communications materials

PERSPECTIVE

https://doi.org/10.1038/s43246-022-00283-x OPEN

Why big data and compute are not necessarily the path to big materials science

Naohiro Fujinuma [1,2], Brian DeCost [3✉], Jason Hattrick-Simpers[4] & Samuel E. Lofland [5]

Applied machine learning has rapidly spread throughout the physical sciences. In fact, machine learning-based data analysis and experimental decision-making have become commonplace. Here, we reflect on the ongoing shift in the conversation from proving that machine learning can be used, to how to effectively implement it for advancing materials science. In particular, we advocate a shift from a big data and large-scale computations mentality to a model-oriented approach that prioritizes the use of machine learning to support the ecosystem of computational models and experimental measurements. We also recommend an open conversation about dataset bias to stabilize productive research through careful model interrogation and deliberate exploitation of known biases. Further, we encourage the community to develop machine learning methods that connect experiments with theoretical models to increase scientific understanding rather than incrementally optimizing materials. Moreover, we envision a future of radical materials innovations enabled by computational creativity tools combined with online visualization and analysis tools that support active outside-the-box thinking within the scientific knowledge feedback loop.

Since Frank Rosenblatt created Perceptron to play checkers[1], machine learning (ML) applications have been used to emulate human intelligence. The field has grown immensely with the advent of ever more powerful computers with increasingly smaller size combined with the development of robust statistical analyses. These advances allowed Deep Blue to beat Grandmaster Gary Kasparov in chess and Watson to win the game show *Jeopardy!* The technology has since progressed to more practical applications such as advanced manufacturing and common tasks we now expect from our phones like image and speech recognition. The future of ML promises to obviate much of the tedium of everyday life by assuming responsibility for more and more complex processes, e.g., autonomous driving.

When it comes to scientific application, our perspective is that current ML methods are just another component of the scientific modeling toolbox, with a somewhat different profile of representational basis, parametrization, computational complexity, and data/sample efficiency. Fully embracing this view will help the materials and chemistry communities to overcome perceived limitations and at the same time evaluate and deploy these techniques with the same level of rigor and introspection as any physics-based modeling methodology. Toward this end, in this essay we identify four areas in which materials researchers can clarify our thinking to enable a vibrant and productive community of scientific ML practitioners:

1. Maintain perspective on resources required
2. Openly assess dataset bias

[1] Department of Chemical Engineering, Rowan University, 201 Mullica Hill Rd, Glassboro, NJ, USA. [2] Sekisui Chemical Co., Ltd, 2-4-4 Nishitemma, Kita-ku, Osaka 530-8565, Japan. [3] Material Measurement Laboratory, National Institute of Standards and Technology, 100 Bureau Dr, Gaithersburg, MD, USA. [4] Department of Materials Science and Engineering, University of Toronto, 27 King's College Cir, Toronto, ON, Canada. [5] Department of Physics and Astronomy, Rowan University, 201 Mullica Hill Rd, Glassboro, NJ, USA. ✉email: brian.decost@nist.gov

3. Keep sight of the goal
4. Dream big enough for radical innovation

Maintain perspective on resources required

The recent high profile successes in mainstream ML applications enabled by internet-scale data and massive computation[2,3] have spurred two lines of discussion in the materials community that are worth examining more closely. The first is an unmediated and limiting preference for large-scale data and computation, under the assumption that successful ML is unrealistic for materials scientists with datasets that are orders of magnitude smaller than those at the forefront of the publicity surrounding deep learning. The second is a tendency to dismiss brute-force ML systems as unscientific. While there is some validity to both these viewpoints, there are opportunities in materials research for productive and creative ML work with small datasets and for the "go big or go home" brute-force approach.

Molehills of data (or compute) are sometimes better than mountains. A common sentiment in the contemporary deep-learning community is that the most reliable means of improving the performance of a deep-learning system is to amass ever larger datasets and apply raw computational power. This sometimes can encourage the fallacy that large-scale data and computation are fundamental requirements for success with ML methods. This can lead to needlessly deploying massively overparameterized models when simpler ones may be more appropriate[4], and it limits the scope of applied ML research in materials by biasing the set of problems people are willing to consider addressing. There are many examples of productive, creative ML work with small datasets in materials research that counter this notion[5,6].

In the small-data regime, high-quality data with informative features often trump excessive computational power with massive data and weakly correlated features. A promising approach is to exploit the bias-variance trade-off by performing more rigorous feature selection or crafting a more physically motivated model form[7]. Alternatively, it may be wise to reduce the scope of the ML task by restricting the material design space or use ML to solve a smaller chunk of the problem at hand. ML tools for exploratory analysis with appropriate features can help us comprehend much higher dimensional spaces even at an early stage of the research, which may be helpful to have a bird's-eye view on our target. For example, cluster analysis can help researchers identify representative groups in large high-throughput datasets, making the process of formulating hypotheses more tractable.

There are also specific ML disciplines aimed at addressing the well-known issues of small datasets, dataset bias, noise, incomplete featurization, and over-generalization, and there has been some effort to develop tools to address them. Data augmentation and other regularization strategies can allow even small datasets to be treated with large deep-learning models. Another common approach is transfer learning, where a proxy model is trained on a large dataset and adapted to a related task with fewer data points[8–10]. Chen et al.[11] showed that multi-fidelity graph networks could be used in comparatively inexpensive low-fidelity calculations to bolster the accuracy of ML predictions for expensive high-fidelity calculations. Finally, active learning methods are now being explored in many areas of materials research, where surrogate models are initialized on small datasets and updated as predictions are used to guide the acquisition of new data generation, often in a manner that balances exploration with optimization[12]. Generally a solid understanding of the uncertainty in the data is critical for success with these strategies, but ML systems can lead us to some insights or perhaps serve as a guide for optimization which might otherwise be intractable.

We assert that the materials community would generally benefit from taking a more model-oriented approach to applied ML, in contrast to the popular prediction-oriented approach that many method-development papers take. With the current prediction-oriented application of ML to the physical sciences, the primary intent of the model is to obtain property predictions, often for screening or optimization workflows. We propose that the community would be better served to instead use ML as a means to generate scientific understanding, using, for instance, inference techniques to quantify physical constants from experiments. To achieve the goals of scientific discovery and knowledge generation, predictive ML must often play a supporting role within a larger ecosystem of computational models and experimental measurements. It can be productive to reassess[13] the predictive tasks we are striving to address with ML methods; more carefully thought out applications may provide more benefit than simply collecting larger datasets and training higher capacity models.

