communications materials

Article

https://doi.org/10.1038/s43246-024-00731-w

Probing out-of-distribution generalization in machine learning for materials

Kangming Li [1,2] ✉, Andre Niyongabo Rubungo[3], Xiangyun Lei[4], Daniel Persaud[1], Kamal Choudhary [5], Brian DeCost [5], Adji Bousso Dieng [3] & Jason Hattrick-Simpers [1,2,6,7] ✉

Scientific machine learning (ML) aims to develop generalizable models, yet assessments of generalizability often rely on heuristics. Here, we demonstrate in the materials science setting that heuristic evaluations lead to biased conclusions of ML generalizability and benefits of neural scaling, through evaluations of out-of-distribution (OOD) tasks involving unseen chemistry or structural symmetries. Surprisingly, many tasks demonstrate good performance across models, including boosted trees. However, analysis of the materials representation space shows that most test data reside within regions well-covered by training data, while poorly-performing tasks involve data outside the training domain. For these challenging tasks, increasing training size or time yields limited or adverse effects, contrary to traditional neural scaling trends. Our findings highlight that most OOD tests reflect interpolation, not true extrapolation, leading to overestimations of generalizability and scaling benefits. This emphasizes the need for rigorously challenging OOD benchmarks.

Machine learning (ML) has emerged as an important tool in accelerating scientific discovery[1–4]. This transition to data-driven science is epitomized by the development of scientific ML that seeks to build generalizable models capable of broad applicability[5]. In chemical and materials sciences, recent studies have been focused on developing universal or foundational deep learning models, which are suggested to achieve unprecedented levels of out-of-distribution (OOD) generalization towards unseen materials that are dissimilar to the training data[6–11].

However, a critical issue that has been overlooked is the potential biases in defining and selecting OOD tasks to demonstrate generalizability. OOD tasks are often defined based on simple heuristics, which, due to their subjective nature, vary between studies and even lead to contradicting interpretations of generalizability. For instance, the generalization to structures with 5+ elements, despite their omission from training, was used to showcase the emergent capability of deep learning models[6]. However, this perspective has been contested with the argument that such generalizations are anticipated from the physical heuristic that interactions in higher-order systems can be inferred from lower-order ones[12–15]. This discrepancy underscores a broader lack of agreement and discussion on what constitutes a genuinely challenging OOD task. Indeed, if the OOD test set falls within the training domain, conclusions about the superiority of a state-of-the-art ML architecture and the real benefits of model scaling may pertain only to interpolation capabilities rather than true extrapolation.

Indeed, various kinds of splitting schemes have been proposed to create OOD test sets for evaluating extrapolation performance[16–19], and there are also some dedicated OOD benchmarks[20,21]. These benchmarks, however, often rely on statistical properties, such as density in latent space or materials property range, to create splits. This approach, while systematic, can result in splits that are less physically meaningful and harder to interpret in the context of deriving physical insights. In contrast, human heuristics are sometimes employed to explain generalization gaps or create OOD splits in certain benchmarks[22–25], but a systematic examination of the validity of these heuristics is lacking. Such an examination could be invaluable in not only identifying gaps from a statistical standpoint but also in providing materials-specific insights by revealing trends unique to chemical or structural properties. More importantly, while these prior studies have focused on defining tasks and benchmarking different models, they lack a deeper inquiry into what constitutes a truly challenging task and whether all OOD tasks are genuinely difficult for models. As noted earlier, this can sometimes lead to misleading conclusions about model generalizability. Additionally, the role of scaling laws[26–28], which are essential in understanding ML performance, has not been explored in the context of OOD tasks, leaving a significant gap in the current discourse on model scaling and generalization.

In this work, we conduct a systematic examination of the performance of various ML methods across over 700 OOD tasks within large materials datasets. These tasks are specifically designed to challenge common

[1]Department of Materials Science and Engineering, University of Toronto, Toronto, ON, Canada. [2]Acceleration Consortium, University of Toronto, Toronto, ON, Canada. [3]Vertaix, Department of Computer Science, Princeton University, Princeton, NJ, USA. [4]Toyota Research Institute, Los Altos, CA, USA. [5]Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, USA. [6]Vector Institute for Artificial Intelligence, Toronto, ON, Canada. [7]Schwartz Reisman Institute for Technology and Society, Toronto, ON, Canada. ✉e-mail: kangming.li@utoronto.ca; jason.hattrick.simpers@utoronto.ca

Table 1 | In-distribution performance for the formation energy prediction

Metric	Dataset	ALIGNN	GMP	LLM-Prop	XGB	RF
MAE (eV/at)	MP	**0.033**	0.052	0.063	0.078	0.090
	JARVIS	**0.036**	0.081	0.068	0.074	0.099
	OQMD	**0.020**	0.038	0.045	0.070	0.065
R^2	MP	**0.996**	0.992	0.981	0.979	0.970
	JARVIS	**0.995**	0.985	0.982	0.981	0.968
	OQMD	**0.998**	0.995	0.995	0.987	0.985

Models are arranged in the ascending order of MAEs of the MP dataset from left to right. Best performance is highlighted in bold. The MP, JARVIS, and OQMD datasets contain 146k, 76k, and 1M entries, respectively.

heuristics based on chemistry or structural symmetry. Our findings indicate that existing models, including those as simple as tree ensembles, exhibit robust generalization across most tasks that feature new chemical or structural groups absent from the training data. Our results highlight that most of these heuristics-based criteria do not constitute truly challenging tasks for ML models. By analyzing the representation spaces of materials, we reveal that test data from well performing tasks largely reside within the training domain, whereas those from poorly performed tasks do not. Furthermore, we find that OOD generalization performance does not necessarily follow traditional scaling laws[26,27]. Notably, scaling up training set size or training time leads to marginal improvement or even degradation in the generalization performance for those challenging OOD tasks. These findings indicate that the purported benefits of scaling and emergent generalizability could be considerably overstated, due to domain misidentification driven by human bias that confounds regimes of interpolation and extrapolation.

Results

Evaluation setup

ML models are usually evaluated for their in-distribution (ID) performance by using random train-test split of the whole dataset. In this work, we assess model performance on OOD tasks. A task is considered OOD if the statistical distributions of one or more attributes differ between the training and test sets. This definition of OOD is chosen to encompass different scenarios considered in the literature[6–11], which often lack explicit definitions or adopt different heuristics-based criteria. We primarily consider six criteria for defining OOD test data: (1) materials containing element X, (2) materials containing any element in the period X, (3) materials containing any element in the group X, (4) materials of space group X, (5) materials of point group X, (6) materials of crystal system X. We consider these leave-one-X-out tasks for all possible values of X but exclude those with fewer than 200 test samples.

Three ab initio-derived materials databases have been selected for this evaluation: Joint Automated Repository for Various Integrated Simulations (JARVIS)[29,30], Materials Project (MP)[31], and the Open Quantum Materials Database (OQMD)[32]. These databases have different data distributions, hence ensuring a robust and generalized conclusion of OOD performance. The combination of datasets and grouping criteria leads to over 700 OOD tasks. OOD performance is evaluated by training a representative set of ML models, including (1) random forest (RF)[33] and XGBoost (XGB)[34] models with Matminer descriptors[35]. (2) single neural network with Gaussian multipole (GMP) expansion on electron density[36] (3) atomistic line graph neural network (ALIGNN)[37], and (4) the large language model (LLM) based LLM-Prop with crystal text descriptions[38]. It therefore covers not only common ML architectures from lightweight tree ensembles and neural networks to graph neural networks and transformer-based large language models, but also distinct input representations from human-devised descriptors and force-field-like features to crystal graphs and text.

