Machine Learning: Science and Technology

PERSPECTIVE • OPEN ACCESS

Scientific AI in materials science: a path to a sustainable and scalable paradigm

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
19 March 2020

REVISED
18 May 2020

ACCEPTED FOR PUBLICATION
5 June 2020

PUBLISHED
14 July 2020

PERSPECTIVE

Scientific AI in materials science: a path to a sustainable and scalable paradigm

BL DeCost[1], JR Hattrick-Simpers[1], Z Trautt[1], AG Kusne[1], E Campo[2,3] and ML Green[1]

[1] National Institute of Standards and Technology, Gaithersburg, MD, United States of America
[2] National Science Foundation, Arlington, VA, United States of America
[3] Campostella Research & Consulting, LLC, Alexandria, VA, United States of America

E-mail: brian.decost@nist.gov

Keywords: materials science, artificial intelligence, reproducible science, cross-discplinary research

Abstract

Recently there has been an ever-increasing trend in the use of machine learning (ML) and artificial intelligence (AI) methods by the materials science, condensed matter physics, and chemistry communities. This perspective article identifies key scientific, technical, and social opportunities that the materials community must prioritize to consistently develop and leverage Scientific AI (SciAI) to provide a credible path towards the advancement of current materials-limited technologies. Here we highlight the intersections of these opportunities with a series of proposed paths forward. The opportunities are roughly sorted from scientific/technical (*e.g.* development of robust, physically meaningful multiscale material representations) to social (*e.g.* promoting an AI-ready workforce). The proposed paths forward range from developing new infrastructure and capabilities to deploying them in industry and academia. We provide a brief introduction to AI in materials science and engineering, followed by detailed discussions of each of the opportunities and paths forward.

1. A brief perspective on AI in materials science

Recent reports, reviews, symposia, and workshops have heralded machine learning (ML) and artificial intelligence (AI) methods as the next scientific paradigm in materials discovery and optimization [1–5]. Applications to materials science have exploded, spanning data analysis, knowledge extraction, and experiment selection [1, 6–9]. The numerous reasons for this trend are related to the omnipresence of ML systems in our everyday lives, the free availability software, and the demonstrated successes in materials discovery and on-the-fly data acquisition inspired by the Materials Genome Initiative (MGI) [1, 10–12]. However, despite their recent prominence, these techniques have been applied in a variety of materials science fields since the early 1960's [13–17].

Some recent examples of the successful implementation of ML to materials science were demonstrated by the high-throughput experimental (HTE, also known as 'combinatorial') community. Parallel material synthesis and rapid characterization introduces a critical bottleneck in the analysis of hundreds to thousands of high-quality measurements correlated in composition, processing and microstructure [18–21]. There have been several international efforts to standardize data formats and create data analysis and interpretation tools for large scale data sets [22–24]. The rise of the HTE community resulted in the creation of new and creative modes of measuring properties and visualizing and interpreting data. As algorithms for automating these tasks mature, decision-making and experimental planning are emerging as new bottlenecks in the materials research process. This means that advances in AI for materials research are as important as ever for accelerating innovation in materials, for example through emerging autonomous experimental systems [25–29].

Although the application of AI is now increasingly commonplace in the materials community, we are approaching the peak of excitement and inflated expectations. Some disillusionment is inevitable, but we believe that by pursuing the following opportunities the community will more rapidly reach a steady state of widespread productive application of AI. Figure 1 summarizes the scope of our proposed paths towards achieving this goal through 1) methodological development in scientific AI, 2) significant investment in cyber-physical infrastructure, 3) commitment to measures improving trust in AI systems, and 4) workforce development.

2. Opportunities in scientific AI

The transition from expectations to practice for AI will require development of robust Scientific AI systems that can go beyond generating leads, i.e. nudges in the right direction, to providing rich functionality that enables scientific discovery. Two opportunities to close this gap are:

(a) developing Scientific AI systems that combine ML techniques with physical mechanisms
(b) innovative applications of AI systems to directly derive scientific insight

A robust community of interdisciplinary materials science and engineering (MSE) and ML researchers is needed to enable the algorithmic development to support these two goals. Distributed automated laboratory systems will facilitate this development by equalizing access to cutting edge experimental materials science, providing a substrate for high-impact interdisciplinary collaboration. Materials have always been technology enablers, and currently there are many key technology areas that await materials discovery and processing solutions. Addressing these opportunities will drive and propel the required developments.

2.1. Incorporating physical mechanisms into ML models

The brute force strategy of collecting massive annotated datasets, such as those that enabled the current wave of advances in image recognition, natural language understanding, and neural translation, is untenable due to the relative scarcity of many types of materials data, and the high cost of obtaining materials data. Instead, the materials community needs to address underdeveloped material and processing representations to improve model quality and expand application of AI methods, leveraging the so-called bias-variance tradeoff [30]. Simple models fail to capture the complexity of hierarchical materials structures (i.e. they underfit due to high bias), while high capacity models often yield pathological or trivial results for small and medium-sized datasets (i.e. they overfit due to high variance). The challenge is to introduce the **right** kind of bias into high capacity models by designing input representations and model forms to reflect known invariances, equivariances, and symmetries in the domain [31–35]. In the context of scientific AI, this means incorporation of **mechanistic biases** to create interpretable models and learning algorithms, explicitly incorporating physical heuristics, theories, and laws into the model form. For example, the Physically Inspired Neural Network interatomic potential [36] uses a neural network to adaptively parametrize a classical interatomic potential form instead of directly modeling forces and energies. More expansively, universal differential equations [34] directly incorporate neural networks into mechanistic differential equation models.

