NOT AS SIMPLE AS WE THOUGHT: A RIGOROUS EXAMINATION OF DATA AGGREGATION IN MATERIALS INFORMATICS

A PREPRINT

Federico Ottomano*
Department of Computer Science
University of Liverpool
Liverpool, L69 3DR, UK
federico.ottomano@liverpool.ac.uk

Giovanni De Felice*
Department of Computer Science
University of Liverpool
Liverpool, L69 3DR, UK
g.de-felice@liverpool.ac.uk

Vladimir V. Gusev
Department of Computer Science
University of Liverpool
Liverpool, L69 3DR, UK

Taylor D. Sparks†
Department of Material Science and Engineering
University of Utah
Salt Lake City, Utah 84112, USA

August 8, 2023

ABSTRACT

Recent Machine Learning (ML) developments have opened new perspectives on accelerating the discovery of new materials. However, in the field of materials informatics, the performance of ML estimators is heavily limited by the nature of the available training datasets, which are often severely restricted and unbalanced. Among practitioners, it is usually taken for granted that more data corresponds to better performance. Here, we investigate whether different ML models for property predictions benefit from the aggregation of large databases into smaller repositories. To do this, we probe three different aggregation strategies prioritizing training size, element diversity, and composition diversity. For classic ML models, our results consistently show a reduction in performance under all the considered strategies. Deep Learning models show more robustness, but most changes are not significant. Furthermore, to assess whether this is a consequence of a distribution mismatch between datasets, we simulate the data acquisition process of a single dataset and compare a random selection with prioritizing chemical diversity. We observe that prioritizing composition diversity generally leads to a slower convergence toward better accuracy. Overall, our results suggest caution when merging different data sources and discourage a biased acquisition of novel chemistries when building a training dataset.

Keywords Materials Informatics · Machine Learning · Data Aggregation

1 Introduction

In recent years, following the increased availability of computational material databases [Jain et al., 2013, Kirklin et al., 2015, Blokhin and Villars, 2018], Machine Learning (ML) and data-driven approaches have opened new frontiers for accelerating materials discovery. These aim at overcoming the limitations imposed by the expensive physical simulations adopted in density functional theory (DFT), which allow only for a narrow exploration of the chemical space. Furthermore, DFT suffers from systematic errors due to numerical approximations occurring in any solver [Schleder et al., 2019]. Besides the computational advantages, ML models can also discover novel patterns that are otherwise hard to identify by only leveraging traditional chemical knowledge [Mansouri Tehrani et al., 2018, Tewari et al., 2020].

*Equal contribution
†Work done while at Liverpool

While, on the one hand, such approaches have shown remarkable success [Wang et al., 2022, Khakurel et al., 2021, Cao et al., 2019, Li et al., 2019], it is important to acknowledge their limitations and potential downsides. One significant challenge is the difficulty in assessing the quality of performance outside the distribution of training data. As it happens, ML models can learn patterns that are too specific to the training data and fail to extrapolate to unseen data (overfitting). Such approaches also heavily rely on the size of the training data and scarcity can lead to models with limited capabilities or inaccurate predictions. Experimental datasets of specific chemical properties, such as thermoelectric properties [Gaultois et al., 2013, Katsura et al., 2019], are very unbalanced and rare throughout the literature. This is a consequence of the popular material repositories predominantly relying on DFT calculations [Curtarolo et al., 2012, Jain et al., 2013, Kirklin et al., 2015], which tend to provide a constrained selection of chemical attributes. This hampers the ability to effectively target specific material classes. Different approaches have been adopted to mitigate the natural presence of bias in materials data. LOCO-CV [Meredig et al., 2018] has been proposed as a modification of the standard KFold evaluation strategy to measure the extrapolation error of machine learning models on unseen chemical clusters. Furthermore, an entropy-based metric has been recently proposed to mitigate the imbalance of a crystal structures dataset by improving the diversity of underrepresented crystal systems [Zhang et al., 2023]

On a general level, three main strands are usually considered to improve the predictive accuracy of ML models:

- **Better model:** in a *model-centric* approach, the primary emphasis is on creating better algorithms to extract valuable insights from the available data. Lately, especially in the area of Deep Learning, this is mostly done by designing novel architectures. Here, a popular approach is to strengthen the algorithm by tailoring the architecture to the specific application, usually by leveraging symmetries that exist in the data, e.g. crystal structures [Klipfel et al., 2022];
- **Better data:** in a *data-centric* approach, the focus is instead on the quality of the inputs for the model. Notable examples are the refinement of the measurement strategy and preprocessing, e.g. data balancing or outlier filtering. Also falling under this category are methods that leverage domain knowledge to design better data features, more commonly known as 'feature engineering' Ward et al. [2016] Lee et al. [2023]
- **More data:** in this branch of the *data-centric* approach, the attention is shifted to increasing the number of data points. This is generally considered to be more significant in view of a better-performing statistical model [Goodfellow et al., 2016, Zha et al., 2023] and a compelling alternative to vast domain knowledge [Murdock et al., 2020].

As this last point is generally taken for granted, little attention has been dedicated to it in the literature on materials informatics. Given this and the limited availability of experimental data, it is natural for practitioners to consider the aggregation of diverse data sources. However, data aggregation in materials informatics presents unique challenges compared to most ML datasets [Himanen et al., 2019]. Unlike many other domains, chemical datasets can often be unbalanced, small in size, or collected under diverse experimental conditions. Additionally, the ranges of values can be considerably large, and the data space can exhibit pronounced discontinuities. As an example, in the context of thermoelectric materials, the introduction of chemistry defects through *doping* can lead to substantial alterations in electronic properties [Kdasap, 2002, Na et al., 2021]. These challenges emphasize the need for careful consideration when aggregating different datasets in materials informatics research.

In this work, we deepen the aggregation of different datasets reporting chemical formulas and associated properties. In particular, we study whether the predictive accuracy of different ML models can benefit from the aggregation of local repositories with databases with larger availability. In order to do that, we consider three different aggregation strategies in which we prioritize training size, element diversity, and composition diversity. Our main findings are summarized as follows:

- We report that classical ML methods undergo a noticeable degradation in accuracy subsequent to a concatenation with popular databases. Additionally, we show that the incorporation of data points focusing on maximizing chemical diversity also leads to a decline in the performance of such models.
- We establish that Deep Learning (DL) models exhibit a much higher level of robustness. However, the majority of changes in the accuracy, whether improvements or degradations, are not statistically significant.
- We simulate the data acquisition process on a single dataset by utilizing both the DiSCoVeR algorithm and a random acquisition approach. We proceed to compare the results obtained from these two methods on both a randomly generated test set and a biased test set, which was previously constructed using DiSCoVeR. Notably, our observations demonstrate that a biased acquisition strategy for new stoichiometries deteriorates the learning process of the model, regardless of the test set scenario.

