Matter* **4**, 2702–2726 (2021).
7. Choudhary, K. et al. Recent advances and applications of deep learning methods in materials science. *npj Comput. Mater.* **8**, 59 (2022).
8. Schleder, G. R., Padilha, A. C., Acosta, C. M., Costa, M. & Fazzio, A. From DFT to machine learning: recent approaches to materials science–a review. *J. Phys. Mater.* **2**, 032001 (2019).
9. Green, M. L., Maruyama, B. & Schrier, J. Autonomous (ai-driven) materials science. *Appl. Phys. Rev.* **9**, 030401 (2022).
10. Kalinin, S. V. et al. Machine learning in scanning transmission electron microscopy. *Nat. Rev. Methods Primers* **2**, 1–28 (2022).
11. Krenn, M. et al. On scientific understanding with artificial intelligence. *Nat. Rev. Phys* **4**, 761–769 (2022).
12. Wilkinson, M. D. et al. The fair guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 1–9 (2016).
13. Jain, A. et al. A high-throughput infrastructure for density functional theory calculations. *Comput. Mater. Sci.* **50**, 2295–2310 (2011).
14. 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* **65**, 1501–1509 (2013).
15. Garrity, K. F. & Choudhary, K. Database of wannier tight-binding hamiltonians using high-throughput density functional theory. *Sci. Data* **8**, 1–10 (2021).
16. Horton, M. K., Montoya, J. H., Liu, M. & Persson, K. A. High-throughput prediction of the ground-state collinear magnetic order of inorganic materials using density functional theory. *npj Comput. Mater.* **5**, 1–11 (2019).
17. Armiento, R., Kozinsky, B., Fornari, M. & Ceder, G. Screening for high-performance piezoelectrics using high-throughput density functional theory. *Phys. Rev. B* **84**, 014103 (2011).
18. Jain, A. et al. Commentary: the Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
19. Curtarolo, S. et al. AFLOW: An automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **58**, 218–226 (2012).
20. Choudhary, K. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Comput. Mater* **6**, 173 (2020).
21. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
22. De Jong, M. et al. A statistical learning framework for materials science: application to elastic moduli of k-nary inorganic polycrystalline compounds. *Sci. Rep.* **6**, 1–11 (2016).