Chemistry-based OOD generalization

In the following, our analysis primarily focuses on formation energy data, given its fundamental importance in materials science. Table 1 shows the in distribution (ID) performance obtained from a random 8:2 train-test split of the whole dataset, providing a baseline for comparison with OOD performance. We utilize two complementary performance metrics: mean absolute error (MAE) and coefficient of determination (R^2). While MAE measures the expected error on the original physical scale, its interpretation can be scale-dependent and less intuitive for assessing the goodness of predictions. In contrast, R^2 is a dimensionless accuracy measure with a value range between 1 (for a perfect model) and infinitely negative (for arbitrarily bad models) and can be conveniently compared across different OOD test sets.

Figure 1a illustrates the leave-one-element-out generalization performance of the ALIGNN and XGB models on the MP dataset, demonstrating robust OOD performance across much of the periodic table. Notably, 85% of the tasks for the ALIGNN model and 68% for the simpler XGB model achieved R^2 scores above 0.95. This broad generalization capability is surprising, given that the training set does not contain any information regarding the bonding between the left-out element and other atoms. Our results suggest that effective OOD generalization across chemistry may be more achievable than previously assumed, for both low- and high-capacity models.

Tasks with low R^2 scores are mainly associated with nonmetals such as H, F, and O. Nonetheless, ML models may still prove useful in ranking materials even in the worst performing tasks. As shown in the parity plots, the ALIGNN model systematically overestimates the formation energies of H compounds (or O compounds), with a Pearson correlation coefficient of 0.7 (or 0.9). The poor R^2 is therefore a result of systematic biases in the OOD predictions. Such bias is also shared among other ML models: all ML models considered here overestimate formation energies of H and F compounds; the three deep learning models tend to overestimate formation energies of O compounds, while tree models tend to overestimate low-energy and underestimate high-energy O compounds. Similar systematic biases, which can be addressed by simple linear corrections, have also been reported elsewhere for OOD property predictions[39,40]. Future investigation to understand and mitigate these biases can be useful for developing chemically transferable models.

The causes behind the poor OOD performance for nonmetals, particularly H, F, and O, have been examined. Initially, we considered training set size as a potential factor, but this was ruled out based on the weak correlation between training set size and OOD performance, evidenced by a Spearman's rank correlation coefficient of −0.2. Despite investigating various element attributes, none could fully account for the R^2 trends. For example, while elements with low R^2 scores tend to be more electronegative or positioned at the boundary of the periodic table, exceptions such as S, Cl, Br (which are electronegative) and Cs, Bi (positioned near the corner of the periodic table) exhibit high ALIGNN R^2 scores above 0.96, suggesting other influencing factors.

The biases leading to this poor OOD performance could be of either compositional or structural origin. The former is related to chemical dissimilarity between elements, whereas the latter refers to the failure mode where the test set associated with a given element contains materials that are structurally distinct from the training set. We propose a SHAP-based[41] (SHapley Additive exPlanations) method to identify the sources of biases. This involves training a second model to correct the model for each leave-one-element-out task and evaluating the contributions from compositional and structural features to the corrections (see Method). Figure 1b shows the violin plots of the feature contributions from the XGB model for the test data in selected leave-one-element-out tasks. Compared to the well-performing case of Cl, the compositional contributions in the cases of H, F, and O are much stronger in magnitude than structural contributions. The predominately negative compositional contributions are consistent with the corrections needed for the overestimation shown in the parity plots. The observed significant change in chemical contributions suggests that the poor OOD performance in these cases is likely associated more with differences in chemistry than with geometry between the training and test sets.

To compare the overall performance between different models, we count for each model the number of leave-one-element-out tasks with an accuracy

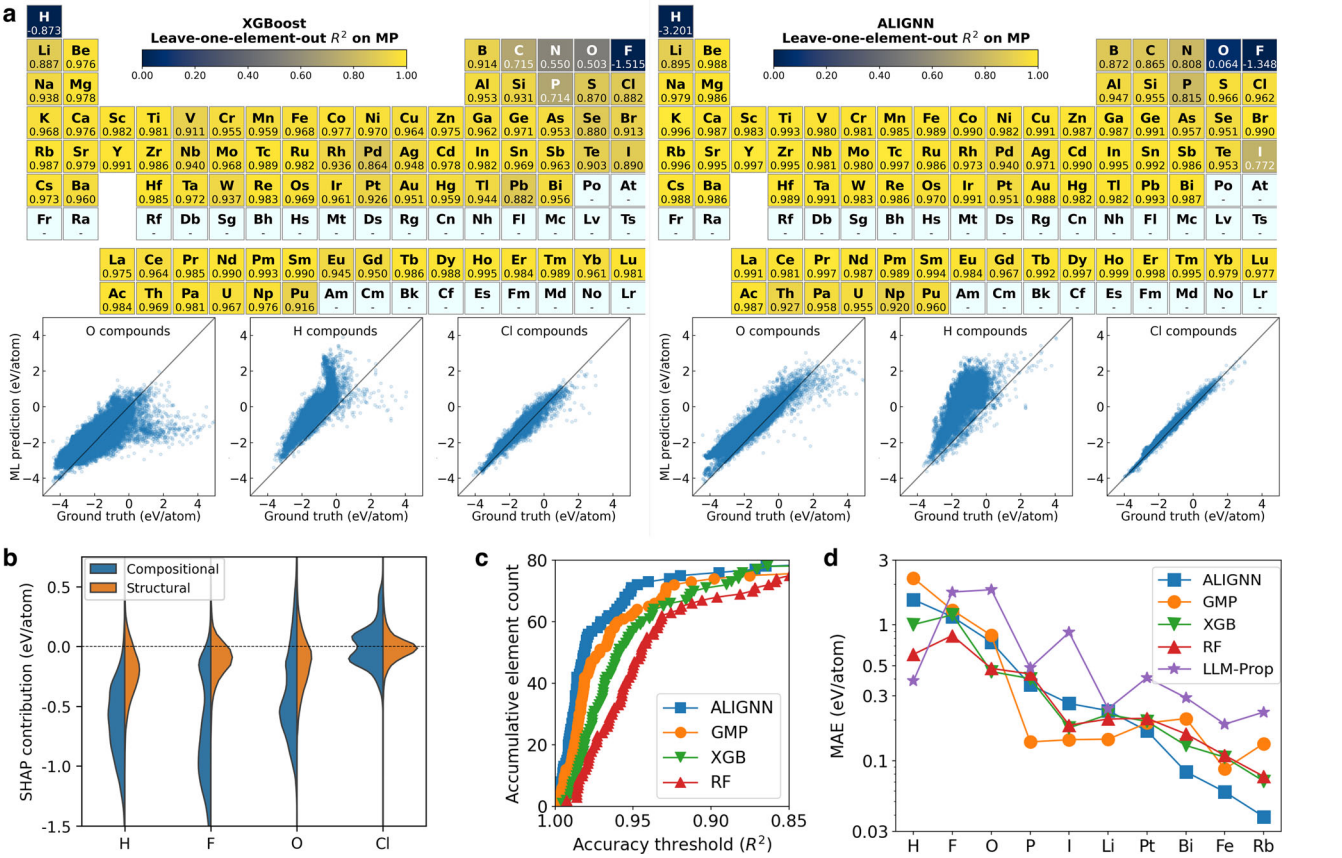

Fig. 1 | Leave-one-element-out performance on Materials Project. a Leave-one-element-out performance across the periodic table for the XGB and ALIGNN models. Despite the presence of negative R^2 for H and F, the colorbar range is bounded between 0 and 1 for better visibility. The OOD tests are performed only for elements with more than 200 data. Results for other models and datasets are provided in Supplementary Figs. 2–12. The periodic tables are plotted using pymatviz[59]. **b** Violin plots of SHAP contributions from compositional and structural features for selected left-out elements. **c** Number of leave-one-element-out tasks with R^2 higher than a given threshold as a function of threshold. LLM-Prop is not included due to the large number of tasks and high training cost. **d** Comparison of performance between ML models for selected leave-one-element-out tasks.

better than a given threshold. As shown in Fig. 1c, models with better ID performance have a better overall performance in the leave-one-element-out tasks. However, this correlation does not imply uniform superiority of any model across all tasks. Figure 1d details performance comparisons in 10 representative tasks, highlighting nuanced outcomes. In the most challenging tasks involving H, F, and O, the RF model, despite its lower ID performance, ranks as the best or second best. Conversely, for moderately challenging tasks (P, I, and Li), the GMP model outperforms others. Meanwhile, in less demanding tasks (Bi, Fe, and Rb) where MAEs are below 0.1 eV/atom, the ALIGNN model achieves a much lower MAE than the other models.