Massive computation can be useful but is not everything. On the other hand, characterizing brute computation as "unscientific" can lead to missed opportunities to meaningfully accelerate and enable new kinds or scales of scientific inquiry[14]. Even without investment in massive datasets or specialized ML models, there is evidence that simply increasing the scale of computation applied can help compensate for small datasets. For example, ref. [15] show that simply by increasing the number of training iterations, large-object detection and segmentation models trained from random initialization can match the performance of the conventional transfer learning approach. In many cases, advances enabled in this way do not directly contribute to scientific discovery or development, but they absolutely change the landscape of feasible scientific research by lowering the barrier to exploration and increasing the scale and automation of data analysis.

A perennial challenge in organic chemistry is predicting the structure of proteins, but recent advances in learned potential methods[16] have provided paradigm-shifting improvements in performance made possible by sheer computational power. In addition, massive computation can enable new scientific applications through scalable automated data analysis systems. Recent examples include phase identification in electron backscatter diffraction[17] and X-ray diffraction[18], and local structural analysis via extended x-ray absorption fine structure[19,20]. These ML systems leverage extensive precomputation through the generation of synthetic training data and training of models; this makes online data analysis possible, removing barriers to more adaptive experiments enabled by real-time decision making.

In light of the potential value of large-scale computation in advancing fundamental science, the materials field should make computational efficiency[21] an evaluation criterion alongside accuracy and reproducibility[22]. Comparison of competing methods with equal computational budgets can provide insight into which methodological innovations actually contribute to improved performance (as opposed to simply boosting model capacity) and can provide context for the feasibility of various methods to be deployed as online data analysis tools. Careful design and interpretation of benchmark tasks and performance measures are needed for the community to avoid chasing arbitrary targets that do not meaningfully facilitate scientific discovery and development of novel and functional materials.

Openly assess dataset bias

Acknowledging dataset bias. It is widely accepted that materials datasets are distinct from the datasets used to train and validate ML systems for more "mainstream" applications in a number of ways. While some of this is hyperbole, there are some genuine differences that have a large impact on the overall outlook for ML in materials research. For instance, there is a community-wide perception that all ML problems involve data on the scale of the classic image recognition and spam/ham problems. While there are over 140,000 labeled structures in the Materials Project Database[23] and the MNIST[24] dataset contains about twice that amount, other popular ML benchmark datasets are much more modest in size. For instance, the Iris Dataset contains only 50 samples each of three species of Iris and is treated as a standard dataset for evaluating a host of clustering and classification algorithms. As noted above dataset size is not necessarily the major hurdle for the materials science community in terms of developing and deploying ML systems; however, the data, input representation, and task must each be carefully considered.

Viewed as a monolithic dataset, the materials literature is an extremely heterogeneous multiview corpus with a significant fraction of missing entries. Even if this dataset were accessible in a coherent digital form, its diversity and deficiencies would pose substantial hurdles to its suitability for ML-driven science. Most research papers narrowly focus on a single or a small handful of material instances, address only a small subset of potentially relevant properties and characterization modalities, and often fail to adequately quantify measurement uncertainties. Perhaps most importantly, there is a strong systemic bias toward positive results[25]. All of these factors negatively impact the generalization potential of ML systems.

Two aspects of publication bias play a particularly large role: domain bias and selection bias (Fig. 1b) . Domain bias results when training datasets do not adequately cover the input space. For example, ref. [26] recently demonstrated that the "tried and true" method of selecting reagents following previous successes artificially constrained the range of chemical space searched, providing the ML with a distorted view of the viable parameter space. Severe domain bias can lead to overly optimistic estimates of the performance of ML systems[27,28] or in the worst case even render them unusable for real-world scientific application[29,30].

Selection bias arises when some external factor influences the likelihood of a data points inclusion in the dataset. In scientific research, a major source of such selection bias is the large number of unreported failures (Fig. 1a). For instance the Landolt-Bornstein collection of ternary amorphous alloys lists 71% of the alloys as being glass formers while the actual occurence of glass-forming compounds is estimated to be about 5%[31]. This further complicates the already challenging task of learning from imbalanced datasets by skewing the prior probability of glass formation through dataset imbalance. Schrier et al.[32] reported on how incorporating failed experiments into ML models can actually improve upon the overall predictive power of a model.

Furthermore, the annotations or targets used to train ML systems do not necessarily represent true physical ground truth. As an example, in the field of metallic glasses the full width half-maximum (FWHM) of the strongest diffraction peak at low wavevector is often used to categorize thin-film material as being metallic glass, nanocrystalline, or crystalline. Across the literature the FWHM value used as the threshold to distinguish between the first two classes varies from 0.4 to 0.7 $Å^{-1}$ (with associated uncertainties) depending upon the research group. Although compendiums invariably capture the label ascribed to the samples, they almost ubiquitously omit the threshold used for the classification, the uncertainty in the measurement of the FWHM, and the associated synthesis and characterization metadata. Comprehensive studies often report only reduced summaries for the datasets presented and include full details only for a subset of "representative data". These shortcomings are common across the primary materials science literature. Given that even experts can reasonably disagree on the interpretation of experimental results, the lack of access to primary datasets prevents detailed model critique, posing a substantial impediment to model validation[29,33]. The push for creating F.A.I.R. (Findable, Accessible, Interoperable, and Reusable[34]) datasets with human/computer readable data structures notwithstanding, most of the data and meta-data for materials that have ever been made and studied have been lost to time.

