Contents lists available at ScienceDirect

Current Opinion in Solid State & Materials Science

journal homepage: www.elsevier.com/locate/cossms

Machine learning in materials research: Developments over the last decade and challenges for the future

Anubhav Jain

Energy Technologies Area, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, United States

ARTICLE INFO

Keyword:
Machine learning in materials science

ABSTRACT

The number of studies that apply machine learning (ML) to materials science has been growing at a rate of approximately 1.67 times per year over the past decade. In this review, I examine this growth in various contexts. First, I present an analysis of the most commonly used tools (software, databases, materials science methods, and ML methods) used within papers that apply ML to materials science. The analysis demonstrates that despite the growth of deep learning techniques, the use of classical machine learning is still dominant as a whole. It also demonstrates how new research can effectively build upon past research, particular in the domain of ML models trained on density functional theory calculation data. Next, I present the progression of best scores as a function of time on the matbench materials science benchmark for formation enthalpy prediction. In particular, a dramatic improvement of 7 times reduction in error is obtained when progressing from feature-based methods that use conventional ML (random forest, support vector regression, *etc.*) to the use of graph neural network techniques. Finally, I provide views on future challenges and opportunities, focusing on data size and complexity, extrapolation, interpretation, access, and relevance.

1. Introduction

The use of machine learning techniques in materials research has grown in the last decade from a small niche topic to an entire subfield within materials science & engineering. Indeed, there have been over 2000 papers on the topic of materials machine learning in the year 2023 alone, and over the past decade there has been a 1.67 times yearly growth in the number of papers (Fig. 1). A 2020 review by Morgan and Jacobs [1] found that not only were the number of papers on the topic exponentially increasing, but that the number of *review* papers per year on the topic had already reached nearly 40 by 2019. Indeed, there already exist many excellent reviews on various aspects of materials machine learning, including its applications in simulation and modeling [2–4], synthesis and characterization [5–7], manufacturing [8,9], and literature mining [10]. Reviews also exist for specific topics such as structural materials [11] or best practices for research reporting [12].

This review both looks back and looks ahead. Looking back, it examines what has enabled the field of machine learning to advance so rapidly. Indeed, about five years ago it was unclear whether the field would enter a “trough of disillusionment” or an “AI winter” [13]. However, the development of the Transformer architecture [14] in the computer science domain and the crystal graph neural network [15,16] in the materials science domain around the same time revitalized much research and led to major advancements in performance. More recently, materials machine learning is rapidly building upon advances in natural language processing, and in particular large language models such as Generative Pre-trained Transformer (GPT) models [17] (e.g., GPT-3, GPT-4, and ChatGPT). Thus, the field has largely avoided any periods of stagnation thus far and the pace of innovation appears to only be increasing.

One factor for the rapid growth of the field is the ability for materials machine learning research to rapidly build upon past work such as databases, software, ML methods, or domain-specific techniques. New research papers can build upon data sets from prior papers – bypassing expensive data collection and focusing on method development – or transfer methods developed by the computer science community to the materials domain. Such methods, packaged in reusable software libraries, can then be applied directly to tackling specific materials problems, often with minimal additional method development, data collection, or software programming.

This review is divided into three sections. In the first section, it presents an analysis of the cross-fertilization between machine learning methods, materials science methods, data and software by analyzing common citations between papers. The second section presents a

E-mail address: ajain@lbl.gov.

https://doi.org/10.1016/j.cossms.2024.101189
Received 27 February 2024; Received in revised form 12 August 2024; Accepted 1 September 2024
Available online 11 September 2024

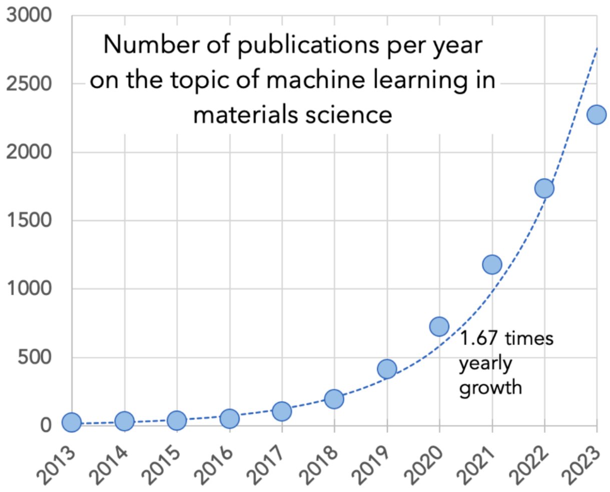

Fig. 1. The number of publications per year on the topic of machine learning in materials science. Data was retrieved on Jan 23, 2024 via http://api.elsevier.com and http://www.scopus.com using the pybliometrics Python library. See Code and Data Availability for data collection code and extracted information.

quantitative analysis of the amount of progress achieved in a particular subfield of materials machine learning – structure-based property prediction. The final section presents remaining challenges and opportunities, particularly in the areas of data size and complexity, extrapolation, interpretation, access, and relevance.

2. Rapid growth by building upon prior work

The rapid rise in publications on the topic of machine learning in materials science is fueled by advancements in software, material databases, domain-specific materials science methods, and domain-agnostic machine learning methods (Fig. 2). To examine this relationship further beyond a simple publication count increase, I compiled a data set of 6795 research papers on the topic of materials machine learning and subsequently analyzed the citations for each of these papers using the Scopus API and pybliometrics [15] Python library. I then analyzed the data set to determine all papers that were cited at least 100 times *within* materials science machine learning papers to determine commonly used tools and techniques within this subdomain. We note that this method only counts citations *within* the set of 6795 materials machine learning papers, and therefore is lower than a full citation count which may include citations from many domains or study types.

First, I focus on the results for software (Fig. 3a). The most commonly used software and the mostly highly cited work within the data set overall is the scikit-learn [18] Python package. This may stem from several reasons. First, scikit-learn implements a variety of techniques that can be useful for many different types of materials machine learning. For example, it can be applied to predict the band gaps of solids [19], to predict the strength of cement composites [20], to associate processing conditions with final properties in batteries [21], to predict the fatigue life of powder metallurgy components [22], or for many other materials tasks. Furthermore, although scikit-learn is missing the capability to implement more complex deep learning models, the small data set sizes of many materials problems often make it practical to use more conventional machine learning algorithms that have fewer parameters to train. Overall, it is interesting to note that the most commonly cited paper within the entire data set is a general-purpose machine learning tool rather than a domain-specific tool.