A key opportunity is to systematically integrate the vast implicit and explicit materials knowledge in the published literature on a per-task basis through model form specification and learning algorithm design choices. This principle is applied in a limited way in the materials AI community, but much research is needed to more fully incorporate physical intuition before ML models can extrapolate to new regions of material space. Development of knowledge graphs and ontologies that capture subject matter expertise will help to provide more actionable material representations and hierarchical material models. Differentiable programming [35] (and probabilistic programming more generally) is a promising new set of tools for coordinating and unifying complementary sources of mechanistic physical information.

In addition to incorporating mechanistic biases at the level of individual modeling tasks, scientific AI systems for materials development will require the development of hybrid machine learning systems that bridge time and length scales as well as experimental and computational paradigms. Outside of the interatomic potentials community, there are few demonstrations of representing material structure representations tailored for dynamic processes. It is difficult to encode certain types of metadata (environment, processing paths, heat transfer characteristics that depend on geometry, etc) that are known to influence material properties. Vector-valued and time-varying material processing attributes (such as loading and annealing schedules) are often reduced to categorical and tabular representations. Importantly, much effort is required to address technologically important materials systems, where the complexity of material processing far exceeds that of laboratory-scale studies.

2.2. Deriving scientific insight from AI models

In many ways, current applications of AI in materials science focus more on solving engineering and design problems than on directly deriving fundamental scientific insight from data. Current materials science AI applications predominantly focus on lead generation and black-box optimization. To realize the full potential of AI to help us more efficiently and effectively practice scientific inquiry, the materials community must develop AI systems that can represent, evaluate, and perform inference about physical mechanisms underlying observational data.

In the short term, creative application of existing ML methods is enabling new avenues to accelerate scientific discovery. Active learning, for example, might be applied to identify a set of optimal experiments to disambiguate a list of potential physical theories, as is being explored in the social sciences [37]. Similarly, algorithmically driven experimentation could be used to search for counterexamples to heuristic models or physical theories, potentially providing materials scientists with valuable insights into why these heuristics and theories break down [38]. Furthermore, much of the existing materials knowledge base is in the form of implicit institutional knowledge and expert intuition. Thus, development of 'human-in-the-loop' methodologies leveraging real-time model visualization, introspection, and feedback must not be overlooked.

An important next step in scientific AI is the development of new AI methods tailored for scientific discovery. This includes methods that can infer physical relationships, mechanisms, and principles from data, potentially drawing from the fields of causal discovery [39] and probabilistic programming [40]. At the 'Strong AI' extreme of this line of inquiry, hypothetical AI systems will be expected to formulate and test scientific theories to credibly identify new scientific paradigms. Even if such systems can be constructed, they will still need to overcome the *Pauling Problem* [41], where physical bias overwhelms new evidence of worldview-breaking phenomena such as superconductivity, 2D materials, or quasicrystals.

2.3. Paths forward

Cross-disciplinary Collaboration

- Generate funding opportunities targeted towards funding cross-disciplinary research at the cutting edge of MSE, ML, AI, and Robotics to promote communication skills to identify and frame mutually interesting research.
- Collaborate to develop multiscale materials and knowledge representations and generative modeling techniques
- Create career opportunities at the research associate and technician levels in applied ML and Software Engineering.
- Explore probabilistic programming methods to meld physical and phenomenological modeling with machine learning.
- Develop objective methods for identification and evaluation of the most informative or unusual datum in any given scientific dataset.

Autonomous research platforms:

- Develop open autonomous research platforms to provide a substrate for developing and deploying materials AI methods on large-scale materials design problems.
- Provide opportunities for the broad materials and AI communities to have access to these platforms, lowering the barrier to entry to materials discovery and design.

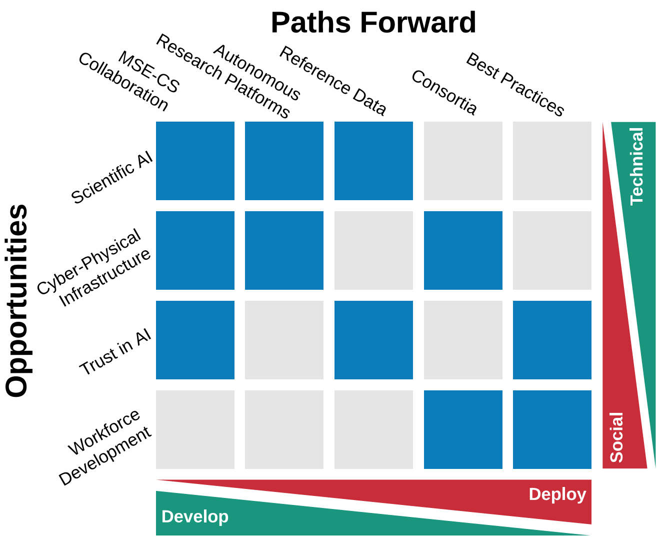
Reference data:

- Develop challenge problems to focus innovation and collaboration on difficult scientific discovery problems, i.e. the materials discovery and design analog to Large Scale Visual Recognition Challenge [42].
- Compile materials datasets with annotated physical rules and heuristics.