Systematic errors in datasets are not restricted to experimental results alone. Theoretical predictions from high-throughput density functional theory (DFT) databases, for example, are a valuable resource for predicted material (meta-) stability, crystal structures, and physical properties, but DFT computations contain several underlying assumptions that are responsible for known systematic errors, e.g., calculated band gaps. DFT experts are well aware of these limitations and their implications for model building; however, scientists unfamiliar with the field may not be able to reasonably draw conclusions about the potential viability of a model's predictions given these limitations. Discrepancy between DFT and experimental data will expand

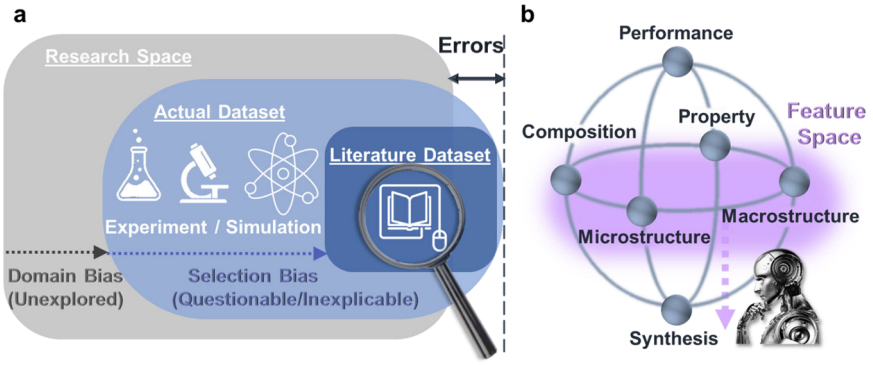

Fig. 1 Impact of datasets and feature sets in implementing ML for materials research. a Materials literature with a heterogeneous dataset due to domain bias and selection bias. Domain bias results when training datasets do not adequately cover the research space. Selection bias arises when some external factors such as questionability and inexplicability restrict the likelihood of a data inclusion in the datasets; such data can be either experimental, theoretical, or computational. **b** Holistic description of the synthesis, composition, microstructure, and macrostructure of materials, which are related to material properties and performance. Identifying a sufficient feature space with essential variables such as synthesis parameters requires careful observation and lateral thinking.

as systems get increasingly more complex, a longstanding trend in applied materials science. A heterogeneous model, in particular, may cause large uncertainty depending on the complexity of the input structure, and many times little to no information is detailed about the structure or the rationale for choosing it.

Finally, even balanced datasets with quantified uncertainties are not guaranteed to generate predictive models if the features used to describe the materials and/or how they are made are not sufficiently descriptive. Holistically describing the synthesis, composition, microstructure, macrostructure of existing materials for their property/performance (Fig. 1b) is a challenging problem and the feature set used (e.g., microstructure 2-point correlation, compositional descriptors and radial distribution functions for functional materials, and calculated physical properties) is largely community driven. This presupposes that we know and can measure the relevant features during our experiments. Often identifying the parameters that strongly influence materials synthesis and the structural aspects highly correlated to function is a matter of scientific inquiry in and of itself. For example, identifying the importance of temperature in cross-linking rubber or the effect of moisture in the reproducible growth of super-dense, vertically aligned single-walled carbon nanotubes requires careful observation and lateral thinking to connect seemingly independent or unimportant variables. If these parameters (or covariate features, e.g., chemical vapor deposition system pump curves) are not captured from the outset, then there is no hope of algorithmically discovering a causal model, and weakly predictive models are likely to be the best case output.

There is no silver bullet that will solve the issue of dataset bias, but there are several concrete steps that can be taken to begin addressing it. For instance, as a community we can commit to re-balancing the data pool against selection bias by including in our supplementary material one failed (or subpar) result for every successful result in the main text. Domain bias is best addressed by first acknowledging its existence and then encouraging researchers (possibly through funding) to spend time exploring outside of the well-known regions within their respective fields (perhaps resulting in additional data points to address selection bias). In terms of the need to capture all relevant material features, we accept that (happily) new insights will constantly crop up, and when they do, public datasets should be updated to contain the newly important features. Even if the new field is left empty for historical records, its existence will draw attention to its relevance for model builders. Finally, individuals applying ML in their research should analyze and discuss sources of bias in the data used to train and evaluate models and their potential impact on reported results.

Productivity in spite of dataset bias. Bias in historical and as-collected datasets should be acknowledged, but it does not entirely preclude their use to train an ML targeted toward scientific inquiry. Instead one can continue to gain productive insights from ML by taking the appropriate approach and thinking analytically about the results of the model.

Especially with small datasets, it is important to characterize the extent of dataset bias and perform careful model performance analysis to obtain realistic estimates of the generalization of ML models. Rauer and Bereau[28] provide compelling examples of these effects of dataset bias by comparing the empirical distribution in chemical space of three similar molecular property datasets. Dataset bias can cause common measures of a model's generalization ability to become overconfident; typically generalization ability is measured through cross-validation where a portion of the data is withheld from the training data. Recent research in the chemical and materials informatics literature has focused on developing dataset unbiasing techniques that aim to find cross-validation splits that more faithfully serve as a check against overfitting. For example, the Asymmetric Validation Embedding method[27] quantifies the bias of a dataset split by using a nearest-neighbor model to memorize the training data. If the nearest-neighbor lookup can achieve a good validation accuracy, then the training and validation sets are deemed to be too similar. Searching for cross-validation splits that minimize this bias metric can improve the robustness of the benchmark, but the Asymmetric Validation Embedding metric is specific to classification tasks. In contrast, the leave-one-cluster-out cross-validation[35] is more general, using only distances in the input space to define cross-validation groups to reduce information leakage between folds. Extending these kinds of debiasing methods to additional material classification and prediction tasks will have an outsized impact on applied artificial intelligence for practical scientific advances and discoveries because by nature these goals depend on excellent generalization and extrapolation performance.

One method for maintaining "good" features and models is to adapt an active human intervention in the ML loop. For example, we have recently demonstrated that Random Forest models that are tuned to aggressively maximize only cross-validation accuracy may produce low-quality, unreliable feature ranking explainability[36]. Carefully tracking which features (and data points) the model is most dependent on for its predictions allows a researcher to ensure that the model is capturing physically relevant trends, identify new potential insight into material behavior, and spot possible outliers. Similarly, when physics-based models are used to generate features and training data for ML models, subsequent comparison of new predictions to theory-based results offers the opportunity for improvement of both models[37]. The preceding examples are all a human-initiated post-hoc investigation of model outputs. Kusne et al.[38] recently demonstrated the inverse example where the ML model can request expert input, such as performing a measurement or calculation, that is expected to lower predictive uncertainties.