Following scikit-learn, the next most highly cited work is the VASP software (Fig. 3a). Indeed, the next five software libraries (VASP [23], pymatgen [24], Phonopy [25], matminer [26], and AFLOW [27]) are generally used to calculate (VASP, Phonopy) or analyze (pymatgen, matminer, AFLOW) materials properties through density functional theory (DFT). The final software, OVITO [28], is also used to visualize simulation results. The high presence of these software libraries within the group of materials machine learning papers suggests that a large fraction of materials machine learning is being performed on simulation data. Analysis of top databases cited within MSE-ML papers (Fig. 3b) supports this trend: all three databases (Materials Project [29], OQMD [30], and AFLOWLib.org [31]) focus primarily on density-functional-theory-generated data sets.

When examining the most popular domain-agnostic machine learning methods that are applied to the materials domain (Fig. 3c), traditional tree-based machine learning techniques are found to be the most dominant. The top two cited works are related to tree-based methods (random forest [32] and XGBoost [33]), and the third (Gradient Boosting [34]) is also frequently applied to tree-based methods. The use of deep learning [35] is becoming more popular; nevertheless, the larger data sets needed to train these algorithms likely inhibits more widespread usage. Finally, methods for interpreting machine learning methods, including SHAP model explanations [36] and t-SNE [37], are also popular in the materials community.

Finally, I examine commonly used methods developed for the domain of materials science (Fig. 3d). As with software and databases, simulation methods (DFT-GGA [38], Molecular Dynamics [39], DFT-Monkhorst Pack [40], and Ab Initio Molecular Dynamics–AIMD [41]) make up much of the list. The next category of methods are descriptors

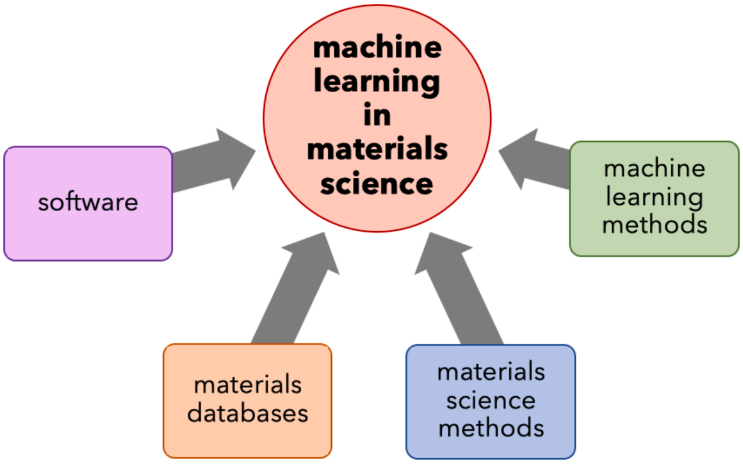
machine learning in materials science
software
machine learning methods
materials databases
materials science methods

Fig. 2. Progress in machine learning in materials science is stimulated by advances in software, materials databases, materials science methods, and machine learning methods.

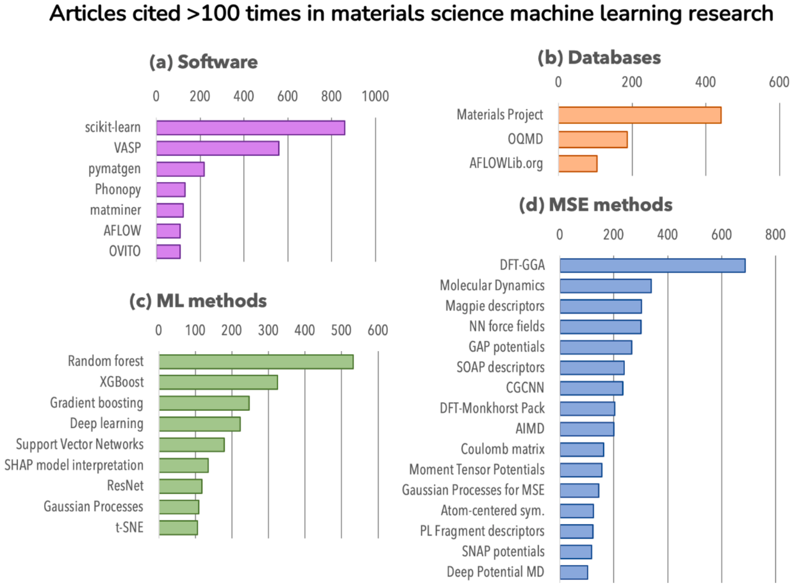

Fig. 3. Analysis of papers with >100 citations within the collection of 6795 research papers collected on the topic of materials machine learning. Data was retrieved on Jan 23, 2024 via http://api.elsevier.com and http://www.scopus.com using the pybliometrics Python library. See Code and Data Availability for data collection code and extracted information.

for materials – *i.e.*, methods that use crystal structure or composition as an input and produce a library of features that describe the input for use in machine learning techniques. Popular methods include Magpie descriptors [42], SOAP descriptors [43], Coloumb matrix [44], moment tensor potentials [45], Atom-centered symmetry [46], and PL fragment descriptors [47]. A third category is methods for ML force fields, including Behler-Parrinello potentials [48], GAP potentials [49], SNAP potentials [50], and Deep potentials [51]. Finally, two of the results are separate from the categories of simulation method, materials descriptor, or force field. The first is the crystal graph neural network [15] which popularized a neural-network-based approach to structure–property relationships by solving the problem of crystal representation as a periodic graph. The second is Gaussian Processes for iterative exploration [52], which is becoming an increasingly popular technique as machine learning is being integrated into for materials discovery campaigns.

Overall, what conclusions can we draw from such analysis? First, it is worth pointing out that despite the seeming dominance of deep learning techniques in achieving good performance in tasks like language modeling or image generation, in the materials domain such techniques are still overall less popular than conventional techniques. As found in the matbench study [53], deep learning becomes much more attractive for larger data sets (a threshold of approximately 10,000 data points was found in that study). Many materials ML problems simply do not have the data to make effective use of deep learning techniques. A second conclusion is that common databases, software libraries, and techniques are now readily available for simulation-based machine learning. However, there still remains a need for similar large and concerted efforts in other domains of materials science such as synthesis, characterization, and materials processing data. Although such databases have been developed [54–56], they do not factor heavily in our literature review of materials ML and are still in the process of realizing a large, coordinated research community. This may potentially be due to complexity and heterogeneity of data in experimental domains, making data preprocessing (e.g., cleaning and standardization) cumbersome for subsequent ML. It is possible that the push for automation in experiments may help in this endeavor, however the complexity of describing materials samples in experiments makes this far from straightforward.