23. 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. *Phys. Rev. Mater.* **2**, 083802 (2018).
24. Schütt, K. T. et al. How to represent crystal structures for machine learning: towards fast prediction of electronic properties. *Phys. Rev. B* **89**, 1–5 (2014).
25. Faber, F., Lindmaa, A., Von Lilienfeld, O. A. & Armiento, R. Crystal structure representations for machine learning models of formation energies. *Int. J. Quantum Chem.* **115**, 1094–1101 (2015).
26. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2**, 1–7 (2016).
27. Ward, L. et al. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Phys. Rev. B* **96**, 024104 (2017).
28. Ward, L. et al. Matminer: an open source toolkit for materials data mining. *Comput. Mater. Sci.* **152**, 60–69 (2018).
29. Choudhary, K., DeCost, B. & Tavazza, F. Machine learning with force-field-inspired descriptors for materials: fast screening and mapping energy landscape. *Phys. Rev. Mater.* **2**, 083801 (2018).
30. Jha, D. et al. Elemnet: deep learning the chemistry of materials from only elemental composition. *Sci. Rep.* **8**, 1–13 (2018).
31. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
32. Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **31**, 3564–3572 (2019).
33. De Breuck, P. P., Hautier, G. & Rignanese, G. M. Materials property prediction for limited datasets enabled by feature selection and joint learning with MODNet. *npj Comput. Mater.* **7**, 1–8 (2021).
34. Choudhary, K. & DeCost, B. Atomistic line graph neural network for improved materials property predictions. *npj Comput. Mater.* **7**, 185 (2021).
35. Schmidt, J., Pettersson, L., Verdozzi, C., Botti, S. & Marques, M. A. Crystal graph attention networks for the prediction of stable materials. *Sci. Adv.* **7**, eabi7948 (2021).
36. Ihalage, A. & Hao, Y. Formula graph self-attention network for representation-domain independent materials discovery. *Adv. Sci.* **9**, 1–15 (2022).
37. Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Comput. Mater.* **6**, 1–10 (2020).
38. Chen, C. & Ong, S. P. Atomsets as a hierarchical transfer learning framework for small and large materials datasets. *npj Comput. Mater.* **7**, 1–9 (2021).
39. Choudhary, K. et al. Unified graph neural network force-field for the periodic table: solid state applications. *Dig. Discov.* 25–33 (2023). https://doi.org/10.1039/D2DD00096B.
40. Chen, C. & Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comput. Sci.* **2**, 718–728 (2022).
41. Kong, S. et al. Density of states prediction for materials discovery via contrastive learning from probabilistic embeddings. *Nat. Commun.* **13**, 949 (2022).
42. Stein, H. S. Advancing data-driven chemistry by beating benchmarks. *Trends Chem.* **4**, 682 (2022).
43. Kirkpatrick, P. & Ellis, C. Chemical space. *Nature* **432**, 823–823 (2004).
44. Davies, D. W. et al. Computational screening of all stoichiometric inorganic materials. *Chem* **1**, 617–627 (2016).
45. Jia, X. et al. Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature* **573**, 251–255 (2019).
46. Griffiths, R.-R., Schwaller, P. & Lee, A. A. Dataset bias in the natural sciences: a case study in chemical reaction prediction and synthesis design. Preprint at https://arxiv.org/abs/2105.02637 (2021).
47. De Breuck, P.-P., Evans, M. L. & Rignanese, G.-M. Robust model benchmarking and bias-imbalance in data-driven materials science: a case study on MODNet. *J. Phys. Condens. Matter* **33**, 404002 (2021).
48. Kumagai, M. et al. Effects of data bias on machine-learning-based material discovery using experimental property data. *Sci. Technol. Adv. Mater. Methods* **2**, 302–309 (2022).
49. Kauwe, S. K., Graser, J., Murdock, R. & Sparks, T. D. Can machine learning find extraordinary materials? *Comput. Mater. Sci.* **174**, 109498 (2020).
50. Xiong, Z. et al. Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. *Comput. Mater. Sci.* **171**, 109203 (2020).
51. Zahrt, A. F., Henle, J. J. & Denmark, S. E. Cautionary guidelines for machine learning studies with combinatorial datasets. *ACS Comb. Sci.* **22**, 586–591 (2020).
52. Ren, F. et al. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Sci. Adv.* **4**, eaaq1566 (2018).
53. Meredig, B. et al. Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. *Mol. Syst. Des. Eng.* **3**, 819–825 (2018).
54. Zhao, Z.-W., del Cueto, M. & Troisi, A. Limitations of machine learning models when predicting compounds with completely new chemistries: possible improvements applied to the discovery of new non-fullerene acceptors. *Digit. Discov.* **3** (2022).
55. Bartel, C. J. et al. A critical examination of compound stability predictions from machine-learned formation energies. *npj Comput. Mater.* **6**, 1–11 (2020).
56. George, E. P., Raabe, D. & Ritchie, R. O. High-entropy alloys. *Nat. Rev. Mater.* **4**, 515–534 (2019).
57. Chen, T. & Guestrin, C. XGBoost. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016).
58. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
59. Zhang, H., Nettleton, D. & Zhu, Z. Regression-enhanced random forests. *JSM Proceedings, Section on Statistical Learning and Data Science,* 636–647 (American Statistical Association, 2017).
60. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861 (2018).
61. Takahashi, A., Kumagai, Y., Miyamoto, J., Mochizuki, Y. & Oba, F. Machine learning models for predicting the dielectric constants of oxides based on high-throughput first-principles calculations. *Phys. Rev. Mater.* **4**, 103801 (2020).
62. Zhang, H., Chen, W. W., Rondinelli, J. M. & Chen, W. Et-al: Entropy-targeted active learning for bias mitigation in materials data. Preprint at https://arxiv.org/abs/2211.07881 (2022).
63. Abdar, M. et al. A review of uncertainty quantification in deep learning: techniques, applications and challenges. *Inf. Fusion* **76**, 243–297 (2021).
64. Meinshausen, N. & Ridgeway, G. Quantile regression forests. *J. Mach. Learn. Res* **7**, 983–999 (2006).

ACKNOWLEDGEMENTS

We acknowledge funding provided by Natural Resources Canada's Office of Energy Research and Development (OERD). ©His Majesty the King in Right of Canada, as represented by the Minister of Natural Resources, 2022.

AUTHOR CONTRIBUTIONS

K.L., B.D., and J.H.-S. conceived and designed the project. K.L. trained ML models, analyzed results, and drafted the manuscript. J.H.-S. supervised the project. K.L., B.D., K.C., and J.H.-S. discussed the results. B.D., K.C., M.G., and J.H.-S. reviewed and edited the manuscript. All authors contributed to the manuscript preparation.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-023-01012-9.

Correspondence and requests for materials should be addressed to Kangming Li.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.