3. Opportunities in cyber-physical infrastructure

Realization of scientific AI's potential in materials science and engineering will require advanced cyber-physical infrastructure. We have identified four major opportunities to facilitate this development:

(a) Improved standards and coordination in materials data infrastructure
(b) Development of open and interoperable API-enabled experimental tools



	MSE-CS Collaboration	Autonomous Research Platforms	Reference Data	Consortia	Best Practices
Scientific AI	🟦	🟦	🟦	⬜	⬜
Cyber-Physical Infrastructure	🟦	🟦	⬜	🟦	⬜
Trust in AI	🟦	⬜	🟦	⬜	🟦
Workforce Development	⬜	⬜	⬜	🟦	🟦

Figure 1. Opportunities and proposed paths forward that will enable broader and more effective use of AI in materials science and engineering. The blue intersections represent areas of particular importance to be discussed in this paper.

(c) Development of scalable on-demand synthesis/characterization capabilities
(d) Democratization of research platforms

An improved materials data infrastructure will enable data stewardship throughout the research data lifecycle, which will greatly improve the accessibility of data and metadata to both AI systems and human researchers. Fully automate-able synthesis and characterization tools that execute standardized experimental protocols will improve reproducibility while seamlessly capturing provenance. This will decrease the cost of generating new data and knowledge, and will support real-time distributed and autonomous experimentation. Development of new impedance matched, on-demand synthesis and characterization techniques will be critical to expand the applicability of this approach. The fundamental question is how do we rethink the ‘synthesize-then-characterize’ framework when actionable knowledge can be generated at a rate faster than it takes to transfer the specimens? Finally, we must develop organizational frameworks to democratize access to these new experimental, computational, and data resources, something, comparable to the user facility paradigm at high performance computing centers. Ultimately, this framework would enable scientists and engineers to focus more of their time on conceiving, planning, and executing scientific studies.

3.1. Standards and coordination in materials data infrastructure

Over the past decade, several reports have identified materials data infrastructure as critical gaps limiting innovation in materials research [43–45]. These reports consistently highlight the need for long-term support of shared data services, improved coordination among government agencies, publication of all research data (novel as well as null) with robust metadata, and improved development of community standards for these data and metadata. Findable, Accessible, Interoperable, and Reusable (FAIR) data principles [46] can guide the materials science and engineering community in developing infrastructure suited to collaborative and adaptive research. However, the complexity of materials science and engineering data poses unique challenges to the adoption of FAIR principles. International groups, such as the Research Data Alliance, are fully embracing FAIR Data Principles and are extending them beyond data and metadata, to data types, instruments, and physical samples. Currently, the materials science and engineering community does not have robust frameworks for assigning persistent identifiers to data types, instruments, physical samples, and data and metadata within a larger dataset. Furthermore, once persistent identifiers are assigned on smaller units within a larger dataset, the community will face challenges in effectively and uniformly citing data.

3.2. Open and interoperable API-enabled experimental tools

Critical bottlenecks for adaptive science and autonomous control of experimental systems are (i) a widespread absence of application programming interfaces (API) to interact with laboratory equipment, (ii) lack of a unified language for experimental workflow protocols, and (iii) lack of standardized and open data formats to facilitate accessibility and interoperability. Currently, downstream researchers are developing ad hoc hardware interfaces, duplicating effort and often incurring substantial technical debt. Materials synthesis and characterization workflows are typically manifested in custom software rather than in composable and machine-actionable data representations. Finally, experimental equipment is supported by a diverse collection of vendor-specific interfaces and formats, which may not be well-documented, and may be difficult to use independently from vendor-developed software frameworks. This presents an unnecessary impediment to innovation in real-time data analysis and adaptive experimental planning and control. There is a significant need to facilitate industry-lead development of standards for open and machine-actionable instrument APIs, executable protocols for experimental workflows, and file formats.

3.3. Scalable on-demand synthesis/characterization capabilities

Current materials synthesis and characterization tools are not designed for low latency and high agility between experiments, leading to a significant time-constant mismatch with the algorithmic decision-making that is enabling autonomous experimentation. Currently, an individual high-throughput experimental campaign is restricted to depositing a monolithic combinatorial library under (typically) identical processing conditions and characterizing each sample within the library for the composition, structure, and multiple figures of merit. While high-throughput synthesis techniques enabled revolutionary improvements in the rapid exploration of process-structure-property relationships [47–51], library generation now presents a major bottleneck due to its high latency and the intensity of human labor involved. Therefore, low latency, automated synthesis platforms, integrated with multimodal characterization tools, should be developed. AI also presents unprecedented opportunities for novel adaptive experiments enabled by *in situ* automated perception and data analysis, e.g. through real-time identification, tracking, and subsequent fine-grained analysis of features of interest [52]. For low-latency decision-making, it may be necessary to leverage edge computing [53], e.g. running a deep learning model directly on detector output.

3.4. Resource democratization

Large materials research user facilities (e.g. Advanced Photon Source, NERSC) have demonstrated a model for decoupling the construction and operation of experimental tools and computing infrastructure from the use of those tools by scientific subject matter experts. Similarly, the adaptive synthesis and integrated multimodal characterization platforms described above will require significant capital investment to invent, develop, build, and operate. Therefore, the materials community, and the greater community at large, is presented with an opportunity to develop an organizational and technological framework to facilitate collaboration between theoretical and experimental research groups, and to lower the barrier for cross-material-system, cross-synthesis-method, and cross-modality studies. This framework would also provide increased access to cutting edge experimental materials capabilities to new user communities from underrepresented groups and smaller institutions.

