Integrating Materials and Manufacturing Innovation (2024) 13:83–91
https://doi.org/10.1007/s40192-023-00330-6

THEMATIC SECTION: HARNESSING THE POWER OF MATERIALS DATA

Beyond Combinatorial Materials Science: The 100 Prisoners Problem

J. Elliott Fowler[1] · Matthew A. Kottwitz[1] · Nat Trask[2,4] · Rémi Dingreville[3]

Received: 20 October 2023 / Accepted: 28 November 2023 / Published online: 8 January 2024

Abstract
Advancements in high-throughput data generation and physics-informed artificial intelligence and machine-learning algorithms are rapidly challenging the status quo for how materials data is collected, analyzed, and communicated with the world. Machine-learning algorithms can be executed in just a few lines of code by researchers with minimal data science expertise. This perspective addresses the reality that the ecosystems which have been constructed to nurture new materials discovery and development are not yet well equipped to take advantage of the radically more powerful and accessible computational and algorithmic tools which have the immediate potential to enhance the pace of scientific advancement in this field. A novel architecture for managing materials data is proposed and discussed from the standpoint of how historical and emerging subfields of materials science could have been or might still significantly improve the impact of materials discoveries to the many human societal needs for new materials.

Keywords Combinatorial materials science · Data management · Machine learning · Materials characterization

Introduction

For decades, there has been a push in the materials science community to develop novel materials and fabrication protocols that meet future societal challenges ranging from sustainability to clean energy to microelectronics [1–6]. Despite billions of dollars in public and private research investments over that period, it has been challenging to move the needle and enable the realization of "Moore's Law for Scientific Discovery" [7]. Breakthroughs in computational power and the resulting improvement in efficiency of engineering hardware led to a renaissance in combinatorial chemistry and materials science in the early 2000s and 2010s, respectively. Excitement for opportunity, however, quickly transitioned to realization within the field of materials science that vastly more data did not natively translate to vastly more information. Rather, unguided combinatorics often led to libraries of data on infeasible, or inconsequential materials [8, 9]. Thus, the age of Machine Learning and Artificial Intelligence-led chemistry and materials science has dawned in the 2020s, with promises of curing the woes of unbridled combinatorics with guided exploration and exploitation of materials. Figure 1 shows how the body and momentum of literature on combinatorial and artificial intelligence-focused chemistry and materials science has evolved over the last 50 years.

The 100 Prisoners Problem

Material correlations which capture the complex relationships investigated in materials science experimentation are commonly referred to as process–structure–property–performance (psp^2) relationships. Owing to the realities of stochasticity, heterogeneity, and the enormous divergence in length scales between the number of atoms that form a molecule of material and the number of atoms in relevant engineering quantities of a material's use, psp^2 relationships are extraordinarily high dimensional. As an analogy, consider the recently re-popularized mathematical riddle commonly

✉ J. Elliott Fowler
jfowle@sandia.gov

1 Sandia National Laboratories, Materials Performance and Characterization Department, Albuquerque, NM 87123, USA

2 Sandia National Laboratories, Computational Mathematics Department, Albuquerque, NM 87123, USA

3 Sandia National Laboratories, Nanostructure Physics Department, Center for Integrated Nanotechnologies, Albuquerque, NM 87123, USA

4 Mechanical Engineering and Applied Mechanics, University of Pennsylvania, Philadelphia, PA 19104, USA

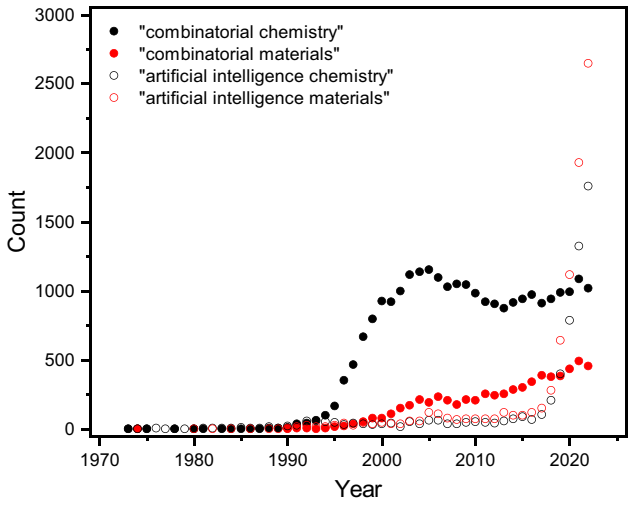

Fig. 1 Scale and trajectory of Web of Science search results for 'combinatorial' and 'artificial intelligence' associated with 'chemistry' and 'materials', as of March 2023

known as the "100 Prisoners Problem" [10, 11]. A version of this riddle adapted from Analytic Combinatorics reads [11]:

> "The director of a prison offers 100 death row prisoners, who are numbered from 1 to 100, a last chance. A room contains a cupboard with 100 drawers. The director randomly puts one prisoner's number in each closed drawer. The prisoners enter the room, one after another. Each prisoner may open and look into 50 drawers in any order. The drawers are closed again afterwards. If, during this search, every prisoner finds his number in one of the drawers, all prisoners are pardoned. If just one prisoner does not find his number, all prisoners die. Before the first prisoner enters the room, the prisoners may discuss strategy—but may not communicate once the first prisoner enters to look in the drawers. What is the prisoners best strategy?"