3. Steady gains in accuracy

The ability of materials machine learning to draw upon prior work has led to steady gains in accuracy for a variety of tasks. In particular, the degree of improvement can be quantified for the field of materials property prediction from crystal structures. As I will later show, the state-of-the-art-models today can outperform those from 7 years ago by over a factor of 7 improvement in accuracy. With the remarkable gains in accuracy comes the potential to do more science with greater confidence using machine learning models.

To study the improvement of structure–property models in greater detail, I make use of the matbench [53] leaderboard (https://matbench.materialsproject.org). The leaderboard measures the performance of various algorithms on a series of benchmark tasks. The current top performers for each of the 13 tasks is presented in Table 1. The tasks are ordered by the number of samples (*i.e.*, data points) in the task. There is a clear delineation in the leading algorithms based on the number of samples in the task. The MODNet algorithm [57] is the clear leader on tasks with <10,000 samples, leading 5 out of 7 tasks in this category. However, these particular tasks do not have a structure provided and only chemical composition is known. Meanwhile, for tasks with >10,000 samples, the (related) coNGN [58] and coGN [58] algorithms lead 5 out of 6 tasks. This separation based on sample size (with different leading algorithms at that time) was observed in the original matbench paper [53]. However, a more careful interpretation of the data shows that today, in all tasks where a structure is provided (even the phonons task with only 1,265 samples), some form of crystal graph neural network is the dominant algorithm. Algorithms such as MODNet [57] and AMMExpress [53] come into play only in tasks in which the structure is not provided, for which the crystal graph neural networks cannot be applied at all. In such cases, using classical (*i.e.*, scientifically designed) features and classical machine learning algorithms can be a valid decision. It should also be pointed out that in many scenarios, such algorithms may give “good enough” results for the task at hand even if they are not the optimally scoring model.

I next examine progress over time for one of the matbench tasks. In Fig. 4, I plot the improvement in mean absolute error over time in predicting formation enthalpies of compounds as calculated by density functional theory simulations and tabulated in the Materials Project database. The most significant feature of Fig. 4 is the sharp drop in mean

Table 1
Current snapshot of matbench leaderboard (Jan 23, 2024).

Task name	Samples	Algorithm	Verified MAE (unit) or ROC-AUC	Structure Required
matbench_steels	312	MODNet (v0.1.12)	87.763 (MPa)	
matbench_jdft2d	636	MODNet (v0.1.12)	33.192 (meV/atom)	
matbench_phonons	1265	MegNet (kgcnn v2.1.0)	28.761 (cm^{-1})	X
matbench_expt_gap	4604	MODNet (v0.1.12)	0.333 (eV)	
matbench_dielectric	4764	MODNet (v0.1.12)	0.271 (unitless)	
matbench_expt_is_metal	4921	AMMExpress v2020	0.921	
matbench_glass	5680	MODNet (v0.1.12)	0.960	
matbench_log_gvrh	10,987	coNGN	0.067 ($\log_{10}$(GPa))	X
matbench_log_kvrh	10,987	coNGN	0.049 ($\log_{10}$(GPa))	X
matbench_perovskites	18,928	coGN	0.027 (eV/ unit cell)	X
matbench_mp_gap	106,113	coGN	0.156 (eV)	X
matbench_mp_is_metal	106,113	CGCNN v2019	0.952	X
matbench_mp_e_form	132,752	coGN	0.0170 (eV/ atom)	X

absolute error when progressing from the random forest model with Magpie descriptors [42] to the crystal graph neural network [15] (CGCNN). The switch from more conventional machine learning techniques with a series of hand-tuned features to that of a neural network architecture with features that are largely learned on-the-fly led to a tremendous, immediate jump in performance (from >100 meV/atom to <35 meV/atom). Indeed, for large data tasks, it is generally the case that hand-tuned features are not necessary for good performance [59]. Subsequently, various extensions and improvements to the CGCNN architecture have steadily reduced the error further. State-the-art-models today can reproduce formation enthalpies of compounds within the Materials Project database to <20 meV/atom, which is in most cases lower than the error of the density functional calculations as compared to experiments [60] and of various experiments amongst themselves [61]. Further examination of the chemical and structural spaces in which such models produce low versus high errors would be a fruitful area of research. Overall, however, the matbench task can likely be considered “solved” and more difficult tasks need to be designed.

4. Future challenges and opportunities

Next, I briefly review progress and outline challenges in 5 areas that are outlined in Fig. 5: data size and complexity, extrapolation, interpretation, access, and relevance.

Data size and complexity: Data is the essential raw material for machine learning. Unfortunately, materials data can be limited in quantity and high in complexity. Nevertheless, new advancements may help make progress in this area. The use of natural language processing techniques to parse the scientific literature has resulted in many new structured data sets being compiled from previous literature [62]. New advancements in large language models may allow researchers to extract data sets by simply providing a few examples of structured output from unstructured text [63]. Nevertheless, even though algorithmic improvements are rapid, gaining access to the raw literature data for parsing remains difficult. In parallel, researchers continue to compile data sets outside the materials modeling domain through user contributions [54,55,64], with tools such as Foundry-ML [65] aiming to simplify data access. From the analysis side, progress on small data remains more challenging as compared to large data sets. A previous study has generally found a power law scaling of performance with respect to data size for various graph neural network models [66]. Such scaling relies on larger data and is often ineffective for small data sets. Therefore, techniques such as multi-fidelity modeling [67,68], hybrid featurization and neural networks [57], and transfer learning [69] may be needed to leverage big data sets when analyzing small data. While we

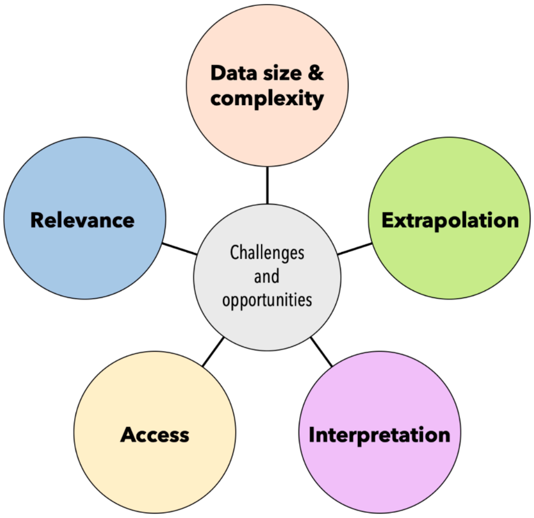

Fig. 5. Overview of major challenges and opportunities in materials ML. Current progress and future needs for each area are discussed in the main text.