In addition to the cyber-physical infrastructure challenges described above, experimental synthesis and characterization methods are very specific to a given class of materials. There is unlikely to be even one brick and mortar facility to allow researchers to study several materials classes. The Materials Innovation Platforms at the National Science Foundation [54] provide one avenue for resource democratization spanning from predictive synthesis to characterization. These topical platforms are well suited to serve as highly-connected experimental nodes in a research network where information is shared through repositories with community-designed schemas and communication protocols. MIPs might also provide a means of performing the expansive microstructure and interface characterization needed to explore property/performance landscapes across a diverse set of critical materials systems, where microstructure and interfaces strongly mediate material performance.

3.5. Paths forward

Consortia

- Develop community standards to enable FAIR data and equipment interoperability, while learning from successful examples, such as MTConnect
- Design, deploy, operate, and provide democratized access to distributed autonomous laboratory platforms and broader cyber-physical infrastructure, as advocated in the high throughput experimental materials collaboratory (HTE-MC) concept [55].
- Launch new funding initiatives to support creation of materials-focused AI Research Centers and Mission-Driven AI Laboratories as described in [56].

Autonomous materials science

- Design for automation: Rethink the ‘high throughput’ materials synthesis methodology portfolio in light of new capabilities in real-time automated perception, modeling, decision making, and the need for real-time closed-loop feedback from multiple structure and property probes.

- Leverage automation: identify new opportunities to turn *ex situ* analysis methods into AI-driven *in situ* adaptive techniques.

4. Opportunities in trust

Promoting community-wide trust in Scientific AI results is key to reducing the impact of increased disillusionment. We have identified three important opportunities for improving confidence in scientific AI as applied to materials:

(a) Develop and enforce community wide standards for reporting uncertainty from archival data to final model predictions.
(b) Create a scientific culture that values and promotes reproducibility, validation, and verification of published data
(c) Work towards improving the interpretability of AI models and providing a solid foundation towards trust in their predictions

Creating a robust interdisciplinary community spanning MSE and computer science will create opportunities for real-time methods for exploring materials representations, permitting researchers to have confidence that the final model reflects solid physical principles. Generation of reference data sets and materials data challenge problems will allow benchmarking of new models/algorithms using specially designed performance indicators relevant to the specific AI task. Dissemination of best practices will create a community of informed skeptics that request open code and datasets, look for task-appropriate performance indicators, and are alert to issues of dataset and modeling bias.

4.1. Uncertainty: archival data to final model predictions

Current applications of AI in materials science largely ignore the uncertainty of the raw data used to train models. Leveraging larger scale datasets derived from the open literature and published materials databases will require systematic evaluation of source and reporting of measurement and model uncertainty. Reporting uncertainty is introduced by incomplete collection, storage, and/or publication of relevant data, metadata, experimental uncertainties, and potential spurious covariates. At any point where manual annotations from human experts (or non-experts) enter the process, one must also account for annotation uncertainty.

Robust uncertainty quantification is particularly important for robust data fusion and transfer learning based on multiple experimental and simulation-based information sources. These will play an important role in scientific AI for materials research because of the diversity of information needed for modeling complex materials, the relative expense of experiments, and the dominance of simulation results in the current body of openly available materials data. Each material characterization and simulation technique has known ranges of applicability and sources of bias and uncertainty, but these are not typically expressed in a quantitative form amenable to seamless composition of simulation and experiment. This presents an opportunity to develop methods for propagating such uncertainties through AI models, while providing guard rails that alert users to known limitations of the input data. A familiar example is to use caution when interpreting the role of high-throughput DFT bandgaps (which exhibit well-known systematic biases [57]) as model inputs, especially when modeling properties that arise from unrelated phenomena, such as melting temperature. There is a critical need for validation and verification data (e.g. [58]) to benchmark data fusion and transfer learning efforts, and to assess the physicality of the predictions of scientific AI.

The outputs of ML models also have uncertainty related to the model selection and fitting process. This kind of uncertainty must be systematically propagated through a larger pipeline of interlinked physical and ML models. Unbiased assessments of the full model uncertainty from raw data through final predictions are needed to determine with confidence whether it is reasonable to trust the predictions of a machine learning pipeline. Furthermore, well-calibrated uncertainty estimates are crucial to the performance of active learning systems, which rely on quantification of model uncertainty to identify experiments that are likely to be informative.

4.2. Reproducibility, validation, and verification

Ensuring the reproducibility of scientific AI in materials research depends critically on transparency in publication, attention to correct methodology in evaluating results, and independent testing and verification of model predictions [59]. We must develop a strong culture of scrutinizing modeling assumptions, checking for due diligence in training procedures, and verifying that ML models are not being applied outside their regime of applicability.