The rest of the paper is structured as follows. In Sec. 2, we present the datasets and the downstream ML models that we use to support our claims; in Sec. 3, we evaluate different dataset aggregation strategies and discuss results; in Sec. 4,

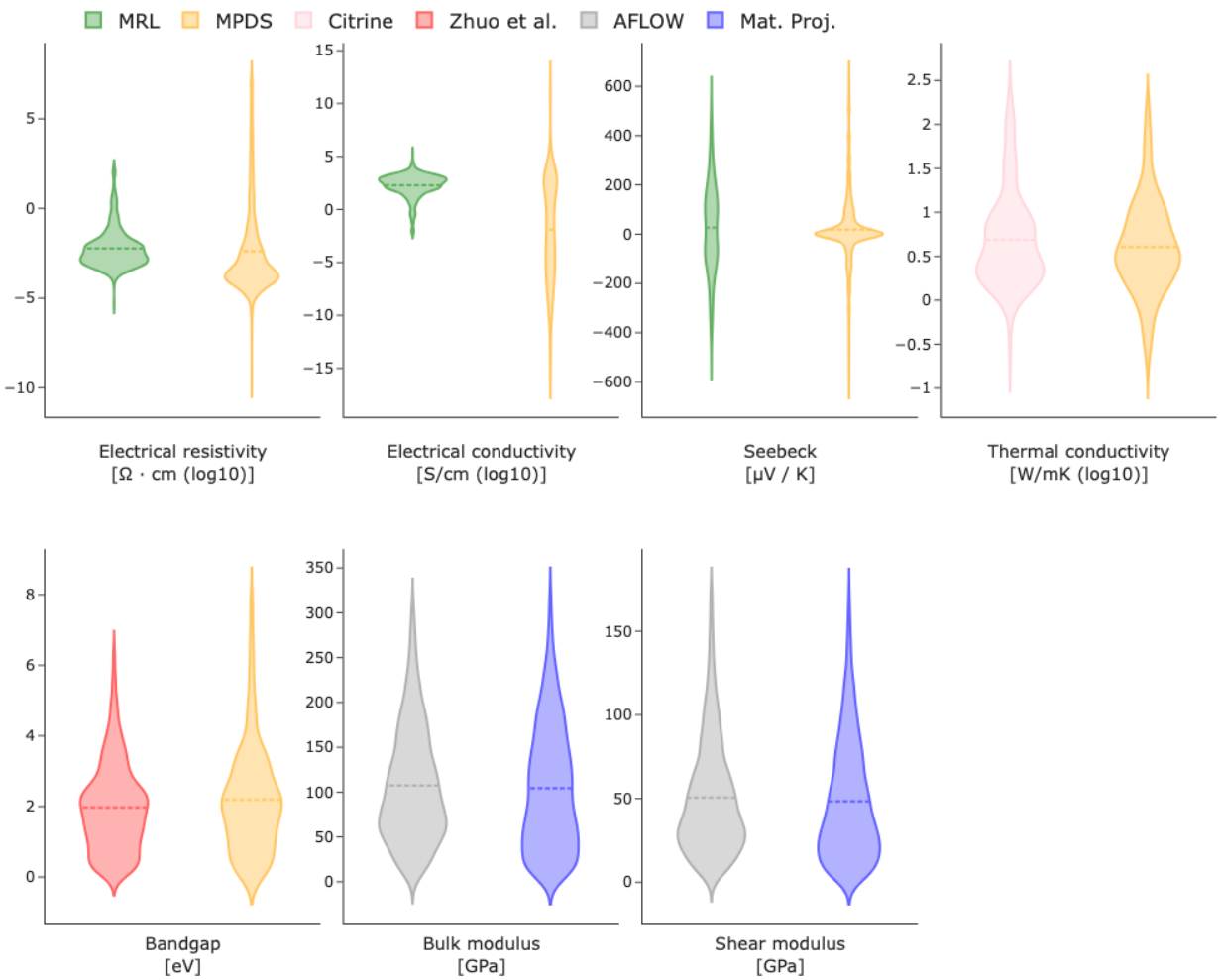

Figure 1: Violin plots of all pairs of datasets. Notably, archives' data covers a wider range of the target property.

we present the result about prioritizing chemical diversity in progressive data acquisition; Sec. 5 concludes the paper with the final remarks.

2 Preliminaries

2.1 Datasets

In our experimental setting, we consider eight different datasets for eight different chemical properties:

- electrical resistivity, electrical conductivity and Seebeck coefficient from the MRL dataset [Gaultois et al., 2013];
- thermal conductivity from the Citrine platform [Mullin, 2017];
- Band gap from Zhuo et al. [2018];
- DFT calculated Bulk modulus and Shear modulus from AFLOW [Curtarolo et al., 2012].

For each property, the respective dataset is aggregated with experimental data coming from the Materials Platform for Data Science (MPDS)Blokhin and Villars [2018], retrieved by using the provided API. MPDS is one of the largest resources currently available for material scientists. It leverages the extensive data available in the Pauling File Villars et al. [2004], a comprehensive database of materials information reporting crystal structures chemical compositions and

Property	units	dataset	nature	size	minimum	maximum	aggregation label
Electrical resistivity	($\Omega\cdot$ cm)	MRL	exp.	400	-5.3 (log)	2.17 (log)	A
		MPDS	exp.	6352	-10 (log)	7.6 (log)	B
Electrical conductivity	(S/cm)	MRL	exp.	401	-2.17 (log)	5.3 (log)	A
		MPDS	exp.	1489	-15 (log)	11 (log)	B
Thermal conductivity	(W/mK)	Citrine	exp.	219	-0.70 (log)	2.37 (log)	A
		MPDS	exp.	878	-0.85 (log)	2.30 (log)	B
Seebeck coefficient	($\mu V/K$)	MRL	exp.	416	-476.68	525.2	A
		MPDS	exp.	2050	-640	674	B
Band gap	(eV)	Zhuo	exp.	2287	0.02	6.43	A
		MPDS	exp.	918	2×10^{-4}	8	B
Bulk modulus	(GPa)	AFLOW	calc.	4822	0.66	312.94	A
		MP	calc.	6221	0.73	324.70	B
		MPDS	exp.	1367	2×10^{-7}	379.4	B
Shear modulus	(GPa)	AFLOW	calc.	4747	0.65	175.81	A
		MP	calc.	6073	0	174.12	B
		MPDS	exp.	358	0.36	293	B

Table 1: Dataset details. Datasets labeled with 'A' are the ones that will be increased through aggregation (denoted 'A') with points from dataset 'B'

phase diagrams, to enable efficient exploration, analysis, and modeling of materials. For the two calculated datasets, we also consider the aggregation with calculated data from the Materials Project (MP) database Jain et al. [2013].