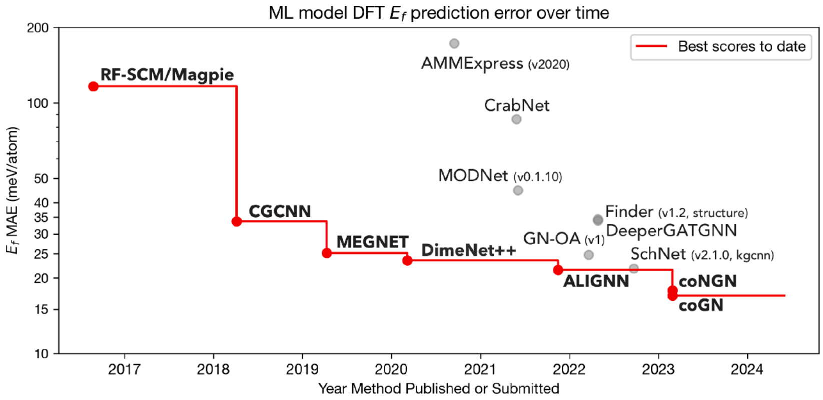

Fig. 4. Progress made on a smaller data task (elastic constants) and a larger data task (formation enthalpy) by various machine learning models.

are likely still far away from true "few shot" learning on general and extrapolative materials science tasks, progress has been reported in using large language models in this manner [70,71]. It is possible that foundational models, trained on large materials data sets and with general purpose "understanding" of materials science, will be able to serve as the basis for training more targeted machine learning algorithms using smaller data sets.

Extrapolation: No clear standard has emerged for how to quantify the extrapolative limits of machine learning models. Standard cross-validation is designed for interpolation; alternate methods such as LOCO-CV [72] are needed to ensure that the test set is significantly different than the training set for evaluating extrapolation. For example, several reports have indicated that the current set of crystal graph neural network algorithms, despite performing well on the matbench data set, may have issues generalizing out of distribution [73,74]. To this end, extensions like matbench-discovery [75] attempts to test on out of sample (*i.e.*, not in Materials Project [29]) structures to better test extrapolation. Nevertheless, the issue is far from solved. As training data sets grow in size and scope, it becomes more difficult to find and evaluate samples that are significantly outside the training domain. More fundamentally, evaluating ML models for open-ended exploration (e.g., as in iterative machine learning or generative models) remains a fruitful area of research because standard performance metrics do not translate directly to these areas [76,77] and validating each new prediction for every algorithm can be expensive. Thus, many opportunities remain to better investigate the extrapolation capability of ML models. In parallel, efforts to clarify and calibrate [78] the uncertainty of such models is needed.

Interpretation: Interpretable models are desirable because they may help uncover physical insights and relations rather than simply make predictions. [79,80] Interpretability can also clarify the domain under which the model is valid. Unfortunately, the general situation today is that the most accurate models also tend to be the most opaque. There exist many model-agnostic interpretation methods that can be applied on top of such models, such as partial dependence plots, individual conditional expectation plots, Shapley additive explanations [36], or surrogate modeling. Nevertheless, these methods do not fully capture nor do they explain the decision-making of the underlying model. As an alternative, some recent work has focused on building interpretation within the model itself [81] or in restricting models to symbolic regression [82]. Unfortunately, without a clear metric for quantifying the interpretability of a model, it remains difficult to measure progress in this area.

Access: Access is increasingly becoming an issue; the popularity of proprietary models trained on proprietary data sets such as OpenAI's GPT have made it difficult to conduct open science using such models. Worse, such models can sometimes only be accessed by an API with the underlying performance and results of the "same" model changing over time [83]. Thus, researchers publishing a certain set of results at the beginning of a project may find that the results have changed when running the same analysis on ostensibly the same model towards the end of that project. Such behavior clearly poses issues for the reproducibility requirements of published scientific research. It is at present unclear what role such models have in the future of science. It is possible that studies incorporating such models are inadmissible outright, or it may be that the models may be considered as black boxes (similar to human intuition or manual data processing) and can form part of the procedure so long as their output is independently verifiable by some other means.

The issue of access is not limited to large language models. For example, Google recently reported a breakthrough in the journal *Nature* for models to predict energies and forces from crystal structures [84]. However, neither the model nor the full data set required for training that model was made available, and attempting to reproducing the data would require exorbitant amounts of computing. This is particularly problematic because the paper also showed that many of the advancements in performance of the model stemmed from increasing the size of the data set rather than improving the architecture of the model. As the resources needed to produce a state-of-the-art result increases, the issue of who will be able to access data and models becomes of greater concern.

Relevance: Finally, as the performance of machine learning models improves, it is worth keeping in mind that these models are intended to be a means to an end rather than the end in themselves. There are many examples where ML was helpful to an outcome even if it was not perfectly optimized [85]. Conversely, there are also examples where ML models obtain good scores on one metric that does not always translate to seemingly related metrics. Examples include formation enthalpy scores not translating to phase stability [86] or MAE scores not translating to materials discovery acceleration factors [75]. Furthermore, materials or design suggestions are not valuable unless they are eventually made and tested. To that end, machine learning in the "virtual" world needs to be more closely integrated with automated laboratories [87,88] and more attention should be paid to ways in which computational predictions might be validated.

5. Conclusion

The application of machine learning to materials research has seen remarkable transformations in a relatively short amount of time. Not too long ago, machine learning in materials was considered a niche field with relatively few publications. Today, with many thousands of articles being published yearly, keeping track of the various developments has become a major challenge. The citation analysis presented in this article provides clues as to the fuel behind the rapid advancements. New publications are built upon algorithms and tools not only from within the materials science community but also from the computer science community. This has led to rapid advancements in performance on benchmark tasks such as those in the matbench protocol.