Dimensionality reduction tools and latent space models are useful to assess the general distribution of a data set. Visualizations from such models can illustrate potential bias and unequal distributions of a dataset by inspecting the internal structure/distribution and the true dimensionality. For instance, ref. [39] used principle component analysis as a method for investigating the role of dataset bias by investigating the density of data points with scores plots. Gomez-Bombarelli et al.[40] have used variational autoencoders to identify sparsely sampled regions in the parameter space by pushing them toward the outside of the latent space distribution. They demonstrated that variational autoencoders can highlight when the model is incapable of recognizing certain classes, indicating the data is outside of the distribution that the model was trained on. A holistic analysis helps gain knowledge about both the ML models and the datasets and thus may lead to more effective research steps.

A culture of careful model criticism is also important for robust applied ML research[41]. A narrow focus on benchmark tasks can lead to false incremental progress, where, over time, models begin overfitting to a particular test dataset and then lack generalizability beyond the initial dataset. Ref. [42] demonstrated that a broad range of computer vision models suffer from this effect by developing extended test sets for the CIFAR-10 and ImageNet datasets extensively used in the community for model development. This can make it difficult to reason about exactly which methodological innovations truly contribute to generalization performance. Because many aspects of ML research are empirical, carefully designed experiments are needed to separate genuine improvements from statistical effects, and care is needed to avoid post-hoc rationalization (Hypothesizing After the Results are Known (HARK)[43]).

That there is historical dataset bias is both unavoidable and unresolvable, but once identified this bias does not necessarily constrain the search for new materials in directions that directly contradict the bias[44]. For instance, ref. [26] identified anthropogenic biases in the design of amine-templated metal oxides, in that a small number of amine complexes had been used for a vast majority of the literature. Their solution was to perform 548 randomly generated experiments to demonstrate that a global maximum had not been reached but also to erode the systemic data bias their models observed. This is not to say that such an approach is a panacea for dataset or feature set bias as such experiments are still designed by scientists carrying their own biases (e.g., using only amines) and may suffer from uncaptured (but important!) features. Of course, a question remains how to best remove human bias from the experimental pipeline.

One potential path forward is deployment of automated systems that perform the ultimate selection of the experiment to be performed and manage data acquisition, functionally to attack the small dataset problem by using automation to fill in the cracks. Using these tools and adopting objective functions that permit random or maximum expected improvement exploration may help researchers avoid biasing their research toward particular solutions, allowing them to focus more on higher-level problem formulation and hypothesis specification. Currently, model prototyping often is done in notebook computing environments, which are convenient for exploring new ideas but make it easy to create unsustainable software. More accessible tools for exploring new ideas while maintaining traceability, reproducibility, flexibility, interactivity, and integration with laboratory equipment will help researchers focus on goal setting, intuition and insights for featurization, and data curation. This is analogous to ML life-cycle management[45], which is used in industrial settings to ensure traceability of predictions to specific models formulations.

Keep sight of the goal

While the implementation of ML in materials science is often focused on a push for better accuracy and faster calculations, these are not always the only objectives or even the most important ones. For the ML novice it is helpful to remember to keep the scientific aim at the forefront when selecting a model and then designing training and validation procedures. Consider the trade-off between accuracy and discovery. If one is optimizing the pseudopotentials to use for DFT[46,47], then design may be centered around accuracy of predicting material characteristics when compared to an existing benchmark set, and this may lead to better predictions for other known compounds. On the other hand, one may want to sacrifice accuracy for exploratory studies. The aforementioned high-accuracy model may fail to predict the novel combination of physical properties of an undiscovered compound. In fact, even if the phase had been recently identified and included in the training set, the model may not be trustworthy due to the inherent lack of benchmark datasets whenever new science appears.

There are clearly cases where ML is the obvious choice to accelerate research, but there can be concerns about the suitability of ML to answer the relevant question. Many applied studies focus only on physical or chemical properties of materials and often fail to include parameters relating to their fundamental utility such as reproducibility, scalability, stability, productivity, safety, or cost[48]. While humans may not be able to find correlations or patterns in high-dimensional spaces, we have rich and diverse background knowledge and heuristics; we have only just begun the difficult work of inventing ways of building this knowledge into ML systems. In addition, for domains with small datasets, limited features, and a strong need for higher-level inference rather than a surrogate model, ML should not necessarily be the default approach. A more traditional approach may be faster due to the error in the ML models associated with sample size, and heuristics can play a role even with larger datasets[49].

One alternative is to employ a hybrid method which may include a Bayesian methodology to analysis[50] or may use ML to guide the work through selective intervention[51]. ML is only a means to model data, and a good fit to the dataset is no guarantee that the model will be useful since it may have little to no relationship to actual science as it attempts to emulate apparent correlations between the features and the targets (Fig. 2). To provide some insight into this issue, Lee and Lundberg[52] developed Shapley additive explanations based on game theory to assess the impact of each feature on ML predictions.

A corollary is that any ML predictions, especially when working with small datasets, may be unphysical. Again, we stress that it doesn't imply that we should never use ML for small datasets. As demonstrated by ref. [53], non-negative matrix factorization can be constrained to provide predictions only within physical spaces. In any case, we need to employ ML tools judiciously and understand their limitations in the context of our scientific goals. For instance, while most ML models are reasonably good at interpolation[54], ML is not nearly as robust when used for extrapolation, although this can be mitigated to some extent by including rigorous statistical analyses on the predictions[55].

A discussion of errors and failure modes can help one understand the bounds of the validity of any ML analysis although it is often lacking or limited. An honest discourse includes not only principled estimates of model performance and detailed studies of predictive failure modes but also notes how reproducible the results within and across research groups. Explanation of model failure modes is required for validating the use of ML for any application.

Finally, one of the biggest potential pitfalls that can occur, even for large, well-curated datasets, is that one can lose sight of the goal by focusing on the accuracy of the model rather than using it to learn new science. There is a particular risk of the community spending disproportionate effort incrementally optimizing models to overfit against benchmark tasks[42], which may or may not even truly represent meaningful scientific endeavors in themselves. We note that in the case of the MatBench benchmark dataset and ML challenge[56], many of the top performing models are neural networks. While these models have impressive predictive capability their interpretability (and thus their ability to inform scientific progress) is limited. This is also the case for the Open Catalyst Challenge[57].

The objective should not be to identify the one algorithm that is good at everything but rather to develop a more focused effort that addresses a specific research question. For ML to reach its true potential to transform research and not just serve as a tool to expedite materials discovery and optimization, it needs to help provide a means to connect experimental and theoretical results instead of simply serving as a convenient vehicle to describe them.