Interestingly, the LLM-Prop model, despite its good ID performance, is the worst performing model in most of the leave-one-element-out tasks. A possible cause might be the lack of inductive bias related to the element similarity, which is encoded in the representation schemes of other models. For instance, ALIGNN and tree ensembles represent atoms by basic attributes such as electronegativity, period, and group. In this way, the models can properly recognize the correlation between the left-out element and the other elements in the training set. By contrast, language models treat each element as a text token in a manner equivalent to one-hot encoding. Element similarity may not be encoded in the initial text embedding and is not guaranteed to be learned for the left-out element during the training. A potential direction to improve chemical transferability of LLMs is therefore to include additional chemical information in text descriptions during pretraining or fine-tuning. Alternatively, language models specialized for materials and chemistry applications might explore custom tokenization schemes for elemental symbols and chemical formulae that directly incorporate aspects of chemical similarity.

It is worth noting that many representation schemes that use one-hot encoding of elements or element-specific functions[42] are in principle incapable of any generalization to new chemistry. By contrast, the good generalization across chemistry shown in Fig. 1 highlights the importance of describing elements with a transferable representation. In this way, an element and its compounds can be seen as lying within the interpolation region in the chemical space of the training data in the leave-one-element-out tasks.

Appropriate representations can achieve chemical transferability even for small training sets with limited chemical diversity. To prove this point, we create a dataset from OQMD by restricting the chemistry to six elements from Cr to Cu. The leave-one-element-out R^2 scores of the XGB model within this dataset (about 1200 structures) range between 0.87 (for Cu) to 0.93 (for Fe), comparable to the in-distribution R^2 score of 0.93, demonstrating that generalization across chemistry does not necessarily require a big model or large dataset.

Figure 2a shows the more challenging OOD tasks where an entire period of elements is excluded from the training set. The worst OOD performance is found for the first period, which contains H, and the second period, which contains many nonmetals with poor leave-one-element-out performance. Overall, the OOD performance tends to be better for the periods in the middle of the periodic table, where ML models can learn from the adjacent periods. For the OOD tasks of the fourth to sixth periods, the R^2 of the ALIGNN model can achieve a R^2 score of 0.95 to 0.97, which is surprisingly good given the number of elements excluded from the training set. Interestingly, the XGB and RF models perform similarly or even better than deep learning models in most of the tasks, highlighting the robustness of tree ensembles in challenging generalization tasks.

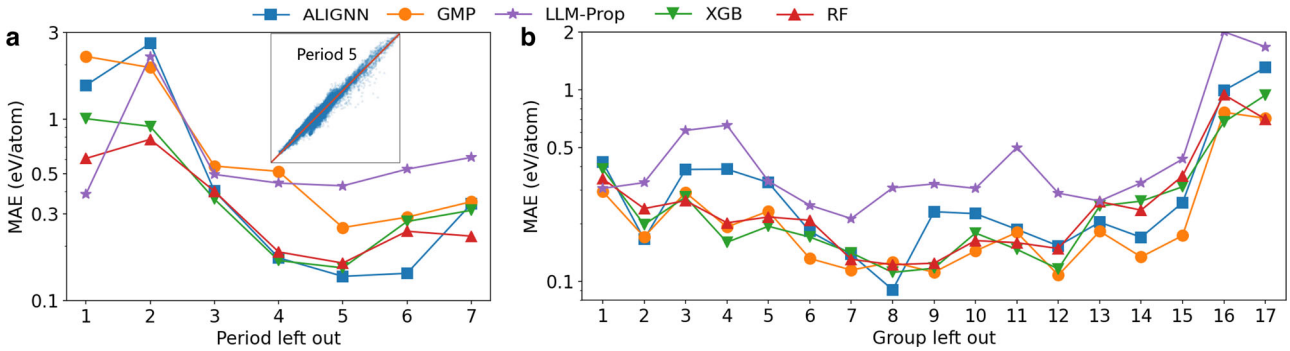

Fig. 2 | Generalization performance for leaving out a number of elements on Materials Project. a Leave-one-period-out performance, where the inset shows the parity plot for the ALIGNN predictions for the period 5 (axis range: -5 eV/atom to 5 eV/atom). **b** Leave-one-group-out performance.