Despite the challenges, outstanding challenges and questions still remain. In an age where model size and data set size translates directly to performance, what will be the role of individual research labs? The situation today is that few academic groups have the resources to reproduce state-of-the-art results from industry in many areas of machine learning, making the issue of access and reproducibility particularly concerning at the current moment. Furthermore, challenges related to data set size and complexity, extrapolation, interpretation, and relevance still require innovative solutions. Nevertheless, researchers continue to make progress in all areas, marking the current time as a particularly exciting era for machine learning in materials research.

6. Data and software availability

Data and analysis scripts for the literature analysis can be found at Github: https://github.com/computron/pybliometrics_ml.

Data for the matbench analysis is derived from the archived benchmark data found in: https://github.com/materialsproject/matbench.

The raw data table used to derive the literature analysis plot is provided in the supplementary material.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the data via the Attach File, and I have shared code via Github linked in the article

Acknowledgements

This work was primarily funded and intellectually led by the Materials Project, which is funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division, under Contract no. DE-AC02-05-CH11231: Materials Project program KC23MP. The author would like to thank Dr. A. Dunn for assistance in the matbench analysis.

During the preparation of this work the author used ChatGPT 4.0 in order to edit the text. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Appendix A. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.cossms.2024.101189.

References

[1] D. Morgan, R. Jacobs, Opportunities and challenges for machine learning in materials science, Annu. Rev. Mater. Res. 50 (1) (2020) 71–103, https://doi.org/10.1146/annurev-matsci-070218-010015.

[2] O.A. Von Lilienfeld, K. Burke, Retrospective on a decade of machine learning for chemical discovery, Nat. Commun. 11 (1) (2020) 4895, https://doi.org/10.1038/s41467-020-18556-9.

[3] K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science, Nature 559 (7715) (2018) 547, https://doi.org/10.1038/s41586-018-0337-2.

[4] C. Chen, Y. Zuo, W. Ye, X. Li, Z. Deng, S.P. Ong, A critical review of machine learning of energy materials, Adv. Energy Mater. 10 (8) (2020) 1903242, https://doi.org/10.1002/aenm.201903242.

[5] J.M. Rickman, T. Lookman, S.V. Kalinin, Materials informatics: from the atomic-level to the continuum, Acta Mater. 168 (2019) 473–510, https://doi.org/10.1016/j.actamat.2019.01.051.

[6] S.V. Kalinin, C. Ophus, P.M. Voyles, R. Erni, D. Kepaptsoglou, V. Grillo, A. R. Lupini, M.P. Oxley, E. Schwenker, M.K.Y. Chan, J. Etheridge, X. Li, G.G.D. Han, M. Ziatdinov, N. Shibata, S.J. Pennycook, Machine learning in scanning transmission electron microscopy, Nat. Rev. Methods Primers 2 (1) (2022) 1–28, https://doi.org/10.1038/s43586-022-00095-w.

[7] A. Baskaran, E.J. Kautz, A. Chowdhary, W. Ma, B. Yener, D.J. Lewis, Adoption of image-driven machine learning for microstructure characterization and materials design: a perspective, JOM 73 (11) (2021) 3639–3657, https://doi.org/10.1007/s11837-021-04805-9.

[8] K.S. Aggour, V.K. Gupta, D. Ruscitto, L. Ajdelsztajn, X. Bian, K.H. Brosnan, N. Chennimalai Kumar, V. Dheeradhada, T. Hanlon, N. Iyer, J. Karandikar, P. Li, A. Moitra, J. Reimann, D.M. Robinson, A. Santamaria-Pang, C. Shen, M.A. Soare, C. Sun, A. Suzuki, R. Venkataramana, J. Vinciquerra, Artificial intelligence/machine learning in manufacturing and inspection: a GE perspective, MRS Bull. 44 (7) (2019) 545–558, https://doi.org/10.1557/mrs.2019.157.

[9] V. Bhuvaneswari, M. Priyadharshini, C. Deepa, D. Balaji, L. Rajeshkumar, M. Ramesh, Deep learning for material synthesis and manufacturing systems: a review, Mater. Today: Proc. 46 (2021) 3263–3269, https://doi.org/10.1016/j.matpr.2020.11.351.

[10] E.A. Olivetti, J.M. Cole, E. Kim, O. Kononova, G. Ceder, T.-Y.-J. Han, A. M. Hiszpanski, Data-driven materials research enabled by natural language processing and information extraction, Appl. Phys. Rev. 7 (4) (2020) 041317, https://doi.org/10.1063/5.0021106.

[11] T.D. Sparks, S.K. Kauwe, M.E. Parry, A.M. Tehrani, J. Brgoch, Machine learning for structural materials, Annu. Rev. Mater. Res. 50 (1) (2020) 27–48, https://doi.org/10.1146/annurev-matsci-110519-094700.

[12] N. Artrith, K.T. Butler, F.-X. Coudert, S. Han, O. Isayev, A. Jain, A. Walsh, Best practices in machine learning for chemistry, Nat. Chem. 13 (6) (2021) 505–508, https://doi.org/10.1038/s41557-021-00716-z.

[13] K.G. Reyes, B. Maruyama, The machine learning revolution in materials? MRS Bull. 44 (7) (2019) 530–537, https://doi.org/10.1557/mrs.2019.153.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems 30 (2017).

[15] T. Xie, J.C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, Phys. Rev. Lett. 120 (14) (2018) 145301, https://doi.org/10.1103/PhysRevLett.120.145301.

[16] P. Reiser, M. Neubert, A. Eberhard, L. Torresi, C. Zhou, C. Shao, H. Metni, C. Van Hoesel, H. Schopmans, T. Sommer, P. Friederich, Graph neural networks for materials science and chemistry, Commun. Mater. 3 (1) (2022) 93, https://doi.org/10.1038/s43246-022-00315-6.

[17] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, arXiv (2020), https://doi.org/10.48550/arXiv.2005.14165.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, Scikit-learn: machine learning in Python, Mach. Learn. IN Python (2011).

[19] U. Kumar, K.A. Mishra, A.K. Kushwaha, S.B. Cho, Bandgap analysis of transition-metal dichalcogenide and oxide via machine learning approach, J. Phys. Chem. Solid 171 (2022) 110973, https://doi.org/10.1016/j.jpcs.2022.110973.