Dream big enough for radical innovation

To date, ML has increased its presence in materials science for mainly three applications: (1) automating data analysis that used to be done manually; (2) serving as lead-generation in a materials-screening funnel, illustrated by the Open Quantum Materials Database and Materials Project; and (3) optimizing existing materials, processes, and devices in a broadly incremental manner. While these applications are critically important in this field, radical innovation historically has often been accomplished

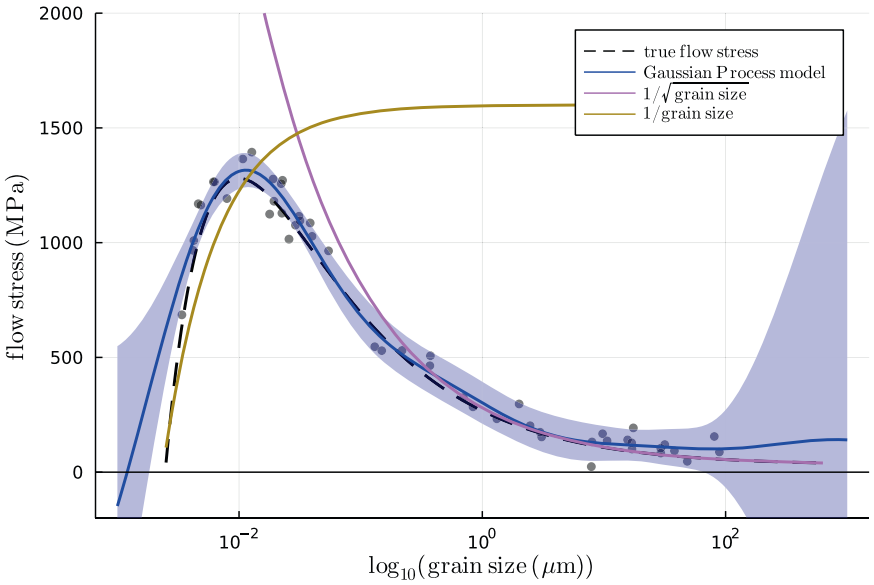

Fig. 2 Comparison of theoretical and ML Models of the Hall-Petch effect. The success of a given ML model may have little or no relationship to the actual physical processes as the model is merely interpolating between observations. For example, a Gaussian Process model can "capture" the changeover in the behavior of the flow stress in metals from being dependent on grain boundary density in large-grain metals[78] to being dominated by grain boundary sliding in nanocrystalline alloys[79] even though the model is unaware of either mechanism. However, outside the range of acquired data the lack of encoding scientific understanding results in rapidly increasing uncertainties, even in well-calibrated systems. Code for reproducing this figure is available at https://github.com/usnistgov/ml-materials-reflections[80].

outside of the context of these three general research frameworks, driven by human interests or serendipity along with stubborn trial and error. For instance, graphene was first isolated during Friday night experiments when Geim and Novoselov would try out experimental science that was not necessarily linked to their day jobs. Escobar et al.[58] discovered that peeling adhesive tape can emit enough x-rays to produce images. Shirakawa[59] discovered a conductive polyacetylene film by accidentally mixing doping materials at a concentration a thousand times too high.

Design research has argued that every radical innovation investigated was done without careful analysis of a person's or even a society's needs[60]. If this is the case, an ultimate question about ML deployment in materials science would be, can ML help humans make the startling discovery of "novel" materials and eventually new science? The new science often relies on a discrete discovery possibly outside the context of an existing theory, which is noticeably different from current ML applications which tackle problems like chess and *Jeopardy!*.

According to a proposed categorization in design research[60], one can position their research based on scientific and application familiarity (Fig. 3a). Here, incremental areas (blue region) can provide easier data acquisition and interpretation of results but may hinder new discovery. In contrast, an unexplored area may more likely provide such unexpected results but presents a huge risk of wasting research resources due to the inherent uncertainty. Self-aware resource allocation and inter-area feedback will be needed to balance novelty with the probability of successful research outcomes. Although there is currently a lack of ML methods that can directly navigate one in the radical change/radical application region to discover new science, we expect that there are methodologies that can harness ML to increase the chance of radical discovery.

Active outside-the-box exploration driven by ML-assisted knowledge acquisition. Human interests motivate outside-the-box research that may lead to a radical discovery, and these interests are fostered by theoretical or experimental knowledge acquisition. Therefore, any applied ML and automated research systems may contribute to discrete discovery by accelerating the knowledge feedback loop (Fig. 3b). Such ML-involved research loop can include a proposal of hypotheses, theoretical and experimental examination, knowledge extraction, and generalization, which may lead to an opportunity for radical thinking. Analysis and online visualization tools can help better interpret the result and mechanism of ML-involved research, which facilitates new hypotheses and generalization through knowledge extraction. Such interactive analysis/visualization can be implemented in various steps of the research loop such as feature selection, ML model investigation, and ML interpretation.

For ML to play a meaningful role in expediting this loop, one also should maintain exploratory curiosity at each step and be inspired or guided by any outputs while attentively being involved in the loop. In addition, at the very beginning of proof-of-concept research, either in a current research loop or outside-the-box search, the fear of reproducibility should not prevent the attempt at new ideas because the scientific community needs to integrate conflicting observations and ideas into a coherent theory[61].

One can harken back to Delbruck's principle of limited sloppiness[62], which reminds us that our experimental design sometimes tests unintended questions, and hidden selectivity requires attention to abnormality. In this context, ML may help us notice the anomaly or even hidden variables with a rigorous statistical procedure, leading to new pieces of knowledge and outside-the-box exploration. For instance, ref. [63] used automated experiments and statistical analysis to clarify the effect of trace water (a hidden variable) on crystal/domain growth of halide perovskite (an important property), which had often been communicated only in intra-lab conversation. Since such correlation analysis can only shed light on a domain where features are input, researchers still need comprehensive experimental records containing both data and metadata to be fed, possibly regardless of their initial interests. Also, an unbiased and flexible scientific attitude based upon observation may be crucial to reforming a question after finding the abnormality.