Several steps of preprocessing are applied to the raw datasets. First, we filter our values outside 15 K of the room temperature, noble gases and radio-isotopes ($A > 93$). If input duplicates are found, we store their median. Finally, we discard all the data points outside 3 standard deviations from the overall mean. Fig. 1 compares the distributions of the mentioned local repositories with the corresponding dataset from which we gather the additional data. Except for sporadic cases, we observe a general agreement in shape between the considered pairs of datasets. As expected, local repositories generally cover a smaller range of values with respect to the data gathered from the archives. Further details about sizes and value range for all datasets are given in Tab. 1. With the only exception of the *band gap* datasets pair, the size of the archives' data are always larger.

2.2 ML estimators

Throughout the paper, we evaluate data aggregation by comparing the performance of different ML estimators before and after increasing the dataset size. These models include baselines and SOTA for chemical properties prediction given the stoichiometry, with representatives of both classical and Deep Learning (DL) approaches. In more detail, we consider *ridge regression* as a simple baseline model, *random forest regression* as a robust model for low-data regimes [Murdock et al., 2020], *Roost* [Goodall and Lee, 2020] as a DL model based on graph representations and *CrabNet* [Wang et al., 2021] as a transformer-based approach and representative of the SOTA. Performance is assessed through the ordinary procedure of train-test split and on the *mean absolute error* (MAE), a typical metric used for regression that quantifies the absolute deviation between models' predictions and true corresponding values.

Finally, we adopt a classification task, inspired by recent work investigating machine learning extrapolation capabilities in materials informatics [Kauwe et al., 2020]. We first label material instances with the corresponding property in the top 20% of the distribution as extraordinary. Here, the term 'top' is defined based on the specific property under consideration. In some cases, 'top' refers to the highest values, while in other cases, 'top' denotes the lowest values, depending on the tail of the distribution. We finally consider *logistic regression* as a simple binary classifier to differentiate ordinary from extraordinary materials.

The regularization strength for the ridge regression and logistic regression model is optimized via Cross Validation (CV) from a range of logarithmically spaced values between $[10^{-4}, 10^{3}]$. Finally, results are averaged across 5 iterations with different random seeds controlling the initialization of all stochastic components.

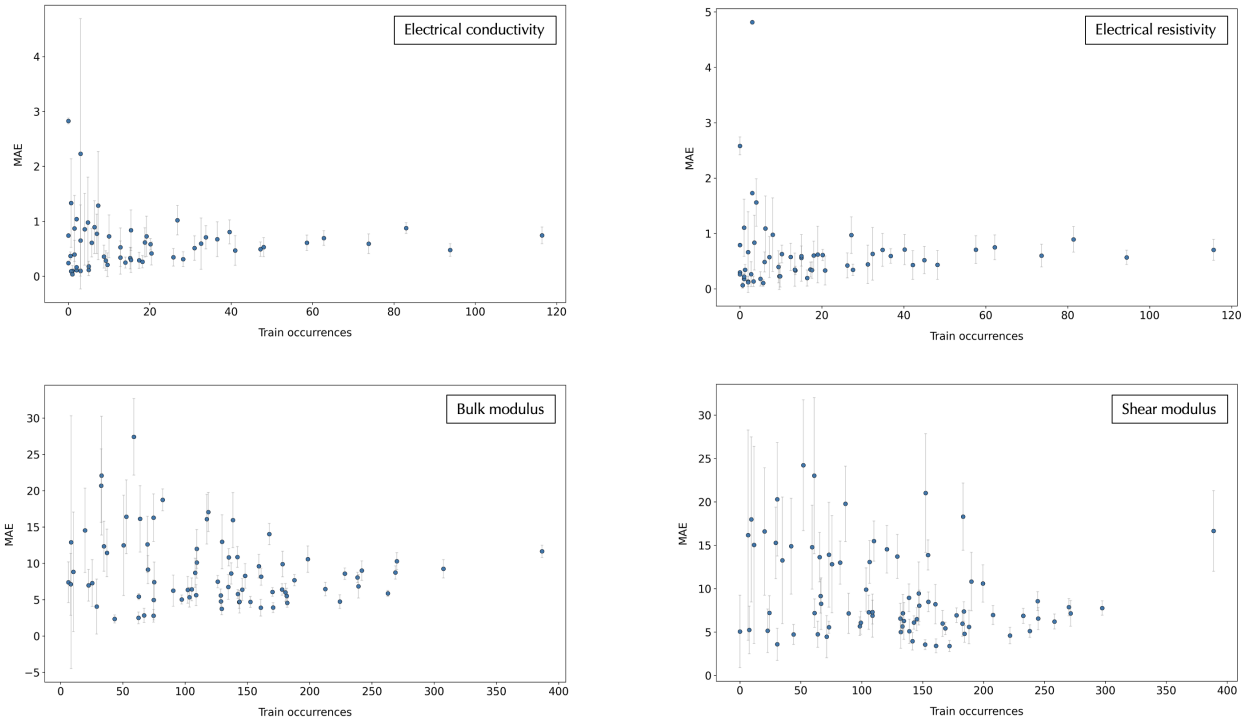

Figure 2: Imbalance of MAE. For different datasets A in the baseline setting (see Sec. 3), the Mean absolute error (MAE) of a Random Forest model is plotted against the occurrences of individual elements in compositions of the training set. Error bars represent 1σ over 5 different random seeds. It can be observed how larger errors and deviations are mostly found in correspondence with low train occurrences. Similar patterns can be observed for most other properties and models.