[20] J. Zhang, W. Niu, Y. Yang, D. Hou, B. Dong, Machine learning prediction models for compressive strength of calcined sludge-cement composites, Constr. Build. Mater. 346 (2022) 128442, https://doi.org/10.1016/j.conbuildmat.2022.128442.

[21] A. Turetskyy, J. Wessel, C. Herrmann, S. Thiede, Battery production design using multi-output machine learning models, Energy Storage Mater. 38 (2021) 93–112, https://doi.org/10.1016/j.ensm.2021.03.002.

[22] D.S. Leininger, F.-C. Reissner, J. Baumgartner, New approaches for a reliable fatigue life prediction of powder metallurgy components using machine learning, Fatigue Fract. Eng. Mater. Struct. 46 (3) (2023) 1190–1210, https://doi.org/10.1111/ffe.13921.

[23] G. Kresse, J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B 54 (16) (1996) 11169–11186, https://doi.org/10.1103/PhysRevB.54.11169.

[24] S.P. Ong, W.D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K.A. Persson, G. Ceder, Python materials genomics (pymatgen): a robust, open-source python library for materials analysis, Comput. Mater. Sci. 68 (2013) 314–319, https://doi.org/10.1016/j.commatsci.2012.10.028.

[25] A. Togo, I. Tanaka, First principles phonon calculations in materials science, Scr. Mater. 108 (2015) 1–5, https://doi.org/10.1016/j.scriptamat.2015.07.021.

[26] L. Ward, A. Dunn, A. Faghaninia, N.E.R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K.A. Persson, G. J. Snyder, I. Foster, A. Jain, Matminer: an open source toolkit for materials data mining, Comput. Mater. Sci. 152 (2018) 60–69, https://doi.org/10.1016/j.commatsci.2018.05.018.

[27] S. Curtarolo, W. Setyawan, G.L.W. Hart, M. Jahnatek, R.V. Chepulskii, R.H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M.J. Mehl, H.T. Stokes, D.O. Demchenko, D. Morgan, AFLOW: an automatic framework for high-throughput materials discovery, Comput. Mater. Sci. 58 (2012) 218–226, https://doi.org/10.1016/j.commatsci.2012.02.005.

[28] A. Stukowski, Visualization and analysis of atomistic simulation data with OVITO–the open visualization tool, Modell. Simul. Mater. Sci. Eng. 18 (1) (2009) 015012, https://doi.org/10.1088/0965-0393/18/1/015012.

[29] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, Commentary: the materials project: a materials genome approach to accelerating materials innovation, APL Mater. 1 (1) (2013) 011002, https://doi.org/10.1063/1.4812323.

[30] J.E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD), JOM 65 (11) (2013) 1501–1509, https://doi.org/10.1007/s11837-013-0755-4.

[31] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R.H. Taylor, L.J. Nelson, G.L. W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, O. Levy, AFLOWLIB.ORG: a distributed materials properties repository from high-throughput Ab initio calculations, Comput. Mater. Sci. 58 (2012) 227–235, https://doi.org/10.1016/j.commatsci.2012.02.002.

[32] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, https://doi.org/10.1023/A:1010933404324.

[33] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.

[34] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. 29 (5) (2001) 1189–1232.

[35] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444, https://doi.org/10.1038/nature14539.

[36] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems; NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 4768–4777.

[37] L. van der Maaten, G. Hinton, Visualizing data using T-SNE, J. Mach. Learn. Res. 9 (86) (2008) 2579–2605.

[38] J.P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. 77 (18) (1996) 3865–3868, https://doi.org/10.1103/PhysRevLett.77.3865.

[39] S. Plimpton, Fast parallel algorithms for short-range molecular dynamics, J. Comput. Phys. 117 (1) (1995) 1–19, https://doi.org/10.1006/jcph.1995.1039.

[40] H.J. Monkhorst, J.D. Pack, Special points for brillouin-zone integrations, Phys. Rev. B 13 (12) (1976) 5188–5192, https://doi.org/10.1103/PhysRevB.13.5188.

[41] G. Kresse, J. Hafner, Ab initio molecular dynamics for liquid metals, Phys. Rev. B 47 (1) (1993) 558–561, https://doi.org/10.1103/PhysRevB.47.558.

[42] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, npj Comput. Mater. 2 (1) (2016) 16028, https://doi.org/10.1038/npjcompumats.2016.28.

[43] A.P. Bartók, R. Kondor, G. Csányi, On representing chemical environments, Phys. Rev. B 87 (18) (2013) 184115, https://doi.org/10.1103/PhysRevB.87.184115.

[44] M. Rupp, A. Tkatchenko, K.-R. Müller, O.A. von Lilienfeld, Fast and accurate modeling of molecular atomization energies with machine learning, Phys. Rev. Lett. 108 (5) (2012) 058301, https://doi.org/10.1103/PhysRevLett.108.058301.

[45] A.V. Shapeev, Moment tensor potentials: a class of systematically improvable interatomic potentials, Multiscale Model. Simul. 14 (3) (2016) 1153–1173, https://doi.org/10.1137/15M1054183.

[46] J. Behler, Atom-centered symmetry functions for constructing high-dimensional neural network potentials, J. Chem. Phys. 134 (7) (2011) 074106, https://doi.org/10.1063/1.3553717.

[47] O. Isayev, C. Oses, S. Curtarolo, A. Tropsha, Universal fragment descriptors for predicting electronic properties of inorganic crystals, Nat. Commun. (2016) 1–12.

[48] J. Behler, M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces, Phys. Rev. Lett. 98 (14) (2007) 146401, https://doi.org/10.1103/PhysRevLett.98.146401.

[49] A.P. Bartók, M.C. Payne, R. Kondor, G. Csányi, Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons, Phys. Rev. Lett. 104 (April) (2010) 1–4, https://doi.org/10.1103/PhysRevLett.104.136403.

[50] A.P. Thompson, L.P. Swiler, C.R. Trott, S.M. Foiles, G.J. Tucker, Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials, J. Comput. Phys. 285 (2015) 316–330, https://doi.org/10.1016/j.jcp.2014.12.018.

[51] L. Zhang, J. Han, H. Wang, R. Car, W. E, Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics, Phys. Rev. Lett. 120 (14) (2018) 143001, https://doi.org/10.1103/PhysRevLett.120.143001.