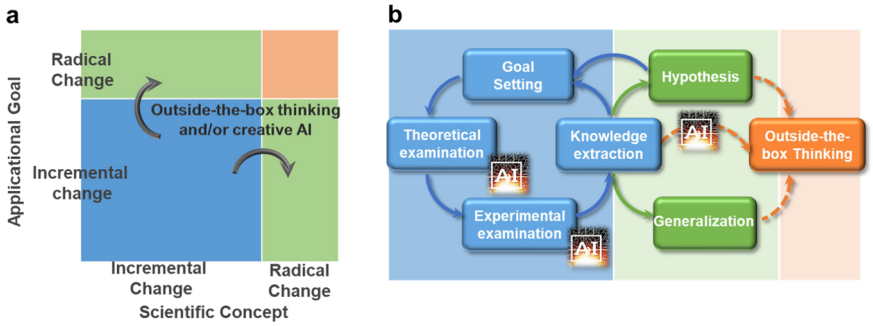

Fig. 3 Use of outside-the-box thinking in advancing scientific research with ML. a Conceptual research domain defined by a scientific concept and an applicational goal where the arrows represent a radical shift in research driven by outside-the-box thinking and/or creative artificial intelligence (AI). **b** Machine-learning-involved research loop in conjunction with possible generalization and outside-the-box thinking pathways. Blue arrows illustrate research flows in an incremental domain, green arrows show knowledge-based new research steps, and orange arrows illustrate radical shifts based on new hypotheses and generalizations in the loop.

Deep generative inverse design to assist in creating material concepts. Functionality-oriented inverse design[64] is an emerging approach for searching chemical spaces[65] for small molecules and possibly solid-state compounds[66]. Here, generative models simultaneously learn how to map existing materials to a set of few key variables and how to generate "new" materials from those key "latent" variables. One can then optimize a material by finding latent variables that should maximize the property and then generating a new material from those coordinates. Novel compounds likely to have desired properties can then be sampled from the generative model[67]. While the design spaces, such as the 166 billion molecules mapped by chemical space projects[68], are far beyond the human capability to understand them comprehensively, ML may distill patterns connecting functionalities and compound structures spanning the space. This approach can be a critical step in conceptualizing materials design based upon desired functionalities and further accelerating the ML-driven research loop. One application of such inverse design is to create a property-first optimization loop which includes defining a desired property, proposing a material and structure for that property, validating the results with (automated) experiments, and refining the model.

While these generative methods may start to approach creativity, they still explicitly aim to learn an empirical distribution based on the available data. Therefore, extrapolation outside of the current distribution of known materials is not guaranteed to be productive. For instance, these methods would probably not generate a carbon nanotube given only pre-nanotube-era structures for training or generate ordered superlattices if there is none in the training data. In addition, these huge datasets are mainly constructed based on simulation, and we need to be careful about a gap between simulated and actual experimental data as discussed previously. Still, a new concept extracted from inverse design may inspire researchers to jump into a new discrete subfield of material design by actively interpreting the abstracted property-structure relationship.

Creative artificial intelligence for materials science. The essence of scientific creativity is the production of new ideas, questions, and connections[69]. The era of artificial intelligence as an innovative investigator in this sense has yet to arrive. However, since human creativity has been captured by actively learning and connecting dots highlighted by our curiosity, it may be possible that machine "learning" can be as creative as humans in order to reach radical innovation.

While conventional supervised natural language processing[70] has required large hand-labeled datasets for training, a recent unsupervised learning study[71] indicates the possibility of extracting knowledge from literature without human intervention to identify relevant content and capturing preliminary materials science concepts such as the underlying structure of the periodic table and structure-properties relationships. This unsupervised learning was demonstrated by encoding latent literature into information-dense word embeddings, which recommended some materials for a specific application ahead of human discovery. Since the amount of currently existing literature is too massive for human cognition, such generative artificial intelligence systems may be useful to suggest a specific design or concept given appropriately defined functionalities.

Beyond latent variable optimization, one may consider computational creativity, which is used to model imagination in fields such as the arts[72], music[73], and gaming. This endeavor may start with finding a vector space to measure novelty as a distance[74]. A novelty-oriented algorithm searches the space for a set of distant new objects that is as diverse as possible as to maximize novelty instead of an objective function[75]. Since there would be some bias for measuring the distance along with exploratory space, deep learning novelty explorer (DeLeNox) was recently proposed[76] as a means to dynamically change the distance functions for improved diversity. These approaches could be applied to materials science to diversify research directions and help us pose and consider novel materials and ideas though measuring novelty may be subjective and most challenging for the community, and one always needs to be mindful of ethical and physical materials constraints.

Outlook

Machine learning has been effective at expediting a variety of tasks, and the initial stage of its implementation for materials research has already confirmed that it has great promise to accelerate science and discovery[77]. To realize that full potential, we need to tailor its usage to answer well defined questions while keeping perspective of the limits of the resources needed and the bounds of meaningful interpretation of the resulting analyses. Eventually, we may be able to develop ML algorithms that will consistently lead us to new breakthroughs. In the meantime, a complementary team of humans, ML, and robots has already begun to advance materials science.

Received: 17 December 2021; Accepted: 9 August 2022;
Published online: 30 August 2022