Recent development of open libraries (Matminer, TPOT), data repositories and platforms (MP, JARVIS, AFLOW, OQMD, Citrination, MDF, NOMAD and AIIDA), and paper repositories are significantly increasing the accessibility and reproducibility of materials research. However, manuscripts often do not fully document model hyperparameters, or the model selection and tuning process used, and data and software are not commonly made available. This can make it difficult to evaluate whether the results suffer from overfitting or information leakage, and impossible for independent verification or comparison with other works. Researchers using AI methods should investigate and publish the failure modes of the models they use, as this can promote improvement in, and trust of, AI methods. For any given modeling task, choice of appropriate performance metrics is of paramount importance [60]. Metrics that account for dataset bias are particularly important in the face of systemic publication bias in favor of 'positive' scientific results, community pursuit of 'lead material' derivatives, and in modeling phenomena governed by rare materials features (such as fatigue crack initiation).

Finally, much of the materials AI literature describes proof-of-concept work applied to a single material system, and experimental validation of predictions is often deferred to followup studies. In contrast, many computer science venues expect methods papers to demonstrate generalizable results on multiple datasets and/or multiple tasks. Addressing this problem will entail finding ways to lower the barrier for groups to collaborate. Creation of a reproducibility and validation consortium would facilitate the collaboration process and potentially lead to extensive use of shared resources throughout the materials research landscape.

4.3. Establishing interpretability and trust

Models and theories are fundamental to the scientific method, and scientists expect to be able to rationalize predictions and discoveries by explaining observations through an underlying phenomenon or mechanism. Thus, the interpretability of scientific AI models is necessary to establish sufficient trust in AI methods for widespread scientific application. Interestingly, trust and interpretability currently lack consensus definitions in computer science and psychology [61]. The challenge of interpretability lies in balancing faithful representation of the model's mechanisms and the ease of intuitive understanding by a human [62], while trust corresponds to a user's willingness to accept or reject model predictions relative to the baseline error rate of the model [61].

The most common scenario in AI-driven materials science involves completely opaque previously-generated models, for example in a process-oriented environment [63]. Here the user does not have access to the full descriptor set or material representation, may not know the model form, and only has access to the final prediction. Thus the distribution of user trust levels may have a large variance. Informed trust in this scenario must be gained through meticulous empirical validation procedures. Feature importance ranking provides some level of insight into a model, but does not go far enough to support claims of physical realism; this interpretability tool is not typically robust in the face of correlated or spurious inputs. Great opportunities exist to boost the interpretability of AI models by providing output in the form of either a human interpretable series of selection criteria (e.g. a simple decision tree/process flow diagram) [64], a set of physically meaningful equations, or a textual explanation. At this level of interpretability expert opinions could be built transparently into the framework through extensive interactions and the trust in the model outputs will be increased.

4.4. Paths forward

Cross-disciplinary Collaboration:

- Develop and deploy real-time algorithms for exploring interpretable material representations during research campaigns.
- Design human-in-the-loop methodologies for quantifying interpretability and trust.

Reference Data:

- Develop and adopt common benchmark datasets and performance indicators for measuring and comparing methodological progress.
- Create dedicated funding mechanisms for experimental validation of materials predictions

Best Practices:

- Reviewers insist on full and open access publication of source code, machine-readable training data, and artifacts such as trained models from publicly funded research.

- Reviewers insist on task-appropriate performance indicators and fully-documented research protocol.
- Identify and quantify bias / variance issues in datasets
- Assess dataset and source bias through round-robin type studies to establish reproducible results.
- Create community accepted benchmarks for fusing experimental and computational data with uncertainty and applicability propagation through the model training, testing, and interpretation pipeline.

5. Opportunities in workforce development

There is an urgent need for workforce development to ensure that AI techniques are introduced into the materials science workflow with the appropriate level of scientific rigor. Briefly, there are opportunities in:

(a) Educating the next generation workforce to be conversant in AI techniques and their application to materials science.
(b) Expanding skills within the current workforce, enabling them to effectively mentor the next-generation workforce.
(c) Adopting an open data culture

Here consortia will play an important role by developing open source educational materials, hosting bootcamps, and introducing workshop tracks at professional meetings, creating learning opportunities and materials that can be disseminated up and down the educational tiers. [65] summarizes the current status in these areas. Likewise the definition, publication, and demonstration of best practices in AI (e.g. [5]) will go a long way to increasing awareness and trust within the community.

5.1. Educating the next-generation workforce

Materials science curricula are in need of urgent restructuring to produce a competitive next generation workforce [65, 66]. This restructuring needs to take into account the level of skills transferability needed throughout the overall data landscape, in addition to direct application to materials research. Traditional materials science education contains few required courses in statistical methods and programming. This is a major limitation on the adoption of ML techniques by the materials science community, as graduates lack foundational knowledge and skills.

At the undergraduate level, there are few treatments in the application of AI to materials science available for developing course modules, though graduate programs and standalone summer courses are rapidly expanding [65]. One critical need is the development of open data/code repositories that provide ‘plug and play’ modules to augment the current materials science undergraduate curriculum. GaTech [67] and recently developed ML content on nanoHUB [68–70] are excellent early examples of this educational model. Open educational resources, in addition to formalized courses providing a more rigorous introduction to research computing and statistical methods, are needed to create a BS-level workforce capable of implementing ML.

5.2. Expanding skills within the current workforce

At the graduate and post-graduate level, there is an urgent need for providing salient feedback on the relevance of models and their outputs. Mid to late career materials scientists might feel unprepared to mentor researchers applying ML techniques to their research. This can lead to naively trusting (or dismissing out-of-hand) results from ML workflows, or feeling unequipped to practice informed skepticism.