[52] D. Xue, P.V. Balachandran, J. Hogden, J. Theiler, D. Xue, T. Lookman, Accelerated search for materials with targeted properties by adaptive design, Nat. Commun. 7 (2016) 11241, https://doi.org/10.1038/ncomms11241.

[53] A. Dunn, Q. Wang, A. Ganose, D. Dopp, A. Jain, Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm, Npj Comput. Mater. 6 (1) (2020) 138, https://doi.org/10.1038/s41524-020-00406-3.

[54] B. Puchala, G. Tarcea, E.A. Marquis, M. Hedstrom, H.V. Jagadish, J.E. Allison, The materials commons: a collaboration platform and information repository for the global materials community, JOM 68 (8) (2016) 2035–2044, https://doi.org/10.1007/s11837-016-1998-7.

[55] B. Blaiszik, K. Chard, J. Pruyne, R. Ananthakrishnan, S. Tuecke, I. Foster, The materials data facility: data services to advance materials science research, JOM 68 (8) (2016) 2045–2052, https://doi.org/10.1007/s11837-016-2001-3.

[56] E. Blokhin, P. Villars, The PAULING FILE project and materials platform for data science: from big data toward materials genome, in: W. Andreoni, S. Yip (Eds.), Handbook of Materials Modeling: Methods: Theory and Modeling, Springer International Publishing, Cham, 2018, pp. 1–26, https://doi.org/10.1007/978-3-319-42913-7_62-1.

[57] P.-P. De Breuck, G. Hautier, G.-M. Rignanese, Materials property prediction for limited datasets enabled by feature selection and joint learning with MODNet, Npj Comput. Mater. 7 (1) (2021) 1–8, https://doi.org/10.1038/s41524-021-00552-2.

[58] R. Ruff, P. Reiser, J. Stühmer, P. Friederich, Connectivity optimized nested graph networks for crystal structures, arXiv August 9, 2023. https://doi.org/10.48550/arXiv.2302.14102.

[59] R.J. Murdock, S.K. Kauwe, A.-Y.-T. Wang, T.D. Sparks, Is domain knowledge necessary for machine learning materials properties? Integr. Mater. Manuf. Innov. 9 (3) (2020) 221–227, https://doi.org/10.1007/s40192-020-00179-z.

[60] G. Hautier, S.P. Ong, A. Jain, C.J. Moore, G. Ceder, Accuracy of density functional theory in predicting formation energies of ternary oxides from binary oxides and its implication on phase stability, Phys. Rev. B 85 (15) (2012) 155208, https://doi.org/10.1103/PhysRevB.85.155208.

[61] S. Kirklin, J.E. Saal, B. Meredig, A. Thompson, J.W. Doak, M. Aykol, S. Rühl, C. Wolverton, The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies, Npj Comput. Mater. 1 (1) (2015) 1–15, https://doi.org/10.1038/npjcompumats.2015.10.

[62] J.H. Lee, M. Lee, K. Min, Natural Language processing techniques for advancing materials discovery: a short review, Int. J. Precis. Eng. Manuf.-Green Tech. 10 (5) (2023) 1337–1349, https://doi.org/10.1007/s40684-023-00523-6.

[63] J. Dagdelen, A. Dunn, S. Lee, N. Walker, A.S. Rosen, G. Ceder, K.A. Persson, A. Jain, Structured information extraction from scientific text with large language models, Nat. Commun. 15 (1) (2024) 1418, https://doi.org/10.1038/s41467-024-45563-x.

[64] P. Huck, D. Gunter, S. Cholia, D. Winston, A.T. N'Diaye, K. Persson, User Applications driven by the community contribution framework MPContribs in the materials project, Concurr. Comput.: Pract. Experience 28 (7) (2016) 1982–1993, https://doi.org/10.1002/cpe.3698.

[65] K. Schmidt, A. Scourtas, L. Ward, S. Wangen, M. Schwarting, I. Darling, E. Truelove, A. Ambadkar, R. Bose, Z. Katok, J. Wei, X. Li, R. Jacobs, L. Schultz, D. Kim, M. Ferris, P.M. Voyles, D. Morgan, I. Foster, B. Blaiszik, Foundry-ML - software and services to simplify access to machine learning datasets in materials science, J. Open Source Softw. 9 (93) (2024) 5467, https://doi.org/10.21105/joss.05467.

[66] V. Fung, J. Zhang, E. Juarez, B.G. Sumpter, Benchmarking graph neural networks for materials chemistry, Npj Comput. Mater. 7 (1) (2021) 84, https://doi.org/10.1038/s41524-021-00554-0.

[67] C. Chen, Y. Zuo, W. Ye, X. Li, S.P. Ong, Learning properties of ordered and disordered materials from multi-fidelity data, Nat. Comput. Sci. 1 (1) (2021) 46–53, https://doi.org/10.1038/s43588-020-00002-x.

[68] E. Annevelink, R. Kurchin, E. Muckley, L. Kavalsky, V.I. Hegde, V. Sulzer, S. Zhu, J. Pu, D. Farina, M. Johnson, D. Gandhi, A. Dave, H. Lin, A. Edelman, B. Ramsundar, J. Saal, C. Rackauckas, V. Shah, B. Meredig, V. Viswanathan, AutoMat: automated materials discovery for electrochemical systems, MRS Bull. 47 (10) (2022) 1036–1044, https://doi.org/10.1557/s43577-022-00424-0.

[69] C. Chen, S.P. Ong, AtomSets as a hierarchical transfer learning framework for small and large materials datasets, Npj Comput. Mater. 7 (1) (2021) 173, https://doi.org/10.1038/s41524-021-00639-w.

[70] K.M. Jablonka, P. Schwaller, A. Ortega-Guerrero, B. Smit, Leveraging large language models for predictive chemistry, ChemRxiv October 17, 2023. https://doi.org/10.26434/chemrxiv-2023-fw8n4-v3.