3 A-B data aggregation

As our first and main experiment, we consider the aggregation of each dataset A with data points collected from the respective dataset B. To assess the benefits of the aggregation, we first evaluate the performance of ML estimators before integrating any new data points; this will be indicated as *baseline* setting. This is done, as usual, by training on a subset (80%) of dataset A and computing prediction errors on the corresponding test set (20%). For DL models, 10% of the training size is reserved for a validation set. In the aggregation process, data points collected from B only increase the size of the original training set (and validation, for DL) of A. For consistency, performance is always assessed on the original test set of A. We consider three different aggregation strategies:

Concatenation: a simple concatenation of all points from the train set of A with the whole dataset B. Duplicated instances are removed by taking the median across reported target properties. The primary advantage of this strategy is that the size of the resulting dataset is maximized. This is generally believed to strengthen the robustness of the estimators and potentially discover new patterns. However, a possible drawback is a saturation effect which arises from the compounded presence of redundant data points, hindering model learning and generalization. In particular, different associated values and experimental conditions may have the overall effect of increasing the degree of noise in the dataset.

Element-focused concatenation: to introduce the next strategy, we consider the following illustrative example. In Fig. 2, the mean average error (MAE) of a Random Forest model [Breiman, 2001] is plotted against the occurrences of the chemical elements in compositions of the training dataset A. Two main patterns can be observed: an increase in MAE as fewer representatives are available at the training stage, and an increase in variance (similar patterns are also observed with other models). As a consequence, one might expect to improve the overall accuracy by populating chemical regions with fewer representatives, while, at the same time, avoiding the introduction of noise that would alter the performances on the rest.

Datasets	Ridge regression (regr.)				Random Forest (regr.)			
	Baseline	Concat	ElemConc	DiSCoVeR	Baseline	Concat	ElemConc	DiSCoVeR
Elec. res.	$0.69_{\pm 0.07}$	$1.21_{\pm 0.03}$	$0.75_{\pm 0.04}$	$1.04_{\pm 0.06}$	$0.62_{\pm 0.05}$	$1.11_{\pm 0.06}$	$0.7_{\pm 0.1}$	$1.04_{\pm 0.06}$
Elec. cond.	$0.71_{\pm 0.04}$	$3.73_{\pm 0.06}$	$1.1_{\pm 0.2}$	$1.9_{\pm 0.4}$	$0.65_{\pm 0.08}$	$3.8_{\pm 0.2}$	$1.3_{\pm 0.5}$	$1.7_{\pm 0.3}$
Therm. cond.	$0.25_{\pm 0.02}$	$0.38_{\pm 0.01}$	$0.28_{\pm 0.03}$	$0.31_{\pm 0.03}$	$0.28_{\pm 0.03}$	$0.36_{\pm 0.02}$	$0.25_{\pm 0.02}$	$0.31_{\pm 0.03}$
Seebeck	$106_{\pm 13}$	$123_{\pm 7}$	$103_{\pm 7}$	$104_{\pm 8}$	$83_{\pm 9}$	$109_{\pm 8}$	$83_{\pm 8}$	$98_{\pm 5}$
Band gap	$0.53_{\pm 0.01}$	$0.55_{\pm 0.01}$	$0.53_{\pm 0.01}$	$0.53_{\pm 0.01}$	$0.43_{\pm 0.02}$	$0.46_{\pm 0.02}$	$0.42_{\pm 0.02}$	$0.42_{\pm 0.01}$
Bulk modulus (c)	$21.1_{\pm 0.6}$	$22.7_{\pm 0.8}$	$21.1_{\pm 0.6}$	$21.6_{\pm 0.7}$	$13_{\pm 1}$	$16_{\pm 1}$	$13_{\pm 1}$	$14_{\pm 1}$
Shear modulus (c)	$15.0_{\pm 0.3}$	$15.3_{\pm 0.5}$	$15.0_{\pm 0.3}$	$15.1_{\pm 0.4}$	$10.3_{\pm 0.5}$	$10.6_{\pm 0.4}$	$10.2_{\pm 0.5}$	$10.3_{\pm 0.6}$
Bulk modulus (e)	$21.1_{\pm 0.6}$	$22.7_{\pm 0.4}$	$21.2_{\pm 0.6}$	$21.4_{\pm 0.5}$	$13_{\pm 1}$	$15.9_{\pm 0.7}$	$13_{\pm 1}$	$13_{\pm 1}$
Shear modulus (e)	$15.0_{\pm 0.3}$	$15.4_{\pm 0.2}$	$15.0_{\pm 0.2}$	$15.0_{\pm 0.3}$	$10.3_{\pm 0.5}$	$10.6_{\pm 0.5}$	$10.3_{\pm 0.4}$	$10.4_{\pm 0.5}$
Datasets	**Roost (regr.)**				**CrabNet (regr.)**			
	Baseline	Concat	ElemConc	DiSCoVeR	Baseline	Concat	ElemConc	DiSCoVeR
Elec. res.	$0.56_{\pm 0.05}$	$0.71_{\pm 0.08}$	$0.62_{\pm 0.08}$	$0.6_{\pm 0.1}$	$0.60_{\pm 0.04}$	$0.60_{\pm 0.08}$	$0.67_{\pm 0.04}$	$0.63_{\pm 0.06}$
Elec. cond.	$0.6_{\pm 0.1}$	$0.8_{\pm 0.2}$	$0.60_{\pm 0.05}$	$0.8_{\pm 0.1}$	$0.60_{\pm 0.04}$	$0.8_{\pm 0.2}$	$0.61_{\pm 0.07}$	$0.63_{\pm 0.08}$
Therm. cond.	$0.21_{\pm 0.04}$	$0.24_{\pm 0.03}$	$0.26_{\pm 0.07}$	$0.3_{\pm 0.1}$	$0.20_{\pm 0.03}$	$0.19_{\pm 0.03}$	$0.20_{\pm 0.03}$	$0.20_{\pm 0.02}$
Seebeck	$58_{\pm 6}$	$66_{\pm 7}$	$64_{\pm 8}$	$69_{\pm 9}$	$68_{\pm 10}$	$60_{\pm 6}$	$65_{\pm 7}$	$72_{\pm 7}$
Band gap	$0.42_{\pm 0.02}$	$0.41_{\pm 0.01}$	$0.47_{\pm 0.06}$	$0.40_{\pm 0.03}$	$0.38_{\pm 0.01}$	$0.37_{\pm 0.01}$	$0.39_{\pm 0.01}$	$0.38_{\pm 0.01}$
Bulk modulus (c)	$10.7_{\pm 0.7}$	$10.0_{\pm 1}$	$11_{\pm 1}$	$11.0_{\pm 0.7}$	$9.0_{\pm 0.7}$	$8.6_{\pm 0.8}$	$9.2_{\pm 0.9}$	$8.8_{\pm 0.7}$
Shear modulus (c)	$10.5_{\pm 0.6}$	$8.4_{\pm 0.2}$	$11_{\pm 1}$	$10.4_{\pm 0.2}$	$8.7_{\pm 0.2}$	$7.3_{\pm 0.1}$	$8.8_{\pm 0.3}$	$8.7_{\pm 0.5}$
Bulk modulus (e)	$10.7_{\pm 0.7}$	$12_{\pm 1}$	$11.0_{\pm 0.6}$	$11_{\pm 2}$	$9.0_{\pm 0.7}$	$9.8_{\pm 0.7}$	$9.1_{\pm 0.7}$	$9.3_{\pm 0.7}$
Shear modulus (e)	$10.5_{\pm 0.6}$	$10.6_{\pm 0.5}$	$10.4_{\pm 0.3}$	$10.4_{\pm 0.5}$	$8.7_{\pm 0.2}$	$9.0_{\pm 0.2}$	$8.8_{\pm 0.2}$	$9.0_{\pm 0.6}$