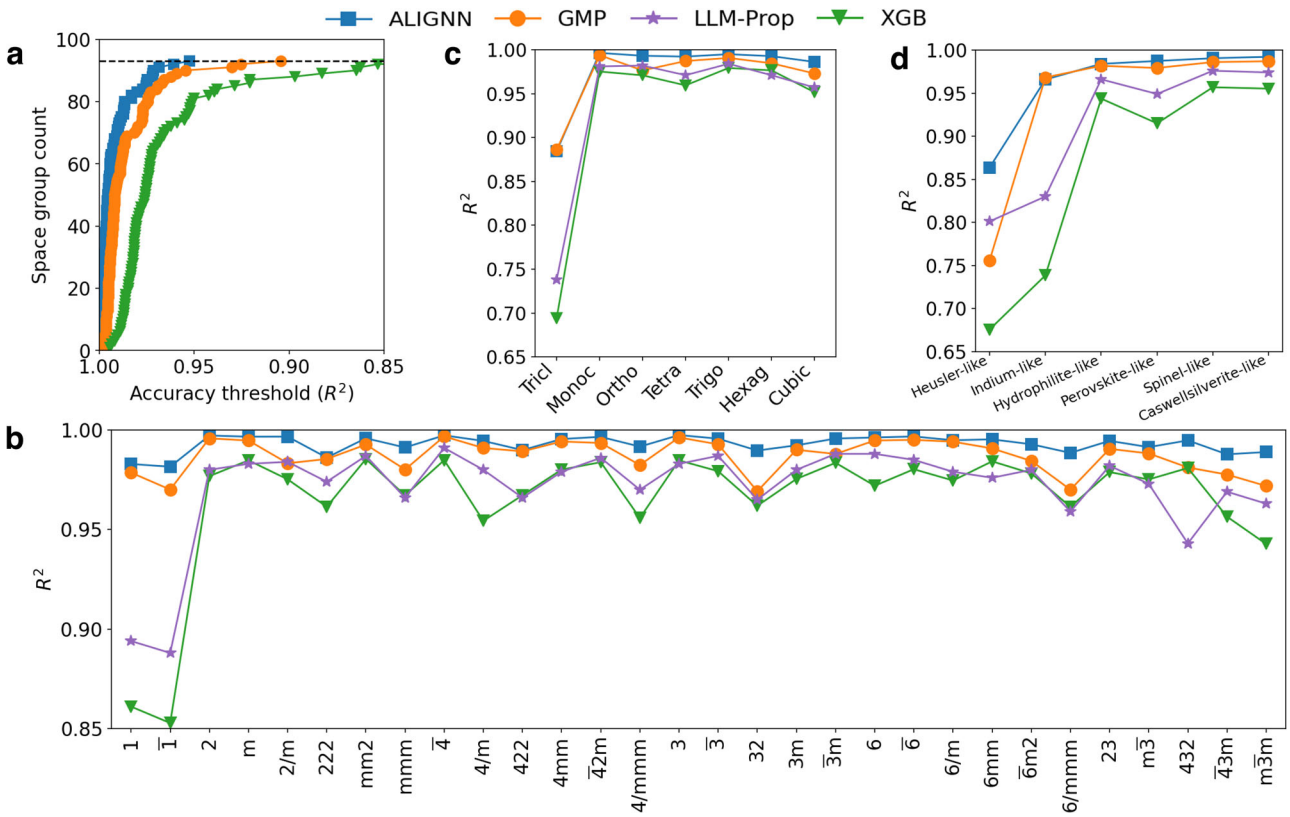

Fig. 3 | Structure-based OOD performance on Materials Project. a Number of leave one space group out tasks with R^2 higher than a given threshold as a function of threshold. The horizontal dashed line indicates the total number of tasks. LLM-Prop is not examined due to the large number of tasks and high training cost. **b** Leave one crystal system out performance in the ascending order of symmetry from left to right. **c** Leave one prototype-like out performance. Structures with fingerprints similar to a prototype are grouped together. **d** Leave one point group out performance in the ascending order of symmetry from left to right.

Similar conclusions are also found in the leave-one-group-out performance shown in Fig. 2b. First, the OOD performance also tends to be better for the groups in the middle of the periodic table. Despite its superior leave-one-element-out performance for elements in groups 3 to 5, the ALIGNN model is outperformed by the tree ensembles in these leave-one-group-out tasks. By contrast, the GMP model demonstrates the best or close-to-best performance in these tasks. The difference in the model performance ranking in the leave-one-element/period/group-out tasks indicates that these ML models have distinct advantages in different types of OOD tasks. Future investigation in understanding their respective advantages can be beneficial for designing models with better generalizability.

In the JARVIS and OQMD datasets, we also find good generalization performance across the majority of the periodic table and poor performance for a few cases of nonmetal elements. Therefore, the trend shown in Figs. 1 and 2 is not a unique artifact of the specific biases in data distribution or computational settings of the MP dataset, nor of the training set size which will be discussed later. Instead, it reveals that the existing featurization schemes can enable transferability between most of the chemical systems, though they may still fail to capture the essential characteristics for certain nonmetals, which leaves rooms for future improvement in materials featurization.

Structure-based OOD generalization

Structures are usually grouped by their crystallographic symmetries. Here we consider three basic symmetry attributes for structural grouping: space groups, point groups, and crystal systems, with each being a subcategory of the next. The structure-based OOD performance on the MP formation energy dataset are shown in Fig. 3.

The OOD performance of the structure-based tasks is much better than that of the chemistry-based OOD tasks. Figure 3a shows the overall OOD performance on all of the leave one space group out tasks. As the best performing model, ALIGNN achieves an R^2 of above 0.95 in all of the tasks with 88% of them having an R^2 score above 0.98. The GMP model is the second best performing one followed by the XGB model. The LLM-Prop model is not examined here due to the high training cost required for a large number of tasks, but it is included for the leave one point group out and leave one crystal system out tasks in Fig. 3b, c. All models demonstrate a good prediction accuracy with an R^2 of above 0.95 (or a MAE of below 0.15 eV/atom) in most of these tasks. The only cases with relatively poor performance are associated with the triclinic crystal system (and its subcategory point groups 1 and $\bar{1}$), which has minimal symmetry constraint and is structurally more diverse, posing greater challenge when generalized from structures with higher symmetry.

Our results suggest that symmetry attributes, which are often used to explain generalization performance or devise active learning algorithms[9,22,43–45], may not be effective in finding data truly dissimilar from the training set. Indeed, two symmetrically different crystals can be structurally very similar. For example, a slight distortion of a cubic lattice can transform the structure into another lattice, as in the famous cubic-to-tetragonal Martensitic transformation in steels. Therefore, the training set still contains similar structures even if all cubic structures are removed. To mitigate this information leakage and make the generalization task more challenging, we create new test sets by using the fuzzy prototype-matching algorithm implemented in the robocrystallographer package[46] to group prototype-like structures. Figure 3d shows the OOD performance on a few leave one prototype-like out tasks. Interestingly, the OOD performance in these tasks is still good, especially for deep learning models. The ALIGNN and GMP models achieve R^2 scores above 0.95 in all tasks except for Heusler-like structures, for which the ALIGNN model can still achieve an R^2 of 0.86 and an MAE of 0.08 eV/atom. The XGB model still perform reasonably well in these tasks, though its MAEs are 100% to 200% higher than those of the ALIGNN model.

Overall, the ranking of model performance in structure-based OOD tasks closely follows the ID performance ranking, in contrast to the situation in chemistry-based OOD tasks. In particular, the ALIGNN model systematically outperforms other models in all the structure-based tasks, highlighting the effectiveness of graph neural networks in capturing complex structural patterns. In addition, the LLM-Prop model also exhibits much better generalization across structural groups than across chemical groups.

Inspecting materials representations

The systematic investigation of various OOD tasks reveals that in most of the cases, generalization beyond the distribution of the training data is not challenging. It therefore seems to suggest that ML models are good at extrapolating across chemical and structural groups, in contrast to the recent findings that ML models generalize poorly even between different versions of the same databases[43,47]. To reconcile this apparent contradiction, we propose to inspect materials not from the perspective of a single human-defined attribute (e.g., element identity), but from the high-dimensional representation space of materials. Furthermore, we propose to define and distinguish two types of OOD. A test set is considered statistically OOD if the statistical distribution of one or more attributes differs from the training set, whereas a test set is considered representationally OOD if the test data are outside the region of the training set when viewed from the high-dimensional representation space. With this definition, all the OOD tasks discussed previously are statistically OOD (which will be simply referred to as OOD) but not necessarily representationally OOD.

Uniform Manifold Approximation and Projection (UMAP)[48] is used to project the high-dimensional materials representations to a two-dimensional plane. For deep learning models, materials representations (embeddings) are automatically learned during the training; embeddings of new materials can then be derived by using the trained models. For each OOD task, we apply a Gaussian kernel to the training data and evaluate the kernel density estimate of the training data for every test data point in the UMAP embedding space. Test data points with high/low density estimates are considered as representationally ID/OOD data.

Figure 4a, b show the UMAP plots of ALIGNN embeddings learned from the leave-Mg-out and leave-O-out tasks in the JARVIS dataset. In the former case, 76% of Mg-containing structures are well covered by the training data (i.e., with an embedding space probability density above the threshold), whereas 24% of Mg-containing structures are representationally OOD with a much lower R^2 score. In the latter task, 87% of O-containing structures are representationally OOD with an R^2 close to 0, whereas there are still some in-domain O-containing structures with an R^2 above 0.9. Figure 4c shows another example from the leave-H-out task in the OQMD dataset, where kernel-density estimates provide again good resolution in separating representationally ID and OOD test data: 13% of H-containing structures are identified as within the training domain with an R^2 score of 0.8. The identification of poorly predicted test data in well-performing OOD tasks and well-predicted test data in poorly-performing OOD tasks demonstrates the effectiveness of the proposed method for classifying representational domains of OOD samples.