There is great need for a new professional track in the materials field, since federally-funded data-intensive centers and facilities will start building both physical and cyber data infrastructure. At present, the availability of data technicians is minimal. Establishing a few training pilots across the country amongst undergraduate institutions and community colleges will provide the needed workforce to accomplish this task. If the MGI/AI visions are to be realized, there is an immediate urgency for workforce training pilots.

5.3. Adopting an open data culture

Data sciences have proven to be a democratizer in a variety of fields, notably in astronomy, bioinformatics, and high energy physics. In an open data culture, data and metadata are rigorously acquired and deposited in standardized open repositories, also accessible to low-capacity research institutions. Adopting the open data paradigm will afford community colleges and low-capacity research institutions membership to the materials research community and greatly contribute to diversity.

5.4. Paths Forward

Consortia:

- Develop open source educational materials (e.g. https://datacarpentry.org/) broadly targeting other opportunities throughout this document. Educational materials should leverage and reference known best practices. The materials, associated data, and code should be promoted in a way that they are indexed by internet search engines to maximize visibility.
- Host bootcamps (e.g. NIST's MLMR), webinars, and hackathons framed around AI usage in materials science
- Introduce a 'workshop track' at major materials conferences for students and researchers to acquire and practice new skills.
- Support internships (e.g. NIST's SURF program) for students and researchers to develop real-world experience.

Best Practices:

- Engage with stakeholders to define, publish, and demonstrate best practices in the use of AI in Materials science and engineering

6. Summary and outlook

In the previous sections, we provided a perspective on the application of AI in materials science and engineering and outlined four overarching opportunities for potential advancement within the community. We have proposed five cross-cutting paths forward for each opportunity:

Cross-disciplinary Collaboration - Reinvigorated collaboration among the domains of materials science and engineering, computer science, and data science will advance state-of-the-art solutions for scientific AI and cyber-physical infrastructure while enabling trust in AI.

Autonomous Research Platforms - The development and deployment of diverse autonomous research platforms will enable implementation and evaluation of new technology in scientific AI and cyber-physical infrastructure by the rapid generation of high-quality experimental materials data. Connecting these platforms will create compound network effects that increase the leverage of any single experiment or calculation.

Reference Data - The creation of new reference and challenge datasets will enable the broader community to develop scientific AI and increase trust in AI, just as the classic MNIST handwritten digit database [71] has enabled these outcomes in the broader STEM community.

Consortia - The creation of new consortia will engage stakeholders in industry, government and academia to provide economically sustainable frameworks for the deployment and operation of cyber-physical infrastructure and expanding the AI skills of the current and future workforce, which will boost consortia member trust in AI.

Best Practices - The creation of stakeholder-lead standards and best practices will enable trust in AI and foster a workforce that understands how to use AI effectively.

If the community makes coordinated efforts in these areas, we can anticipate rapid acceleration of materials discovery and process optimization, which will open new pathways for technological advancement in sustainable development, transportation, water security, medicine, and other technologies central to human welfare.

Disclosures

This article was prepared while E M Campo was employed at the National Science Foundation. The opinions expressed in this article are the author's own and do not reflect the view of the National Science Foundation.

Data availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

ORCID iDs

BL DeCost https://orcid.org/0000-0002-3459-5888
JR Hattrick-Simpers https://orcid.org/0000-0003-2937-3188

Z Trautt https://orcid.org/0000-0001-5929-0354
AG Kusne https://orcid.org/0000-0001-8904-2087
E Campo https://orcid.org/0000-0002-9808-4112
ML Green https://orcid.org/0000-0001-7871-3148

References

[1] Agrawal A and 2016 Choudhary A Perspective: Materials informatics and big data: Realization of the 'fourth paradigm' of science in materials science *APL Mater.* **4** 053208

[2] Kalidindi S R and De Graef M 2015 Materials data science: current status and future outlook *Ann. Rev. Mater. Res.* **45** 171–93

[3] Dimiduk D M, Holm E A and Niezgoda S R 2018 Perspectives on the impact of machine learning, deep learning and artificial intelligence on materials, processes and structures engineering *Integrating Mater. Manufacturing Innovation* **7** 157–72

[4] Schmidt J, Marques M R, Botti S and Marques M A 2019 Recent advances and applications of machine learning in solid-state materials science *npj Computat. Mater.* **5** 1–36

[5] Wang A Y T, Murdock R J, Kauwe S K, Oliynyk A O, Gurlo A, Brgoch J, Persson K A and Sparks T D 2020 Machine learning for materials scientists: An introductory guide towards best practices *Chem. Mater.* **32** 4954–65

[6] Holdren J P *et al* 2014 *Materials Genome Initiative Strategic Plan* (Washington DC: Office of Science and Technology Policy) vol 6

[7] Aspuru-Guzik A and Persson K 2018 Materials acceleration platform: Accelerating advanced energy materials discovery by integrating high-throughput methods and artificial intelligence *Mission Innovation: Innovation Challenge* **6** US Department of Energy

[8] DOE Advanced Manufacturing Office 2017 Workshop on artificial intelligence applied to materials discovery and design *Technical Report* US Department of Energy

[9] DOE Office of Scientific and Technical Information 2019 Workshop report on basic research needs for scientific machine learning: Core technologies for artificial intelligence *Technical Report* US Department of Energy (https://doi.org/10.2172/1478744)