References

1. Rosenblatt, F. Perceptron simulation experiments. *Proc. IRE* **48**, 301–309 (1960).
2. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inform. Proc. Syst.* **33**, 1877–1901 (2020).
3. Deng, J. et al. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, 248–255 (Ieee, 2009).
4. D'Amour, A. et al. Underspecification presents challenges for credibility in modern machine learning. arXiv preprint arXiv:2011.03395 (2020).
5. Hattrick-Simpers, J. R., Choudhary, K. & Corgnale, C. A simple constrained machine learning model for predicting high-pressure-hydrogen-compressor materials. *Mol. Syst. Design Eng.* **3**, 509–517 (2018).
6. Xue, D. et al. Accelerated search for materials with targeted properties by adaptive design. *Nat. Commun.* **7**, https://doi.org/10.1038/ncomms11241 (2016).
7. Childs, C. M. & Washburn, N. R. Embedding domain knowledge for machine learning of complex material systems. *MRS Commun.* **9**, 806–820 (2019).
8. Yamada, H. et al. Predicting materials properties with little data using shotgun transfer learning. *ACS Centr. Sci.* **5**, 1717–1730 (2019).
9. Hoffmann, J. et al. Machine learning in a data-limited regime: augmenting experiments with synthetic data uncovers order in crumpled sheets. *Sci. Adv.* **5**, eaau6792 (2019).
10. Goetz, A. et al. Addressing materials' microstructure diversity using transfer learning. *npj Comput. Mater.* **8**, 1–13 (2022).
11. Chen, C., Zuo, Y., Ye, W., Li, X. & Ong, S. P. Learning properties of ordered and disordered materials from multi-fidelity data. *Nat. Comput. Sci.* **1**, 46–53 (2021).
12. Lookman, T., Balachandran, P. V., Xue, D. & Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Comput. Mater.* **5**, https://doi.org/10.1038/s41524-019-0153-8 (2019).
13. Bartel, C. J. et al. A critical examination of compound stability predictions from machine-learned formation energies. *npj Comput. Mater.* **6** (2020). https://doi.org/10.1038/s41524-020-00362-y. Bartel et al. show that compound stability prediction on the basis of regression models for formation energy cannot be taken at face value.
14. Holm, E. A. In defense of the black box. *Science* **364**, 26–27 (2019).
15. He, K., Girshick, R. & Dollár, P. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4918–4927. https://doi.org/10.1109/ICCV.2019.00502 (2019).
16. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
17. Kaufmann, K., Zhu, C., Rosengarten, A. S. & Vecchio, K. S. Deep neural network enabled space group identification in EBSD. *Microscopy Microanaly.* **26**, 447–457 (2020).
18. Maffettone, P. M. et al. Crystallography companion agent for high-throughput materials discovery. *Nat. Comput. Sci.* **1**, 290–297 (2021).
19. Timoshenko, J. et al. Linking the evolution of catalytic properties and structural changes in copper–zinc nanocatalysts using operando EXAFS and neural-networks. *Chem. Sci.* **11**, 3727–3736 (2020).
20. Schmeide, K. et al. Technetium immobilization by chukanovite and its oxidative transformation products: Neural network analysis of EXAFS spectra. *Sci. Total Environ.* **770**, 145334 (2021).
21. Schwartz, R., Dodge, J., Smith, N. A. & Etzioni, O. Green AI. *Commun. ACM* **63**, 54–63 (2020).
22. Pineau, J. et al. Improving reproducibility in machine learning research: a report from the neurips 2019 reproducibility program. *J. Mach. Learning Res.* **22** (2021). This report summarizes common sources of computational irreproducibility in machine learning research and assesses the impact of a reproducibility checklist on improving quality and transparency of research.
23. Jain, A. et al. The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
24. Grother, P. J. & Flanagan, P. A. NIST special database 19: Handprinted forms and characters database, National Institute of Standards and Technology. https://doi.org/10.18434/T4H01C (1995).
25. Dwan, K. et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS ONE* **3**, e3081 (2008).
26. Jia, X. et al. Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature* **573**, 251–255 (2019). This work illustrates how follow-on-study bias influences the exploration of subsequent chemical studies across an entire field and shows that more time spent performing "bad" experiments enriches our overall understanding of how inorganic synthesis works.
27. Wallach, I. & Heifets, A. Most ligand-based classification benchmarks reward memorization rather than generalization. *J. Chem. Inform. Modeling* **58**, 916–932 (2018).
28. Rauer, C. & Bereau, T. Hydration free energies from kernel-based machine learning: compound-database bias. *J. Chem. Phys.* **153**, 014101 (2020).
29. Griffiths, R.-R., Schwaller, P. & Lee, A.A. Dataset bias in the natural sciences: a case study in chemical reaction prediction and synthesis design (2021).
30. Cubuk, E. D., Sendek, A. D. & Reed, E. J. Screening billions of candidates for solid lithium-ion conductors: a transfer learning approach for small data. *J. Chem. Phys.* **150**, 214701 (2019).
31. Kawazoe, Y., Carow-Watamura, U. & Yu, J.-Z. (eds.) Physical Properties of Ternary Amorphous Alloys. Part 2: Systems from B-Be-Fe to Co-W-Zr (Springer Berlin Heidelberg, 2011). https://doi.org/10.1007/978-3-642-13850-8.
32. Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
33. Hattrick-Simpers, J. R. et al. An open combinatorial diffraction dataset including consensus human and machine learning labels with quantified uncertainty for training new machine learning models. *Integr. Mater. Manufact. Innovat.* **10**, 311–318 (2021).
34. Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* **3**, https://doi.org/10.1038/sdata.2016.18 (2016).
35. Meredig, B. et al. Can machine learning identify the next high-temperature superconductor? examining extrapolation performance for materials discovery. *Mol. Syst. Desig. Eng.* **3**, 819–825 (2018).
36. Lei, K., Joress, H., Persson, N., Hattrick-Simpers, J. R. & DeCost, B. Aggressively optimizing validation statistics can degrade interpretability of data-driven materials models. *J. Chem. Phys.* **155**, 054105 (2021).
37. Liu, N. et al. Interactive human–machine learning framework for modelling of ferroelectric–dielectric composites. *J. Mater. Chem. C* **8**, 10352–10361 (2020).
38. Kusne, A. G. et al. On-the-fly closed-loop materials discovery via bayesian active learning. *Nat. Commun.* **11**, https://doi.org/10.1038/s41467-020-19597-w (2020).
39. Breuck, P.-P. D., Evans, M. L. & Rignanese, G.-M. Robust model benchmarking and bias-imbalance in data-driven materials science: a case study on MODNet. *J. Phys.: Condensed Matter* **33**, 404002 (2021).
40. Gomez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Centr. Sci.* **4**, 268–276 (2018).
41. Lipton, Z. C. & Steinhardt, J. Troubling trends in machine learning scholarship: some ml papers suffer from flaws that could mislead the public and stymie future research. *Queue* **17**, 45–77 (2019).
42. Recht, B., Roelofs, R., Schmidt, L. & Shankar, V. Do imagenet classifiers generalize to imagenet? In International Conference on Machine Learning, 5389–5400 (PMLR, 2019).
43. Gencoglu, O. et al. HARK side of deep learning - from grad student descent to automated machine learning. CoRR abs/1904.07633. http://arxiv.org/abs/1904.07633 (2019).
44. Nguyen, T. N. et al. Learning catalyst design based on bias-free data set for oxidative coupling of methane. *ACS Catalys.* **11**, 1797–1809 (2021).
45. John, M. M., Olsson, H. H. & Bosch, J. Towards mlops: a framework and maturity model. 47th Euromicro Conference on Software Engineering and Advanced Applications. 1–8 (SEAA, 2021).
46. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, https://doi.org/10.1103/physrevlett.98.146401 (2007).
47. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, https://doi.org/10.1103/physrevlett.104.136403 (2010).
48. Olivetti, E. A. & Cullen, J. M. Toward a sustainable materials system. *Science* **360**, 1396–1398 (2018). Discusses materials research in a more general context than simply material properties.
49. George, J. & Hautier, G. Chemist versus machine: Traditional knowledge versus machine learning techniques. *Trends in Chemistry* **3**, 86–95 (2021). Discussion of tradeoffs of conventional research compared to AI-assisted techniques and how the two can be synergistically merged.
50. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. *Bayesian data analysis* (Chapman and Hall/CRC, 1995).
51. Hutchinson, M. L. et al. Overcoming data scarcity with transfer learning. arXiv preprint arXiv:1711.05099 (2017).
52. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inform. Proc. Syst.* **30** (2017).
53. Maffettone, P. M., Daly, A. C. & Olds, D. Constrained non-negative matrix factorization enabling real-time insights of in situ and high-throughput experiments. *Appl. Phys. Rev.* **9**, 041410 (2021).
54. Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction* (Springer open, 2017).
55. Tran, K. et al. Methods for comparing uncertainty quantifications for material property predictions. *Mach. Learning: Sci. Technol.* **1**, 025006 (2020).
56. Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Comput. Mater.* **6**, 1–10 (2020).