[71] K. MaikJablonka, Q. Ai, A. Al-Feghali, S. Badhwar, J.D. Bocarsly, A.M. Bran, S. Bringuier, W.A. de Jong, M.L. Evans, N. Gastellu, J. Genzling, M. Victoria Gil, A. K. Gupta, Z. Hong, A. Imran, S. Kruschwitz, A. Labarre, J. Lála, T. Liu, S. Ma, S. Majumdar, G.W. Merz, N. Moitessier, E. Moubarak, B. Mouriño, B. Pelkie, M. Pieler, M. Caldas Ramos, B. Ranković, S.G. Rodriques, J.N. Sanders, P. Schwaller, M. Schwarting, J. Shi, B. Smit, B.E. Smith, J.V. Herck, C. Völker, L. Ward, S. Warren, B. Weiser, S. Zhang, X. Zhang, G. Ahmad Zia, A. Scourtas, K. J. Schmidt, I. Foster, A.D. White, B. Blaiszik, 14 Examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon, Digital Discovery 2 (5) (2023) 1233–1250, https://doi.org/10.1039/D3DD00113J.

[72] B. Meredig, E. Antono, C. Church, M. Hutchinson, J. Ling, S. Paradiso, B. Blaiszik, I. Foster, B. Gibbons, J. Hattrick-simpers, Can machine learning identify the next YBCO ? Examining extrapolation performance for materials discovery, Mol. Syst. Des. Eng. (2013) 1–3, https://doi.org/10.1039/x0xx00000x.

[73] K. Li, B. DeCost, K. Choudhary, M. Greenwood, J. Hattrick-Simpers, A Critical examination of robustness and generalizability of machine learning prediction of materials properties, Npj Comput. Mater. 9 (1) (2023) 1–9, https://doi.org/10.1038/s41524-023-01012-9.

[74] S.S. Omee, N. Fu, R. Dong, M. Hu, J. Hu, Structure-Based out-of-Distribution (OOD) Materials property prediction: a benchmark study, arXiv January 15, 2024. http://arxiv.org/abs/2401.08032 (accessed 2024-02-20).

[75] J. Riebesell, R.E.A. Goodall, P. Benner, Y. Chiang, B. Deng, A.A. Lee, A. Jain, K.A. Persson, Matbench discovery – a framework to evaluate machine learning crystal stability predictions, arXiv February 4, 2024. https://doi.org/10.48550/arXiv.2308.14920.

[76] C.K.H. Borg, E.S. Muckley, C. Nyby, J.E. Saal, L. Ward, A. Mehta, B. Meredig, Quantifying the performance of machine learning models in materials discovery, Digital Discovery 2 (2) (2023) 327–338, https://doi.org/10.1039/D2DD00113F.

[77] Y. Zhao, E.M.D. Siriwardane, Z. Wu, N. Fu, M. Al-Fahdi, M. Hu, J. Hu, Physics guided deep learning for generative design of crystal materials with symmetry constraints, Npj Comput. Mater. 9 (1) (2023) 1–12, https://doi.org/10.1038/s41524-023-00987-9.

[78] G. Palmer, S. Du, A. Politowicz, J.P. Emory, X. Yang, A. Gautam, G. Gupta, Z. Li, R. Jacobs, D. Morgan, Calibration after bootstrap for accurate uncertainty quantification in regression models, Npj Comput. Mater. 8 (1) (2022) 1–9, https://doi.org/10.1038/s41524-022-00794-8.

[79] H. Choubisa, P. Todorović, J.M. Pina, D.H. Parmar, Z. Li, O. Voznyy, I. Tamblyn, E. H. Sargent, Interpretable discovery of semiconductors with machine learning, Npj Comput. Mater. 9 (1) (2023) 117, https://doi.org/10.1038/s41524-023-01066-9.

[80] J.A. Esterhuizen, B.R. Goldsmith, S. Linic, Interpretable machine learning for knowledge generation in heterogeneous catalysis, Nat. Catal. 5 (3) (2022) 175–184, https://doi.org/10.1038/s41929-022-00744-z.

[81] J. Teufel, L. Torresi, P. Reiser, P. Friederich, MEGAN: Multi-Explanation Graph Attention Network, Communications in Computer and Information Science, in: L. Longo (Ed.), Explainable Artificial Intelligence, Springer Nature Switzerland, Cham, 2023, pp. 338–360, https://doi.org/10.1007/978-3-031-44067-0_18.

[82] S. Sun, R. Ouyang, B. Zhang, T.-Y. Zhang, Data-driven discovery of formulas by symbolic regression, MRS Bull. 44 (7) (2019) 559–564, https://doi.org/10.1557/mrs.2019.156.

[83] L. Chen, M. Zaharia, J. Zou, How is ChatGPT's behavior changing over time? arXiv October 31, 2023. https://doi.org/10.48550/arXiv.2307.09009.

[84] A. Merchant, S. Batzner, S.S. Schoenholz, M. Aykol, G. Cheon, E.D. Cubuk, Scaling deep learning for materials discovery, Nature 624 (7990) (2023) 80–85, https://doi.org/10.1038/s41586-023-06735-9.

[85] J.E. Saal, A.O. Oliynyk, B. Meredig, Machine learning in materials discovery: confirmed predictions and their underlying approaches, Annu. Rev. Mater. Res. 50 (1) (2020) annurev-matsci-090319-010954, https://doi.org/10.1146/annurev-matsci-090319-010954.

[86] C.J. Bartel, A. Trewartha, Q. Wang, A. Dunn, A. Jain, G. Ceder, A critical examination of compound stability predictions from machine-learned formation energies, Npj Comput. Mater. 6 (1) (2020) 97, https://doi.org/10.1038/s41524-020-00362-y.

[87] E. Stach, B. DeCost, A.G. Kusne, J. Hattrick-Simpers, K.A. Brown, K.G. Reyes, J. Schrier, S. Billinge, T. Buonassisi, I. Foster, C.P. Gomes, J.M. Gregoire, A. Mehta, J. Montoya, E. Olivetti, C. Park, E. Rotenberg, S.K. Saikin, S. Smullin, V. Stanev, B. Maruyama, Autonomous experimentation systems for materials development: a community perspective, Matter 4 (9) (2021) 2702–2726, https://doi.org/10.1016/j.matt.2021.06.036.

[88] N.J. Szymanski, B. Rendy, Y. Fei, R.E. Kumar, T. He, D. Milsted, M.J. McDermott, M. Gallant, E.D. Cubuk, A. Merchant, H. Kim, A. Jain, C.J. Bartel, K. Persson, Y. Zeng, G. Ceder, An autonomous laboratory for the accelerated synthesis of novel materials, Nature 624 (7990) (2023) 86–91, https://doi.org/10.1038/s41586-023-06734-w.