Representational domain identification is important for interpreting generalization capabilities. Figure 4d shows the case where the test set contains structures with 5 or more elements. It has been used as an example to demonstrate the emergent ability of deep learning models[6]. This perspective was contested as the same capability was also found in simple boosted trees[12]. In fact, our UMAP analysis shows that 92% the test data are identified as lying within the representational domain of the training data, therefore this test only reflects the interpolation capacity of ML models. This example highlights not only the risk of misinterpreting generalization capabilities, but also the need for designing more challenging tests that can truly demonstrate state-of-the-art.

Domain identification can be achieved not only with the learned embeddings but also with the descriptors used in tree ensembles. An example is shown in Fig. 4e, where the test data are structures that contain any element in the fifth period. Based on kernel-density estimates, 93% of the test data are identified as representationally ID data, thereby explaining the good generalization performance for this seemingly hard task.

It is worth noting that density estimates do not perfectly correlate with prediction errors, as shown in Fig. 4f. However, from a statistical perspective there is a higher proportion of test data with high prediction errors in the low-density region than in the high-density region. Indeed, when calculating the MAE for test data in different density intervals, we find that the MAE tends to decrease with increasing density estimate (dotted line). The density estimate can still enable us to statistically distinguish between the well-predicted in-domain data and poorly-predicted out-of-domain data from an out-of-distribution test set. Future work will be needed to improve domain identification and uncertainty measures with better correlation with OOD prediction errors.

Learning curves for out-of-distribution generalization

Domain identification in the materials representation space is important in correctly interpreting the benefits of the neural scaling strategy for materials discovery applications. This strategy is based on the so-called neural scaling laws, where test errors can be consistently reduced by increasing model size, dataset size, and the amount of compute used for training deep learning models. Recent work on deep learning models for materials follows this rationale to train foundational or universal neural networks[6,8–10,20]. However, neural scaling laws have been empirically tested only for the ID generalization[26,49]. Here we perform an extensive examination on various OOD tasks and, intriguingly, find distinct scaling effects among the OOD tasks. Specifically, neural scaling laws are found to be valid only for representationally ID generalization, while the scaling effects are marginally beneficial or even adverse for representationally OOD generalization.

Figure 5a–c shows the learning curves of MAE versus training time in the leave-one-element-out tasks for F, O, and H. The training and the ID set

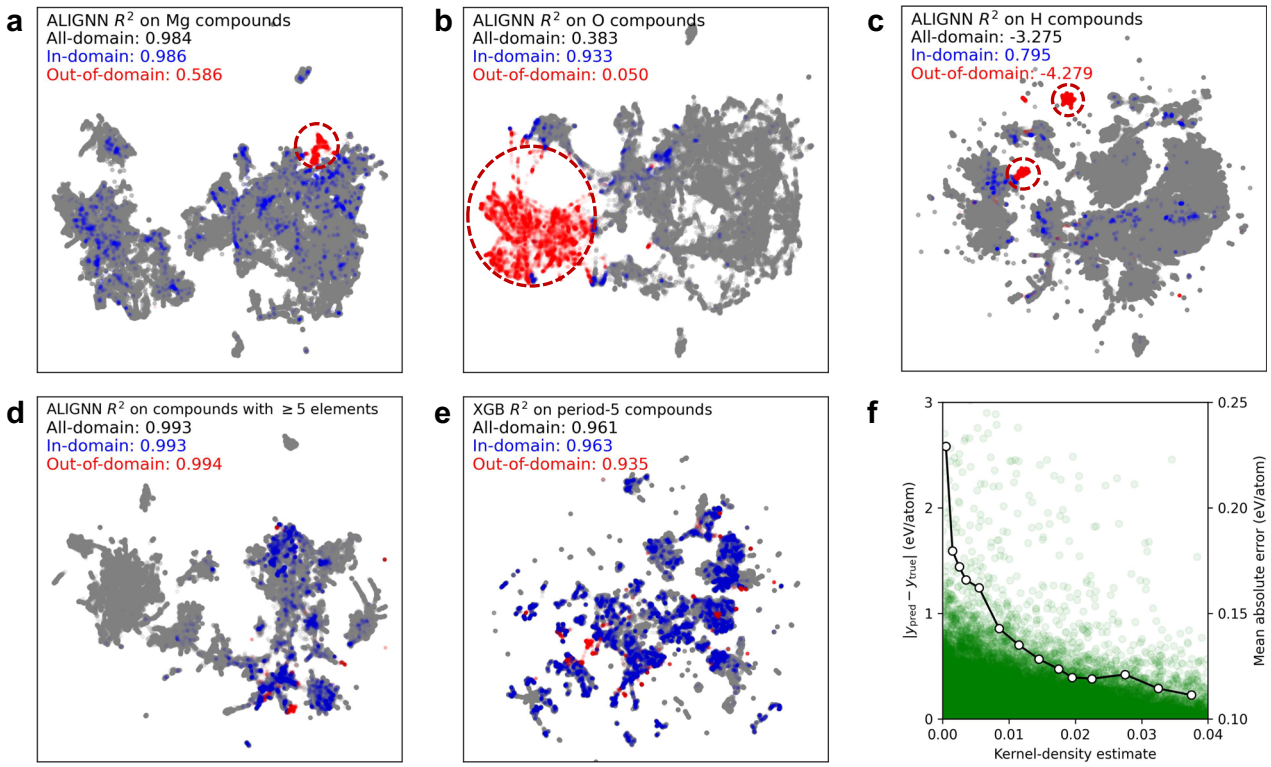

Fig. 4 | UMAP two-dimensional projection of materials representations for domain identification. a, **b** UMAP plots of ALIGNN embeddings learned from the leave-Mg-out and leave-O-out tasks in JARVIS. **c** UMAP plot of ALIGNN embeddings learned from the leave-H-out task in OQMD. **d** UMAP plot of ALIGNN embeddings learned by leaving out structures with 5 or more elements in MP. **e** UMAP plot of the XGB descriptors for the leave-period-5-out task in MP. **f** Absolute errors (left Y axis) of test data as functions of kernel-density estimates of training data for the UMAP plot of (**e**); the solid line denotes the MAEs (right Y axis) for different density intervals. In all UMAP plots, the training data, in-domain test data, and out-of-domain test data are marked in gray, blue, and red, respectively; clusters of out-of-domain test data are circled out in (**a**–**c**); the R^2 scores are indicated for the in-domain, out-of-domain, and all-domain test data.

contain 80% and 20% of the structures without the left-out element, whereas the OOD test set contains the structures with the left-out element. For the leave-F-out task, the training and ID test MAEs decrease with increasing training time of the XGB model. However, the OOD MAE decreases only during the initial 20 training rounds, then continuously increases with more training.

A similar pattern is observed in the leave-O-out and leave-H-out tasks. In fact, for representationally OOD tasks, the test MAE tend to increase or remain constant with increasing training time beyond around 20 epochs. However, for OOD tasks that are representationally ID, the OOD MAE has a tendency similar to the ID MAE, namely the error decreases with increasing training time. As discussed in Fig. 4, a key distinction between the easy and difficult OOD tasks is whether the test data lies outside the training domain in the representation space. Thus, our findings indicate that the response to scaling up training efforts may vary significantly based on whether the test data falls inside or outside the training domain.

Effects of training set size are also investigated. Take the leave-F-out task for example, models are trained with different sizes of the training sets randomly sampled from 80% of the structures without F, and evaluate performance on the held-out ID test set (20% of the structures without F) and the OOD test set (structures with F).