57. Chanussot, L. et al. Open catalyst 2020 (oc20) dataset and community challenges. *ACS Cataly.* **11**, 6059–6072 (2021).
58. Sanderson, K. Sticky tape generates x-rays. *Nature* https://doi.org/10.1038/news.2008.1185 (2008).
59. Guo, X. Conducting polymers forward. *Nat. Mater.* **19**, 921–921 (2020).
60. Norman, D. A. & Verganti, R. Incremental and radical innovation: Design research vs. technology and meaning change. *Design Issues* **30**, 78–96 (2014).
61. Redish, A. D., Kummerfeld, E., Morris, R. L. & Love, A. C. Opinion: Reproducibility failures are essential to scientific inquiry. *Proc. Natl Acad. Sci.* **115**, 5042–5046 (2018).
62. Yaqub, O. Serendipity: Towards a taxonomy and a theory. *Res. Policy* **47**, 169 (2018).
63. Nega, P. W. et al. Using automated serendipity to discover how trace water promotes and inhibits lead halide perovskite crystal formation. *Appl. Phys. Lett.* **119**, 041903 (2021).
64. Zunger, A. Inverse design in search of materials with target functionalities. *Nat. Rev. Chem.* **2**, https://doi.org/10.1038/s41570-018-0121 (2018).
65. Kirkpatrick, P. & Ellis, C. Chemical space. *Nature* **432**, 823–823 (2004).
66. Ren, Z. et al. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter* **5**, 314–335 (2022).
67. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: generative models for matter engineering. *Science* **361**, 360–365 (2018).
68. Reymond, J.-L. The chemical space project. *Acc. Chem. Res.* **48**, 722–730 (2015).
69. Lehmann, J. & Gaskins, B. Learning scientific creativity from the arts. *Palgrave Commun.* **5**, https://doi.org/10.1057/s41599-019-0308-8 (2019).
70. Krallinger, M., Rabal, O., Lourenço, A., Oyarzabal, J. & Valencia, A. Information retrieval and text mining technologies for chemistry. *Chem. Rev.* **117**, 7673–7761 (2017).
71. Tshitoyan, V. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019). Unsupervised learning was demonstrated by encoding latent literature into information-dense word embeddings, which recommended some materials for a specific application by capturing materials science concepts.
72. Ellis, K. et al. Dreamcoder: Growing generalizable, interpretable knowledge with wake-sleep bayesian program learning. CoRR abs/2006.08381 https://arxiv.org/abs/2006.08381 (2020).
73. Briot, J., Hadjeres, G. & Pachet, F. Deep learning techniques for music generation - A survey. CoRR abs/1709.01620 http://arxiv.org/abs/1709.01620 (2017).
74. Berns, S. & Colton, S. *Bridging generative deep learning and computational creativity*. In *Proc. 11th International Conference on Computational Creativity*, 406–409 (2020).
75. Lehman, J. & Stanley, K. O. Abandoning objectives: evolution through the search for novelty alone. *Evol. Comput.* **19**, 189–223 (2011). A novelty-oriented algorithm for finding an instance that differs significantly from previous ones outperformed the objective-based search in some tasks, suggesting that some problems are best solved by methods that ignore the objective.
76. Liapis, A., Martinez, H. P., Togelius, J. & Yannakakis, G. N.Transforming exploratory creativity with delenox. CoRR abs/2103.11715 https://arxiv.org/abs/2103.11715 (2021).
77. Baker, N. et al. Workshop report on basic research needs for scientific machine learning: Core technologies for artificial intelligence. Tech. Rep., USDOE Office of Science, Washington, DC (United States) https://doi.org/10.2172/1478744 (2019).
78. Cordero, Z. C., Knight, B. E. & Schuh, C. A. Six decades of the hall–petch effect – a survey of grain-size strengthening studies on pure metals. *Int. Mater. Rev.* **61**, 495–512 (2016).
79. Trelewicz, J. R. & Schuh, C. A. The hall–petch breakdown in nanocrystalline metals: a crossover to glass-like deformation. *Acta Materialia* **55**, 5948–5958 (2007).
80. Fujinuma, N., DeCost, B., Hattrick-Simpers, J. & Lofland, S. ml-materials-reflections: v0.1. https://doi.org/10.5281/zenodo.6522627 (2022).

Author contributions

N.F. Conceptualization (lead), Visualization, Writing (original draft), Writing (review & editing). B.D.C. Conceptualization, Visualization, Writing (original draft), Writing (review & editing). J.H-S. Conceptualization, Writing (original draft), Writing (review & editing). S.L. Conceptualization, Writing (original draft), Writing (review & editing).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43246-022-00283-x.

Correspondence and requests for materials should be addressed to Brian DeCost.

Peer review information *Communications Materials* thanks Lars Banko, Logan Ward and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Milica Todorović and Aldo Isidori. Peer reviewer reports are available.

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

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