Datasets	Logistic regression (class.)			
	Baseline	Concat	ElemConc	DiSCoVeR
Elec. res.	$0.82_{\pm 0.05}$	$0.58_{\pm 0.04}$	$0.79_{\pm 0.04}$	$0.55_{\pm 0.04}$
Elec. cond.	$0.85_{\pm 0.04}$	$0.82_{\pm 0.03}$	$0.85_{\pm 0.03}$	$0.82_{\pm 0.04}$
Therm. cond.	$0.91_{\pm 0.06}$	$0.86_{\pm 0.05}$	$0.90_{\pm 0.05}$	$0.87_{\pm 0.06}$
Seebeck	$0.85_{\pm 0.02}$	$0.82_{\pm 0.04}$	$0.84_{\pm 0.07}$	$0.84_{\pm 0.05}$
Band gap	$0.893_{\pm 0.003}$	$0.87_{\pm 0.01}$	$0.89_{\pm 0.01}$	$0.89_{\pm 0.01}$
Bulk modulus (c)	$0.91_{\pm 0.01}$	$0.90_{\pm 0.01}$	$0.913_{\pm 0.003}$	$0.91_{\pm 0.01}$
Shear modulus (c)	$0.88_{\pm 0.01}$	$0.81_{\pm 0.02}$	$0.88_{\pm 0.01}$	$0.88_{\pm 0.01}$
Bulk modulus (e)	$0.91_{\pm 0.01}$	$0.91_{\pm 0.01}$	$0.912_{\pm 0.004}$	$0.912_{\pm 0.008}$
Shear modulus (e)	$0.88_{\pm 0.01}$	$0.89_{\pm 0.01}$	$0.88_{\pm 0.01}$	$0.881_{\pm 0.004}$

Table 2: For each model-dataset pair, the MAE is reported before (Baseline) and after 3 different data aggregation strategies (Concat, ElemConc, DiSCoVeR). For calculated datasets A, experiments are repeated using calculated (MP) and experimental (MPDS) dataset B. A green color represents an improvement above one standard deviation with respect to the *Baseline* setting, yellow indicates equivalent performance (variations could simply be attributed to random fluctuations) and red denotes a worsening above one standard deviation. Overall, different aggregation strategies fail to improve performance.

In order to do this, we identify the $k = 5$ elements with the smallest prevalence in A and, for each, we collect $n = 10$ data points at random containing such element from dataset B. This addresses the weakness of previous concatenation strategy. Although targeting specific classes of elements with a narrow prevalence may be attractive, the presence or absence of a certain single element is not a good proxy for the chemical composition. In fact, this approach ignores any high-level relationship between the involved stoichiometries.

DiSCoVeR: DiSCoVeR [Baird et al., 2022] algorithm is a recently proposed ensemble of machine learning methods aimed at facilitating the identification of chemistries lying at the intersection between novelty and performance. In

practice, DiSCoVeR can be used to provide novelty scores of a given pool of data with respect to another and it was recently employed to identify new chemically novel high-temperature superconductors [Seegmiller et al., 2023] The framework employed by DiSCoVeR is structured as follows: first, a distance matrix between all compositions in the dataset is computed by using the *Element Movers Distance* [Hargreaves et al., 2020], a proposed metric which takes into account chemical similarities; subsequently, the obtained distance matrix is used to obtain 2D UMAP embeddings of all data points (A $\cup$ B); the likelihood of each point in B is computed with respect to the density of A, returning a quantitative measure of novelty (*density score*). Compositions in regions of low density are assigned with a higher novelty score. In the original DiSCoVeR implementation, a complementary score *target score* is calculated based on a specific property of interest. Subsequently, these two scores are combined using predetermined weighting factors to highlight materials that lie at the intersection of novelty and performance boundaries. We rely only on the density score to propose the 10% top candidates of B to be merged into the training set of A. To avoid merging a novel data block with all points similar to each other, we iteratively alternate the merging of a small number of candidates and an update of the novelty scores, until 10% of B is integrated into A.

3.1 Discussion for A-B data aggregation

Table 2 shows the average testing errors on the original test of A obtained by training different ML estimators after different AB aggregation strategies. A color scheme is used to guide the interpretation. Our experiments reveal that classical ML approaches fail to leverage the advantages offered by any of the considered aggregation strategies. Among the strategies, the plain *Concatenation* performs the worst, followed by *DiSCoVeR*, and finally *ElemConc*. This observation suggests that the contamination in the original dataset increases as a function of the number of added points, irrespective of the aggregation strategy. Contrary to classical ML approaches, DL models exhibit much greater stability. A possible explanation for this phenomenon can be attributed to the choice of loss function employed at the training stage. Notably, both *Roost* and *CrabNet* utilize a customized variant of the L1 loss referred to as 'robust' [Goodall and Lee, 2020, Wang et al., 2021]. The rationale behind employing this modified loss function is the ability to capture and incorporate the inherent noise associated with individual data points. Therefore, this approach may facilitate a more robust and stable data aggregation process. Despite that, except for sporadic cases, improvements or degradations in accuracy are not significant. Interestingly, the majority of the best overall results are found in the correspondence of the *CrabNet* model after performing a full concatenation with dataset B. Further investigation into the reasons behind this could provide valuable insights for future research. By comparing the results obtained for the calculated datasets, we also observe that maintaining consistency between the nature of datasets A and B led to slightly better performance. In conclusion, different ML algorithms do not consistently benefit from any aggregation strategy. Most interestingly, adding points targeting empty regions of the chemical space does not show a clear advantage. These findings shed light on the strengths and limitations of different approaches in the context of dataset aggregation and provide valuable insights for future studies in this domain.