Figure 5d shows the ALIGNN learning curves for the ID and OOD performance for 3 representative tasks in Materials Project. As the ID learning curves for different OOD tasks are found to be almost the same, we show only one learning curve for the ID performance (dashed line) for simplicity. The ID performance consistently improves with increasing training set size, which is however not always the case for OOD performance. For the leave-F-out task, increasing the training set size from 10^2 to 10^5 only further degrades the already bad OOD performance. For the leave-H-out task, we find a "V" shaped learning curve for the performance on H-containing structures, where the MAE first decreases and then increases with increasing training set size. This violation of neural scaling laws is also found in other representationally OOD tasks (Supplementary Figs. 24–29). By contrast, we find a consistent decrease in MAE with increasing training set size, for the case where the test data are structures with 5 or more elements. This accordance with neural scaling laws is also found in other easy OOD test sets that contain mainly representationally ID data (Supplementary Figs. 24–29).

To avoid establishing our conclusion based on a single dataset, we also perform the same experiments on the OQMD dataset, which is about an order of magnitude larger than the Materials Project dataset. As shown in Fig. 5e, we find the same trend where increasing training set size only leads to marginal improvement or even degradation for test data that are outside the training domain. It should be noted that the marginal improvement is not an artifact of the limited model capacity, considering the continuous decrease in the ID test MAE with increasing training set size. In addition, similar learning curves are also found for other ML models and the JARVIS dataset (Supplementary Figs. 24–27).

It is particularly striking to observe the 5-fold increase in the OOD MAE of the leave-H-out task, when the training set size is increased from 10^4 to 10^6. Figure 5f shows the distributions of the corresponding prediction errors on H-containing structures. When trained on only 1% of structures without H, the error distribution is centered around zero with a relatively narrow spread. With increasing training set size, however, the errors shift towards positive values and exhibit a larger spread. This systematic overestimation is a reflection of biases in models that are enhanced by training on more data from a given domain. The leave-H-out task can therefore serve as an example of losing generalizability due to an overfitting to the training domain, for which future work is needed to identify its root cause and mitigation solutions.

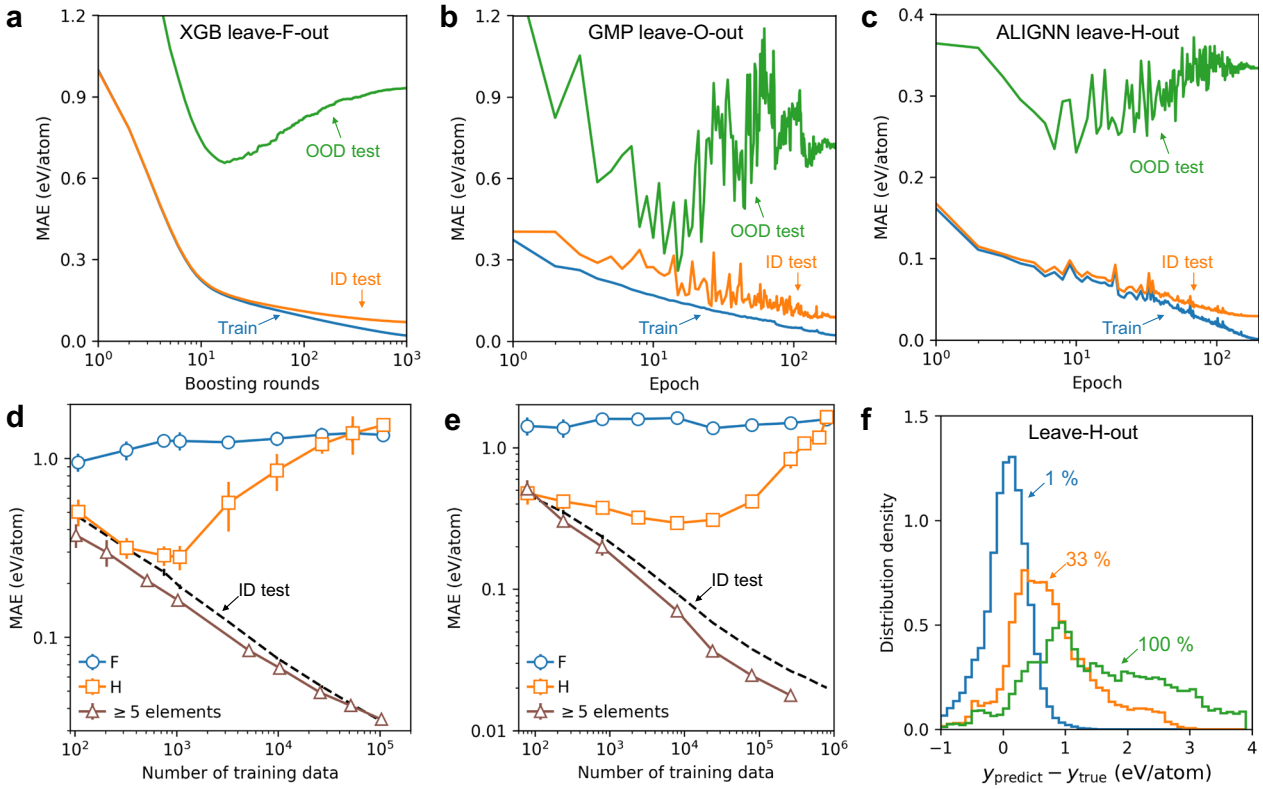

Fig. 5 | Effects of training time and training set size on performance. a–c Training loss, ID test loss, and OOD test loss as functions of training round/epoch on JARVIS. Models and left-out elements are indicated in the top of the plots. **d**, **e** ALIGNN performance on Materials Project and OQMD, respectively. Solid lines with markers denotes the MAEs as functions of training set size for 3 representative OOD tasks: leave F out, leave H out, and leave out structures with ≥5 elements. Dash line marks the ID MAE as the baseline for comparison. **f** Distribution of ALIGNN prediction errors on H-containing structures in the leave-H-out task in OQMD. Results for the models trained on 1%, 33%, and all of the structures without H are shown. Similar analysis for other tasks and ML models is provided in Supplementary Figs. 23–29.

Our results challenge previous interpretations of generalizability. Figure 5d, e include the learning curves for the scenario where the test set comprises structures with five or more elements, while the training set includes those with fewer than five. Contrary to prior claims suggesting such generalization is an emergent capability enabled by neural scaling[6], our analysis indicates that this may merely reflect in-domain generalization. According to the literature, a sudden improvement in performance above a certain threshold is required to classify an ability as emergent[50]; however, this criterion is not met in the learning curve. Additionally, the OOD MAE being lower than the ID test MAE further supports our findings from Fig. 4e that the test data lie well within the representational training domain. Thus, the observed effective generalization in these cases does not demonstrate the models' true generalizability but rather the interpolation within the representational training domain, where errors decrease as the test data become increasingly well-covered by the training data.

OOD generalization for other properties

To further explore the variability in OOD generalization across different material properties, we extended our analysis to include predictions for band gap and bulk modulus (see Supplementary Figs. 30–39). The OOD performance for these properties generally lags behind that of formation energy, which is closely tied to their inherently lower ID performance. For instance, in the ALIGNN models trained on Materials Project data, ID R^2 scores for band gap and bulk modulus are around 0.86 and 0.91, respectively, which are even lower than many of the OOD R^2 scores for formation energy. These results indicate that models already face significant challenges in achieving accurate ID generalization for band gap and bulk modulus. Therefore, the immediate priority for these properties lies in improving their ID performance, as exploring more challenging OOD tasks would be less impactful until ID generalization becomes more robust.

In terms of OOD performance degradation relative to ID performance, the most difficult chemistry-based OOD tasks for band gap are associated with transition metals, while for bulk modulus and energy, they are more commonly related to light nonmetals. These distinctions highlight how OOD performance trends can vary significantly across properties, depending on the underlying chemical characteristics being predicted. Such differences in generalization trends indicate that the thorough analysis done primarily for formation energy could be worth extending to each distinct property, which could bring different insights.