[10] Aziza B 2018 Machine learning and data – where you'd least expect it *Forbes Magazine* (https://www.forbes.com/sites/ciocentral/2018/10/30/machine-learning-data-where-youd-least-expect-it/#5c4fc16e9871)

[11] Lookman T, Alexander F J and Bishop A R 2016 Perspective: Codesign for materials science: An optimal learning approach *APL Mater.* **4** 053501

[12] Ren F, Ward L, Williams T, Laws K J, Wolverton C, Hattrick-Simpers J and Mehta A 2018 Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments *Sci. Adv.* **4** eaaq1566

[13] Hussey S, Placek P and Schack C 1963 An introduction to statistical design of experiments in metallurgical research *Technical Report No.* ADA070811 Bureau of Mines Washington DC

[14] De Wilde W P and Sol H 1987 Anisotropic material identification using measured resonant frequencies of rectangular composite plates *Composite Structures 4: Damage Assessment and Material Evaluation* vol 2 (Dordrecht: Springer Netherlands) pp 317–24

[15] Burati jr J, Antle C and Willenbrock J 1983 Development of a Bayesian acceptance approach for bituminous pavements *Transport. Res. Record* **924** 64–71

[16] Teti R and Caprino G 1994 Prediction of composite laminate residual strength based on a neural network approach *WIT Trans. on Inform. Commun. Technol.* **6** 81–88

[17] Bhadeshia H K D H 1999 Neural networks in materials science *ISIJ Int.* **39** 966–79

[18] Long C, Hattrick-Simpers J, Murakami M, Srivastava R, Takeuchi I, Karen V L and Li X 2007 Rapid structural mapping of ternary metallic alloy systems using the combinatorial approach and cluster analysis *Rev. Sci. Instrum.* **78** 072217

[19] Long C, Bunker D, Li X, Karen V and Takeuchi I 2009 Rapid identification of structural phases in combinatorial thin-film libraries using x-ray diffraction and non-negative matrix factorization *Rev. Sci. Instrum.* **80** 103902

[20] Kusne A G *et al* 2014 On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets *Sci. Rep.* **4** 6367

[21] Suram S K *et al et al* 2016 Automated phase mapping with AgileFD and its application to light absorber discovery in the V–Mn–Nb oxide system *ACS Comb. Sci.* **19** 37–46

[22] Koinuma H 2002 Combinatorial materials research projects in Japan *Appl. Surf. Sci.* **189** 179–87

[23] Lippmaa M, Meguro S, Ohnishi T, Koinuma H and Takeuchi I 2005 On-line data management for high-throughput experimentation *MRS Online Proc. Library Archive* **894** 0894-LL09-07

[24] Chikyow T 2006 Trends in materials informatics in research on inorganic materials *Technical Report* NISTEP Science Technology Foresight Center

[25] Nikolaev P, Hooper D, Webber F, Rao R, Decker K, Krein M, Poleski J, Barto R and Maruyama B 2016 Autonomy in materials research: a case study in carbon nanotube growth *npj Computat. Mater.* **2** 16031

[26] Sanchez-Lengeling B and Aspuru-Guzik A 2018 Inverse molecular design using machine learning: Generative models for matter engineering *Science* **361** 360–5

[27] Dunn A, Brenneck J and Jain A 2019 Rocketsled: a software library for optimizing high-throughput computational searches *J. Phys. Mater.* **2** 034002

[28] Talapatra A, Boluki S, Duong T, Qian X, Dougherty E and Arróyave R 2018 Autonomous efficient experiment design for materials discovery with Bayesian model averaging *Phys. Rev. Mater.* **2** 113803

[29] Gongora A E, Xu B, Perry W, Okoye C, Riley P, Reyes K G, Morgan E F and Brown K A A 2020 Bayesian experimental autonomous researcher for mechanical design *Sci. Adv.* **6** eaaz1708

[30] Kohavi R and Wolpert D H *et al* 1996 Bias plus variance decomposition for zero-one loss functions *ICML* **96** 275–83

[31] Anselmi F, Evangelopoulos G, Rosasco L and Poggio T 2019 Symmetry-adapted representation learning *Pattern Recognit.* **86** 201–8

[32] Senior A W *et al* 2020 Improved protein structure prediction using potentials from deep learning *Nature* www.ncbi.nlm.nih.gov/pubmed/31942072

[33] Chen T Q, Rubanova Y, Bettencourt J and Duvenaud D K 2018 Neural ordinary differential equations *Adv. Neural Inform. Process. Syst.* **31** 6571–83

[34] Rackauckas C, Ma Y, Martensen J, Warner C, Zubov K, Supekar R, Skinner D and Ramadhan A 2020 Universal differential equations for scientific machine learning (Preprint arXiv: 2001.04385)

[35] Innes M, Edelman A, Fischer K, Rackauckus C, Saba E, Shah V B and Tebbutt W 2019 Zygote: A differentiable programming system to bridge machine learning and scientific computing arXiv preprint arXiv: 1907.07587

[36] Pun G P, Batra R, Ramprasad R and Mishin Y 2019 Physically informed artificial neural networks for atomistic modeling of materials *Nat. Commun.* **10** 2339

[37] Ouyang L, Tessler M H, Ly D and Goodman N 2016 Practical optimal experiment design with probabilistic programs *CoRR* arXiv: 1608.05046 [cs.AI]

[38] Jia X 2019 Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis *Nature* **573** 251–5