While overall performance appears degraded, we have inspected the element-wise MAE for the elements that, before the aggregation, had the lowest and highest occurrences in the training set of A. We indeed observed an increase in training instances and sporadic improvements in MAE for the less populated element classes. However, this is often found in correspondence with a noticeable degradation of performance for highly populated elements. This partially explains the previously observed results, as these points weigh much more in the overall MAE. Moreover, as we enrich certain chemical regions, we eventually saturate other chemical elements which eventually coexist within the same chemical formula.

4 A-A Data aggregation

In this section, we conduct a further experiment with the intent of decoupling our results from the use of the archives' data (MPDS and MP) as our resource for gathering additional data. In fact, the use of an external database does not guarantee that the experimental conditions in A are met in B, which can lead to heavy distribution shifts [Wiles et al., 2021]. Instead, here, we simulate a progressive data acquisition of one single dataset. This is done by initially constricting dataset A to a random subset comprising only 5% of the original size. Subsequently, DiSCoVeR is used to integrate new candidates from the remaining 95% of itself. Similarly to the previous experiment, the novelty scores are updated and the new data points are iteratively added until the whole dataset is exhausted. The aforementioned strategy is compared with a random acquisition which iteratively adds random data points ignoring any novelty constraint. We assess the outcomes of the self-acquisition process on top of a test set created by holding out an amount corresponding to 20% of the original dataset: in one case such test set is created randomly; in the other case the DiSCoVeR algorithm is utilized to construct a biased test set with proportionate representatives of ordinary and extraordinary materials, with proportions $2/3$ and $1/3$. The primary objective is to evaluate whether a biased data acquisition approach facilitated by DiSCoVeR enhances the discovery of these new stoichiometries.

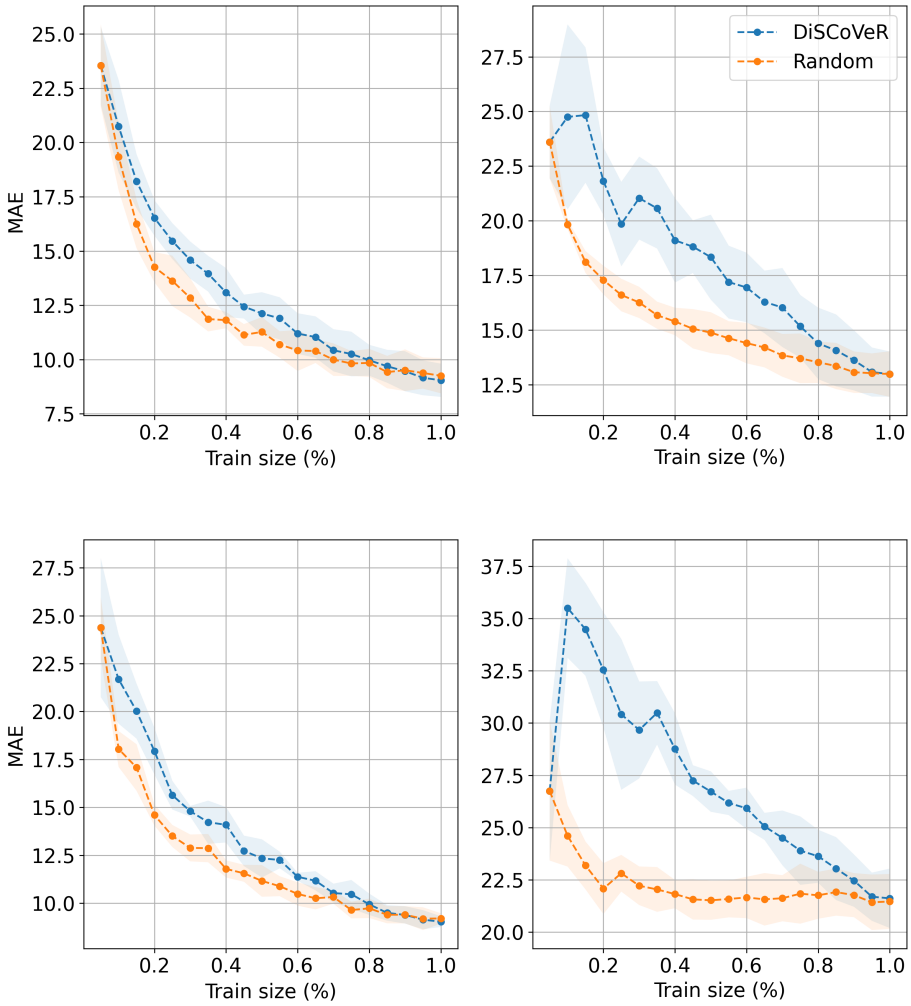

Figure 3: For the *Bulk Modulus* dataset, the plot tracks the MAE of *CrabNet* (left) and *random forest* (right) models under the A-A data integration setting. As explained in the text, the experiment is repeated for a random test set (top) and for a biased one (bottom).

4.1 Discussion for A-A data aggregation

Fig. 3 shows the outcomes of the A-A aggregation process in the case of the *bulk modulus*, which is representative of the results observed across all datasets. As for the regression model, we limit here, for brevity, our presentation to the two SOTA for classic ML and DL methods, i.e. *CrabNet* (left) and *random forest* (right). The figure encompasses the two exposed test scenarios: the randomly selected test set (top) and a biased test set created using the DisCoVer algorithm (below). Notably, our analysis uncovers a consistent pattern across both test configurations. Contrary to our initial expectations, in both cases, where the tests are either random or biased, the application of DisCoVer-guided data acquisition leads to a deceleration in the model learning process with respect to a random acquisition strategy. This observation holds true for both *CrabNet* and the *random forest* model, though with a different intensity. These findings underscore an intriguing phenomenon: the incorporation of bias, even when guided by the DisCoVer algorithm, appears to impede the learning progress of the models. Furthermore, this suggests that the balancing of a dataset in terms of chemical diversity is not to be thought of in correspondence with better ML accuracies. Consequently, a thorough examination of the intricate interplay between data acquisition strategies, model architecture, and test set composition is warranted with the intent of gaining deeper insights and devising more effective approaches for model training and evaluation in the field.