While this expanded analysis provides valuable context, the central conclusions of the study remain focused on formation energy as the most compelling case for examining biases in task design and model evaluation. In the case of formation energy, models already perform well on ID tasks, making it a suitable candidate for exploring more challenging OOD tasks where there is room for improvement. Conversely, for properties like band gap and bulk modulus, improving ID generalization is the more pressing need, and the focus on OOD generalization is of less immediate concern.

Discussion

Our extensive OOD examinations show that existing energy models exhibit remarkable generalization capabilities to new chemical or structural groups outside the training set. These findings can be rationalized by the observations that many OOD test sets are actually within the representational domain of the training data. Only a handful of truly challenging OOD tasks contain significant amounts of test data that lie outside this domain. Importantly, the performance on these representationally OOD tasks does not adhere to conventional neural scaling laws, suggesting that the benefits of scaling for OOD generalization may be overstated or misinterpreted.

These findings underscore the predictive power of machine learning models but also reveal that many OOD tasks commonly considered difficult

based on heuristic definitions may not be genuinely challenging for state-of-the-art models. This situation is reminiscent of the Turing test, once thought to be a benchmark for machine intelligence, but now recognized as inadequate for assessing the capabilities of advanced language models[51]. Similarly, heuristically defined OOD tasks often reflect models' interpolation capabilities rather than true extrapolation.

This study calls for a more rigorous examination of OOD tasks. More attention should be devoted to designing benchmarks that truly push the boundaries of model generalization. Our results suggest that the current overestimation of model generalizability and the inadequacies of scaling assumptions necessitate a reevaluation of what constitutes a genuinely demanding OOD task. Establishing more meaningful and challenging test scenarios will help identify and address the gaps between current capabilities and truly generalizable ML models for materials science.

Methods

Datasets and models

We use the snapshots of the JARVIS, Materials Project, and OMQD databases used in our previous study[47], available on Zenodo at https://zenodo.org/records/8200972. The snapshots correspond the JARVIS 2022.12.12 version, the Materials Project 2021.11.10 version, and the OQMD v1.6 version (published in November 2023), and they have been preprocessed to remove materials with formation energies larger than 5 eV/atom.

Performing hyperparameter tuning for each of over 700 OOD tasks for every model would be computationally impractical. In addition, hyperparameter tuning techniques are developed mainly to optimize the in-distribution rather than OOD performance. Our strategy is to perform the hyperparameter tuning on an in-distribution validation set randomly sampled from the whole dataset; then we apply the same optimal hyperparameters to every OOD task. Here, the criterion for the optimal hyperparameters is based on not only model performance but also training cost, to make the training amenable to extensive benchmarking. Our objective is to demonstrate the overall generalization performance on various OOD tasks rather than ensure optimal performance on each OOD task for every model. We expect future work to develop techniques to optimize OOD performance by focusing on a small set of representative tasks selected based on this study.

For tree-based models, we use a descriptor set that consists of 145 compositional features[52] extracted from chemical formula and 128 structural features[53] extracted from crystal structures. For the RF model[54], we use 100 estimators, 30% of the features for the best splitting, and default settings for other hyperparameters. For the XGB model[34], we use 4 parallel boosting trees; for each boosting tree, we use 1000 estimators, a learning rate of 0.25, an L1 (L2) regularization strength of 0.01 (0.1), and the histogram tree grow method; we set the subsample ratio of training instances to 0.85, the subsample ratio of columns to 0.3 when constructing each tree, and the subsample ratio of columns to 0.5 for each level.

The deep learning models (ALIGNN, GMP, and LLM-Prop) are trained as follows. ALIGNN models are trained with 2 ALIGNN layers, 2 GCN layers, a batch size of 16, 25 epochs, and layer normalization, while keeping other hyperparameters the same as in the original implementation[37]. LLM-Prop models are trained with a batch size of 64, a max token length of 1500, and a drop rate of 0.5, while keeping the tokenizer, preprocessing strategy, and other hyperparameters as in the original implementation[38]. For GMP models, we use the GMP-featurizer package[55] to calculate the Gaussian Multipole features by using 40 radial probes, max 4th-order angular probes, and a Gaussian width of 1.5. The Kresse-Joubert projector augmented-wave pseudopotentials are used to calculated electron density[56]. The model architecture consists of a compressor neural network (hidden layer dimension [512,256,128,64,32]), a Set2Set neural network with 5 hidden layers to convert sets of varying sizes to a fixed size tensor[57], and a predictor neural network (hidden layer dimension [256,128,64,32,16]). GMP models are trained with 200 epochs, a batch size of 128, and a ReduceLROnPlateau scheduler to tune the learning rate on a validation set (10% of the initial training set).

SHAP analysis for leave-one-element-out tasks

The following procedure is proposed to identify the sources of biases in the models trained for the leave-one-element-out tasks. For each task, we use the model trained without element X to make predictions on the whole dataset. Then, we train a second model on the prediction errors of the first model on the whole dataset. By doing so, the second model learns to correct the bias in the first model which is the one used in the leave-one-element-out task. Finally, we calculate the SHAP feature contributions of the second trained model for all data with element X, interpreted as the correction for the feature contribution of the first model for the out-of-distribution materials. Here the first and second models are chosen to be XGB because it uses two types of interpretable features: the compositional features are chemical attributes computed from chemical formula[52], whereas the structural features are characteristics of the local atomic environment calculated from crystal structures[53]. We apply the above procedure to calculate the SHAP contributions for compositional and structural features of the second XGB model and aggregate the contributions into two categories.

UMAP dimension reduction

The 2-component UMAP representation[48] are built using a local neighborhood size of 50 and a minimum distance of 0.1. For deep learning models, the embeddings are used as inputs to construct the UMAP plots. For tree-based models, we sequentially drop highly correlated features with a Pearson correlation threshold of 0.7 and then construct the UMAP plots based on the final feature set. We use the kernel-density estimation method implemented in SciPy[58] to calculate the probability density of training data for every test data point. A probability density threshold of 0.001 is used to separate between in-domain and out-of-domain test data.

Data availability

We use the same snapshots as in our prior work[47] for the JARVIS, MP, and OQMD datasets, which can be found on Zenodo at https://zenodo.org/records/8200972.

Code availability

The code used for ML training and analysis is available on GitHub at https://github.com/mathsphy/probing-ood.

Received: 3 December 2024; Accepted: 30 December 2024;
Published online: 11 January 2025

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. Carleo, G. et al. Machine learning and the physical sciences. *Rev. Mod. Phys.* **91**, 045002 (2019).
3. Krenn, M. et al. On scientific understanding with artificial intelligence. *Nat. Rev. Phys.* **4**, 761–769 (2022).
4. Huang, B., von Rudorff, G. F. & von Lilienfeld, O. A. The central role of density functional theory in the ai age. *Science* **381**, 170–175 (2023).
5. Liu, J. et al. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624* (2021).
6. Merchant, A. et al. Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (2023).
7. Yang, H. et al. Mattersim: a deep learning atomistic model across elements, temperatures and pressures. *arXiv preprint arXiv:2405.04967* (2024).
8. Batatia, I. et al. A foundation model for atomistic materials chemistry. *arXiv preprint arXiv:2401.00096* (2023).
9. Schmidt, J. et al. Machine-Learning-Assisted Determination of the Global Zero-Temperature Phase Diagram of Materials. *Adv. Mater.* **35**, 2210788 (2023).
10. Deng, B. et al. Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nat. Mach. Intell.* **5**, 1031–1041 (2023).