[39] Heckerman D, Meek C and Cooper G 1999 A Bayesian approach to causal discovery *Computat. Causation Discovery* **19** 141–66

[40] Vajda S 1972 *Probabilistic Programming* (New York: Academic Press) (https://doi.org/10.1016/C2013-0-11637-6)

[41] Shechtman D 2013 Quasi-periodic crystals—the long road from discovery to acceptance *Rambam Maimonides Med. J.* **4** e0002

[42] Russakovsky O *et al et al* 2015 Imagenet large scale visual recognition challenge *Int. J. Comput. Vis.* **115** 211–52

[43] Tinkle S, McDowell D, Barnard A, Gygi F and Littlewood P 2013 Sharing data in materials science *Nature* **503** 463–4

[44] Ward C H and Warren J A 2015 *Materials Genome Initiative: Materials Data* US Department of Commerce, National Institute of Standards and Technology)

[45] Jain A, Persson K A and Ceder G 2016 Research update: The materials genome initiative: Data sharing and the impact of collaborative ab initio databases *APL Mater.* **4** 053102

[46] Wilkinson M D *et al* 2016 The fair guiding principles for scientific data management and stewardship *Scientific Data* **3** 160018

[47] Simon C G and Lin-Gibson S 2011 Combinatorial and High-Throughput Screening of Biomaterials *Adv. Mater.* **23** 369–87

[48] Green M L, Takeuchi I and Hattrick-Simpers J R 2013 Applications of high throughput (combinatorial) methodologies to electronic, magnetic, optical and energy-related materials *J. Appl. Phys.* **113** 231101

[49] Potyrailo R and Mirsky V M 2008 Combinatorial and high-throughput development of sensing materials: the first 10 years *Chem. Rev.* **108** 770–813

[50] Maier W F, Stöwe K and Sieg S 2007 Combinatorial and high-throughput materials science *Angewandte Chemie (Int. Ed. English)* **46** 6016–67

[51] Potyrailo R, Rajan K, Stoewe K, Takeuchi I, Chisholm B and Lam H 2011 Combinatorial and high-throughput screening of materials libraries: review of state of the art *ACS Combinatorial Sc.* **13** 579–633

[52] Burnett T and Withers P 2019 Completing the picture through correlative characterization *Nat. Mater.* **1** 1041–9

[53] Shi W, Cao J, Zhang Q, Li Y and Xu L 2016 Edge computing: Vision and challenges *IEEE Internet Things J.* **3** 637–46

[54] NSF 2018 Program Solicitation NSF 19-526: Materials Innovation Platforms *Technical Report* National Science Foundation. (https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505133)

[55] DOE Office of Energy Efficiency and Renewable Energy 2018 Workshop on advanced energy materials discovery, development, and process design utilizing high-throughput experimental methods, artificial intelligence, autonomous systems, and a collaboratory network *Technical Report* US Department of Energy

[56] Gil Y and Selman B A 2019 20-year community roadmap for artificial intelligence research in the us arXiv preprint arXiv: 1908.02624

[57] Cohen A J, Mori-Sánchez P and Yang W 2008 Insights into current limitations of density functional theory *Science* **321** 792–4

[58] Kirklin S, Saal J E, Meredig B, Thompson A, Doak J W, Aykol M, Rühl S and Wolverton C 2015 The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies *npj Computational Materials* **1** 1–15

[59] Hanisch R J, Gilmore I S and Plant A L 2019 Improving reproducibility in research: The role of measurement science *J. Res. Natl Inst. Stand. Technol.* **124** 1–13

[60] Riley P 2019 Three pitfalls to avoid in machine learning Nature **572**, 27–29

[61] Schmidt P and Biessmann F 2019 Quantifying interpretability and trust in machine learning systems arXiv preprint arXiv: 1901.08558

[62] Herman B 2017 The promise and peril of human evaluation for model interpretability arXiv preprint arXiv: (1711.07414)

[63] Holm E A 2019 In defense of the black box *Science* **364** 26–7

[64] Raccuglia P *et al* 2016 Machine-learning-assisted materials discovery using failed experiments *Nature* **533** 73

[65] The Minerals Metals & Materials Society (TMS) 2019 *Creating the Next-Generation Materials Genome Initiative Workforce* (Pittsburgh, PA: TMS) dx.doi.org/10.7449/mgiworkforce_1

[66] Alekseeva L, Azar J, Gine M, Samila S and Taska B 2020 The demand for AI skills in the labor market (https://cepr.org/active/publications/discussion_papers/dp.php?dpno=14320)

[67] Kalidindi S 2018 Materials data sciences and informatics (accessed: 13/ 05/2020) (https://www.coursera.org/learn/material-informatics)

[68] Klimeck G, McLennan M, Brophy S P, Adams III G B and Lundstrom M S 2008 nanohub. org: Advancing education and research in nanotechnology *Computing Sci. Eng.* **10** 17–23

[69] Gastelum J C V, Strachan A and Desai S 2019 Machine learning for materials science: Part 1 (West Lafayette, IN: Purdue University) (https://nanohub.org/resources/mseml)

[70] Gastelum J C V and Strachan A 2019 Citrine tools for materials informatics (West Lafayette, IN: Purdue University) (https://nanohub.org/resources/citrinetools)

[71] Grother P J 1995 NIST Special Database 19 - NIST Handprinted Forms and Characters Database