5 Conclusions

In this paper, we have investigated the aggregation of different datasets in the field of materials informatics and its impact on the performance of machine learning (ML) models for property predictions. In our evaluation, we showed that classical ML models experienced a reduction in performance under all considered aggregation strategies, indicating that the aggregation of diverse datasets can introduce noise and hinder model learning and generalization. Deep learning models exhibited more robustness, but most changes in accuracy were not statistically significant. This suggests that while deep learning models are less affected by the aggregation of datasets, they may not necessarily benefit significantly from it. Furthermore, we simulated a data acquisition process within a single dataset and compared a random data acquisition approach with one guided by the DiSCoVeR algorithm. Surprisingly, we found that prioritizing chemical diversity through the DiSCoVeR-guided approach did not lead to a faster convergence toward better accuracy but rather degraded performance. Our findings highlight the challenges and limitations of data aggregation in materials informatics and emphasize the need for caution when merging different data sources.

Future research efforts should focus on developing more effective approaches for dataset aggregation in materials informatics. As an example, supervised learning algorithms may be used to recognize and aggregate only chemical families with a higher impact on a validation error. Furthermore, to facilitate the integration of diverse datasets and enhance the reproducibility and comparability of research outcomes, the community should consider revising data saving and storing standards, as well as creating automatic ML-driven detectors for nonsense identification. By addressing these challenges, we can enhance the quality, reliability, and efficiency of data aggregation in materials informatics, leading to improved ML models and accelerated materials discovery.

6 Acknowledgements

F.O. acknowledges Pilkington (NSG Group) and EPSRC under grant number EP/V026887 for funding this research. G.D.F. acknowledges the Beckers Group for funding this research. V.V.G. thank Leverhulme Trust for support via the Leverhulme Research Centre for Functional Materials Design and EPSRC under grant number EP/V026887. T.D.S. acknowledges support from NSF Grant Nos. 1936383 and 1651668 as well as the Royal Society Wolfson Visiting Fellowship program.

References

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin a. Persson. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013. ISSN 2166532X. doi:10.1063/1.4812323. URL `http://link.aip.org/link/AMPADS/v1/i1/p011002/s1&Agg=doi`.

Scott Kirklin, James E Saal, Bryce Meredig, Alex Thompson, Jeff W Doak, Muratahan Aykol, Stephan Rühl, and Chris Wolverton. The open quantum materials database (oqmd): assessing the accuracy of dft formation energies. *npj Computational Materials*, 1(1):15010, 2015. doi:10.1038/npjcompumats.2015.10. URL `https://doi.org/10.1038/npjcompumats.2015.10`.

Evgeny Blokhin and Pierre Villars. *The PAULING FILE Project and Materials Platform for Data Science: From Big Data Toward Materials Genome*, pages 1–26. Springer International Publishing, Cham, 2018. ISBN 978-3-319-42913-7. doi:10.1007/978-3-319-42913-7_62-1. URL `https://doi.org/10.1007/978-3-319-42913-7_62-1`.

Gabriel R Schleder, Antonio C M Padilha, Carlos Mera Acosta, Marcio Costa, and Adalberto Fazzio. From dft to machine learning: recent approaches to materials science–a review. *Journal of Physics: Materials*, 2(3):032001, may 2019. doi:10.1088/2515-7639/ab084b. URL `https://dx.doi.org/10.1088/2515-7639/ab084b`.

Aria Mansouri Tehrani, Anton O Oliynyk, Marcus Parry, Zeshan Rizvi, Samantha Couper, Feng Lin, Lowell Miyagi, Taylor D Sparks, and Jakoah Brgoch. Machine learning directed search for ultraincompressible, superhard materials. *Journal of the American Chemical Society*, 140(31):9844–9853, 2018.

Abhishek Tewari, Siddharth Dixit, Niteesh Sahni, and Stéphane PA Bordas. Machine learning approaches to identify and design low thermal conductivity oxides for thermoelectric applications. *Data-Centric Engineering*, 1:e8, 2020.

Teng Wang, Kefei Zhang, Jesse Thé, and Hesheng Yu. Accurate prediction of band gap of materials using stacking machine learning model. *Computational Materials Science*, 201:110899, 2022. ISSN 0927-0256. doi:https://doi.org/10.1016/j.commatsci.2021.110899. URL `https://www.sciencedirect.com/science/article/pii/S0927025621006078`.

Hrishabh Khakurel, M. F. N. Taufique, Ankit Roy, Ganesh Balasubramanian, Gaoyuan Ouyang, Jun Cui, Duane D. Johnson, and Ram Devanathan. Machine learning assisted prediction of the young's modulus of compositionally

complex alloys. *Scientific Reports*, 11(1):17149, 2021. doi:10.1038/s41598-021-96507-0. URL https://doi.org/10.1038/s41598-021-96507-0.

Zhuo Cao, Yabo Dan, Zheng Xiong, Chengcheng Niu, Xiang Li, Songrong Qian, and Jianjun Hu. Convolutional neural networks for crystal material property prediction using hybrid orbital-field matrix and magpie descriptors. *Crystals*, 9(4), 2019. ISSN 2073-4352. doi:10.3390/cryst9040191. URL https://www.mdpi.com/2073-4352/9/4/191.

Xiang Li, Yabo Dan, Rongzhi Dong, Zhuo Cao, Chengcheng Niu, Yuqi Song, Shaobo Li, and Jianjun Hu. Computational screening of new perovskite materials using transfer learning and deep learning. *Applied Sciences*, 9(24), 2019. ISSN 2076-3417. doi:10.3390/app9245510. URL https://www.mdpi.com/2076-3417/9/24/5510.

Michael W. Gaultois, Taylor D. Sparks, Christopher K. H. Borg, Ram Seshadri, William D. Bonificio, and David R. Clarke. Data-driven review of thermoelectric materials: Performance and resource considerations. *Chemistry of Materials*, 25(15):2911–2920, Aug 2013. ISSN 0897-4756. doi:10.1021/cm400893e. URL https://doi.org/10.1021/cm400893e.

Yukari Katsura, Masaya Kumagai, Takushi Kodani, Mitsunori Kaneshige, Yuki Ando, Sakiko Gunji, Yoji Imai, Hideyasu Ouchi, Kazuki Tobita, Kaoru Kimura, and Koji Tsuda. Data-driven analysis of electron relaxation times in pbte-type thermoelectric materials. *Science and Technology of Advanced Materials*, 20(1):511–520, 2019. doi:10.1080/14686996.2019.1603885. URL https://doi.org/10.1080/14686996.2019.1603885.