11. Chen, C. & Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comput. Sci.* **2**, 718–728 (2022).
12. Li, K., Choudhary, K., DeCost, B., Greenwood, M. & Hattrick-Simpers, J. Efficient first principles based modeling via machine learning: from simple representations to high entropy materials. *J. Mater. Chem. A* **12**, 12412–12422 (2024).
13. Chen, W. et al. A map of single-phase high-entropy alloys. *Nat. Commun.* **14**, 2856 (2023).
14. Bokas, G. B. et al. Unveiling the thermodynamic driving forces for high entropy alloys formation through big data ab initio analysis. *Scr. Mater.***202**, 114000 (2021).
15. Chen, H.-L., Mao, H. & Chen, Q. Database development and calphad calculations for high entropy alloys: Challenges, strategies, and tips. *Mater. Chem. Phys.* **210**, 279–290 (2018).
16. 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).
17. 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).
18. Sheridan, R. P. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *J. Chem. Inf. Model.* **53**, 783–790 (2013).
19. Jablonka, K. M., Rosen, A. S., Krishnapriyan, A. S. & Smit, B. An ecosystem for digital reticular chemistry. *ACS Cent. Sci*. **9**, 563–581 (2023).
20. Riebesell, J. et al. Matbench discovery–an evaluation framework for machine learning crystal stability prediction. *arXiv preprint arXiv:2308.14920* (2023).
21. Omee, S. S., Fu, N., Dong, R., Hu, M. & Hu, J. Structure-based out-of-distribution (ood) materials property prediction: a benchmark study. *npj Comput. Mater.* **10**, 144 (2024).
22. Zhang, H., Chen, W. W., Rondinelli, J. M. & Chen, W. Et-al: Entropy-targeted active learning for bias mitigation in materials data. *Appl. Phys. Rev.* **10**, 021403 (2023).
23. Wu, Z. et al. Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).
24. Chanussot, L. et al. Open Catalyst 2020 (OC20) Dataset and Community Challenges. *ACS Catal.* **11**, 6059–6072 (2021).
25. Tran, R. et al. The Open Catalyst 2022 (OC22) Dataset and Challenges for Oxide Electrocatalysts. *ACS Catal.* **13**, 3066–3084 (2023).
26. Frey, N. C. et al. Neural scaling of deep chemical models. *Nat. Mach. Intell.* **5**, 1297–1305 (2023).
27. Kaplan, J. et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).
28. McKenzie, I. R. et al. Inverse scaling: When bigger isn't better. *arXiv preprint arXiv:2306.09479* (2023).
29. Choudhary, K. et al. The joint automated repository for various integrated simulations (jarvis) for data-driven materials design. *npj Comput. Mater.* **6**, 173 (2020).
30. Wines, D. et al. Recent progress in the jarvis infrastructure for next-generation data-driven materials design. *Appl. Phys. Rev.* **10** (2023). https://doi.org/10.1063/5.0159299.
31. Jain, A. et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
32. 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).
33. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
34. Chen, T. & Guestrin, C. XGBoost. In *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min*., 785–794 (ACM, New York, NY, USA, 2016).
35. Ward, L. et al. Matminer: an open source toolkit for materials data mining. *Computational Mater. Sci.* **152**, 60–69 (2018).
36. Lei, X. & Medford, A. J. A universal framework for featurization of atomistic systems. *J. Phys. Chem. Lett.* **13**, 7911–7919 (2022).
37. Choudhary, K. & DeCost, B. Atomistic line graph neural network for improved materials property predictions. *npj Comput. Mater.* **7**, 185 (2021).
38. Rubungo, A. N., Arnold, C., Rand, B. P. & Dieng, A. B. Llm-prop: predicting physical and electronic properties of crystalline solids from their text descriptions. *arXiv preprint arXiv:2310.14029* (2023).
39. Deng, B. et al. Overcoming systematic softening in universal machine learning interatomic potentials by fine-tuning. *arXiv preprint arXiv:2405.07105* (2024).
40. Choudhary, K. & Sumpter, B. G. Can a deep-learning model make fast predictions of vacancy formation in diverse materials? *AIP Adv.* **13**, 095109 (2023).
41. Lundberg, S. M. et al. From local explanations to global understanding with explainable ai for trees. *Nat. Mach. Intell.* **2**, 2522–5839 (2020).
42. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
43. Li, K., DeCost, B., Choudhary, K., Greenwood, M. & Hattrick-Simpers, J. A critical examination of robustness and generalizability of machine learning prediction of materials properties. *npj Comput. Mater.* **9**, 55 (2023).
44. Schrier, J., Norquist, A. J., Buonassisi, T. & Brgoch, J. In pursuit of the exceptional: research directions for machine learning in chemical and materials science. *J. Am. Chem. Soc.* **145**, 21699–21716 (2023).
45. Goodall, R. E., Parackal, A. S., Faber, F. A., Armiento, R. & Lee, A. A. Rapid discovery of stable materials by coordinate-free coarse graining. *Sci. Adv.* **8**, eabn4117 (2022).
46. Ganose, A. M. & Jain, A. Robocrystallographer: automated crystal structure text descriptions and analysis. *MRS Commun.* **9**, 874–881 (2019).
47. Li, K. et al. Exploiting redundancy in large materials datasets for efficient machine learning with less data. *Nat. Commun.* **14**, 7283 (2023).
48. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).
49. Faber, F. A. et al. Prediction errors of molecular machine learning models lower than hybrid dft error. *J. Chem. theory Comput.* **13**, 5255–5264 (2017).
50. Wei, J. et al. *Emergent Abilities of Large Language Models* (TMLR, 2022).
51. Biever, C. Chatgpt broke the turing test-the race is on for new ways to assess ai. *Nature* **619**, 686–689 (2023).
52. 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).
53. Ward, L. et al. Including crystal structure attributes in machine learning models of formation energies via voronoi tessellations. *Phys. Rev. B* **96**, 024104 (2017).
54. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
55. Lei, X. & Montoya, J. Gmp-featurizer: a parallelized python package for efficiently computing the gaussian multipole features of atomic systems. *J. Open Source Softw*. **8**, 5476 (2023).
56. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. b* **59**, 1758 (1999).
57. Vinyals, O., Bengio, S. & Kudlur, M. Order matters: Sequence to sequence for sets. *arXiv preprint arXiv:1511.06391* (2015).
58. Virtanen, P. et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nat. methods* **17**, 261–272 (2020).

59. Riebesell, J., Goodall, R. & Baird, S. G. Pymatviz: visualization toolkit for materials informatics (2022). https://github.com/janosh/pymatviz.

Acknowledgements

This research was undertaken thanks in part to funding provided to the University of Toronto's Acceleration Consortium from the Canada First Research Excellence Fund (Grant number: CFREF-2022-00042). Computational resources were provided by the Calcul Quebec, Westgrid, and Compute Ontario consortia in the Digital Research Alliance of Canada, and the Acceleration Consortium at the University of Toronto. We thank Rhys Goodall for constructive comments and discussion.

Author contributions

K.L. and J.H.-S. conceived the idea. K.L. designed the project, performed the investigation, and wrote the manuscript. A.N.R., X.L., and D.P. assisted in the investigation. J.H.-S. supervised the project. K.C., B.D., and A.B.D contributed to the discussion and revision.

Competing interests

There are no conflicts of interest to declare. Certain commercial products or company names are identified here to describe our study adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the products or names identified are necessarily the best available for the purpose.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43246-024-00731-w.

Correspondence and requests for materials should be addressed to Kangming Li or Jason Hattrick-Simpers.

Peer review information *Communications materials* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: John Plummer. A peer review file is available.