Stefano Curtarolo, Wahyu Setyawan, Gus L.W. Hart, Michal Jahnatek, Roman V. Chepulskii, Richard H. Taylor, Shidong Wang, Junkai Xue, Kesong Yang, Ohad Levy, Michael J. Mehl, Harold T. Stokes, Denis O. Demchenko, and Dane Morgan. Aflow: An automatic framework for high-throughput materials discovery. *Computational Materials Science*, 58:218–226, 2012. ISSN 0927-0256. doi:https://doi.org/10.1016/j.commatsci.2012.02.005. URL https://www.sciencedirect.com/science/article/pii/S0927025612000717.

Bryce Meredig, Erin Antono, Carena Church, Maxwell Hutchinson, Julia Ling, Sean Paradiso, Ben Blaiszik, Ian Foster, Brenna Gibbons, Jason Hattrick-Simpers, Apurva Mehta, and Logan Ward. Can machine learning identify the next high-temperature superconductor? examining extrapolation performance for materials discovery. *Mol. Syst. Des. Eng.*, 3:819–825, 2018. doi:10.1039/C8ME00012C. URL http://dx.doi.org/10.1039/C8ME00012C.

Hengrui Zhang, Wei (Wayne) Chen, James M. Rondinelli, and Wei Chen. ET-AL: Entropy-targeted active learning for bias mitigation in materials data. *Applied Physics Reviews*, 10(2):021403, jun 2023. doi:10.1063/5.0138913. URL https://doi.org/10.1063%2F5.0138913.

Astrid Klipfel, Zied Bouraoui, Yael Fregier, and Adlane Sayede. Equivariant graph neural network for crystalline materials (invited paper). In *STRL@IJCAI*, 2022.

Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2(1):16028, Aug 2016. ISSN 2057-3960. doi:10.1038/npjcompumats.2016.28. URL https://doi.org/10.1038/npjcompumats.2016.28.

Sangjoon Lee, Clio Chen, Griheydi Garcia, and Anton Oliynyk. Machine learning descriptors in materials chemistry: prediction and experimental validation synthesis of novel intermetallic ucd3. 2023.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL http://www.deeplearningbook.org. Book in preparation for MIT Press.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey, 2023.

Ryan J. Murdock, Steven K. Kauwe, Anthony Yu-Tung Wang, and Taylor D. Sparks. Is domain knowledge necessary for machine learning materials properties? *Integrating Materials and Manufacturing Innovation*, 9(3):221–227, Sep 2020. ISSN 2193-9772. doi:10.1007/s40192-020-00179-z. URL https://doi.org/10.1007/s40192-020-00179-z.

Lauri Himanen, Amber Geurts, Adam Stuart Foster, and Patrick Rinke. Data-driven materials science: status, challenges, and perspectives. *Advanced Science*, 6(21):1900808, 2019.

S. O. (Safa O.) Kdasap. *Principles of electronic materials and devices / S. O. Kasap*. McGraw-Hill, Boston, 2nd ed edition, 2002. ISBN 0071122370.

Gyoung S. Na, Seunghun Jang, and Hyunju Chang. Predicting thermoelectric properties from chemical formula with explicitly identifying dopant effects. *npj Computational Materials*, 7(1):106, 2021. doi:10.1038/s41524-021-00564-y. URL https://doi.org/10.1038/s41524-021-00564-y.

Rick Mullin. Citrine informatics. *C&EN Global Enterprise*, 95, 11 2017. doi:10.1021/cen-09544-cover4.

Ya Zhuo, Aria Mansouri Tehrani, and Jakoah Brgoch. Predicting the band gaps of inorganic solids by machine learning. *The Journal of Physical Chemistry Letters*, 9(7):1668–1673, 2018. doi:10.1021/acs.jpclett.8b00124. URL https://doi.org/10.1021/acs.jpclett.8b00124. PMID: 29532658.

P. Villars, M. Berndt, K. Brandenburg, K. Cenzual, J. Daams, F. Hulliger, H. Okamoto, K. Osaki, A. Prince, H. Putz, and Shuichi Iwata. The pauling file. In *European Powder Diffraction EPDIC 8*, volume 443 of *Materials Science Forum*, pages 357–360. Trans Tech Publications Ltd, 1 2004. doi:10.4028/www.scientific.net/MSF.443-444.357.

Rhys E. A. Goodall and Alpha A. Lee. Predicting materials properties without crystal structure: deep representation learning from stoichiometry. *Nature Communications*, 11(1):6280, Dec 2020. ISSN 2041-1723. doi:10.1038/s41467-020-19964-7. URL https://doi.org/10.1038/s41467-020-19964-7.

Anthony Yu-Tung Wang, Steven K. Kauwe, Ryan J. Murdock, and Taylor D. Sparks. Compositionally restricted attention-based network for materials property predictions. *npj Computational Materials*, 7(1):77, May 2021. ISSN 2057-3960. doi:10.1038/s41524-021-00545-1. URL https://doi.org/10.1038/s41524-021-00545-1.

Steven K. Kauwe, Jake Graser, Ryan Murdock, and Taylor D. Sparks. Can machine learning find extraordinary materials? *Computational Materials Science*, 174:109498, 2020. ISSN 0927-0256. doi:https://doi.org/10.1016/j.commatsci.2019.109498. URL https://www.sciencedirect.com/science/article/pii/S0927025619307979.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi:10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Sterling G. Baird, Tran Q. Diep, and Taylor D. Sparks. Discover: a materials discovery screening tool for high performance, unique chemical compositions. *Digital Discovery*, 1:226–240, 2022. doi:10.1039/D1DD00028D.

Colton C Seegmiller, Sterling G Baird, Hasan M Sayeed, and Taylor D Sparks. Discovering chemically novel, high-temperature superconductors. 2023.

Cameron J. Hargreaves, Matthew S. Dyer, Michael W. Gaultois, Vitaliy A. Kurlin, and Matthew J. Rosseinsky. The earth mover's distance as a metric for the space of inorganic compositions. *Chemistry of Materials*, 32(24):10610–10620, 12 2020. doi:10.1021/acs.chemmater.0c03381. URL https://doi.org/10.1021/acs.chemmater.0c03381.

Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre Alvise-Rebuffi, Ira Ktena, Krishnamurthy Dvijotham, and Taylan Cemgil. A fine-grained analysis on distribution shift. *arXiv preprint arXiv:2110.11328*, 2021.