If the prisoners chose fifty drawers at random, their collective probability of success is calculated from their probability of success in any choice ($50/100 = 0.5$) and the total number of choices (100), which simplifies to $P(x) = 0.5^{100}$ or a miniscule 7.8×10^{-31} : 1—*substantially smaller than the probability of finding a single atom of interest among one mole of that substance.*

Even if one were to limit themselves to considering only Newtonian mechanics, the number of chemical, mechanical, electrical, magnetic and other properties for a single material easily surpasses 100 dimensions. Exacerbating the complexity, materials themselves do not always reveal their true behavior easily. Opening the "box" may not be as simple as a quick measurement. Some material properties of interest require extraordinary effort and time to measure. Moreover, materials are naturally stochastic—each one has different properties according to the complex chaotic processes by which it was formed. Material phenomena controlled by statistically rare events have been described as "black swan" events, a reference to their statistically unexpected nature and the natural human bias to underestimate the importance of the tails of a distribution [12, 13]. Considering the stochastic nature of materials, finding the right "box" is not as simple as opening a box and looking inside. Hundreds or thousands of observations may be necessary to isolate rare events, and each of those experiments or simulations can be expensive and time-consuming.

Beyond Combinatorial Materials Science

Recognizing the grim realities of combinatorial explosion, materials and data scientists have begun pursuing a solution which takes advantage of, rather than suffering from dimensionality. That is, complex psp^2 relationships can be represented in reduced order latent spaces such as those used by artificial intelligence and machine learning algorithms to model large sums of data efficiently, and inferences can be made about novel psp^2 relationships that would otherwise be beyond expert cognition [14, 15]. Notably, there has been a recent rise in 'scientific' machine learning. Scientific machine learning aims to incorporate physics constraints and models into AI/ML to enhance efficiency, applicability and explainability of reduced-order relationships for domain-specific problems (such as fluid flow, plasma physics, etc.) [16, 17]. Breakthroughs continue to be made, such as Bengio's Generative Flow Networks which promise to greatly enhance the efficiency of exploring very high-dimensional search spaces [18].

Briefly returning to the riddle, the prisoners derive negligible benefit from 50 opportunities to open drawers without an underlying strategy guiding them to turn data (in the form of drawer contents) into actionable information (to find their number). If the prisoners chose to collaborate on a strategy whereby each prisoner starts by opening the drawer corresponding to their own number and follows that loop (next opening the drawer corresponding to the first number they find, *ad nauseum*) until they either find their number or reach the fifty-drawer limit, the collective probability of success inflates substantially, to roughly 0.31 : 1. In other words, the probability of failure is capped by the probability that any loop the prisoners follow is greater than 50 boxes long. The prisoners, or the materials scientists, still fail more than they succeed, but with this approach intelligently optimize their chances of success.

To effectively navigate this combinatorial complexity, a common requirement for modern AI/ML frameworks is the ability to make informed decisions based on limited information and efficiently explore a vast search space. Brute-force methods such as Monte Carlo or other statistical

methods become intractable as size of the problem increases. Recent advancements in AI/ML have opened new avenues for developing algorithms that can tackle combinatorial problems more effectively. The primary idea is that these algorithms can learn from existing (limited) data, identify patterns within the datasets, and make predictions in order to guide the search process toward promising solutions [19–21]. One promising approach utilizes Bayesian optimization by combining probabilistic modeling with optimization techniques (Fig. 2a). These algorithms build a model of the objective function based on past evaluations and use this probabilistic model to select the next point to evaluate, balancing the exploration and exploitation of the data [22, 23]. This approach has been successfully used for solving constrained multi-objective materials design problems for the discovery of novel alloys over a large compositional space [24]. Another approach is to employ reinforcement learning algorithms, which learn by interacting with an environment and receiving rewards or penalties for their actions. This class of algorithm adapts its search based on feedback to effectively navigate complex design space. This strategy has been recently used for example to predict optimal synthesis protocols for the synthesis of semiconducting monolayer MoS_2 film via chemical vapor deposition [25]. In this case, the reinforcement-learning agent learned deposition conditions in terms of temperature and chemical potentials for onset of chemical reactions and predicted unknown synthesis schedules. Both Bayesian optimization and reinforcement learning algorithms provide a principled framework for decision making to make the 100 prisoners problem in materials science more tractable.

Another emerging approach is to leverage subject matter expertise to design high-throughput experiments (synthesis and characterization) which produce large, information-rich materials databases that can subsequently serve as input(s) into supervised and unsupervised learning models to discover new, hidden relationships between materials processes, structures and properties (Fig. 2b). These hidden relationships (or 'fingerprints') exist due to the sheer complexity and dimensionality of materials descriptors and a general inability of human experts to cognate beyond a few dimensions simultaneously [26]. With this approach, it was shown that novel materials and processes, such as photovoltaic thin-films and high-entropy alloys, were discovered using an existing or freshly generated materials database without requiring a specific objective to optimize towards [27, 28]. For the 100 prisoners, this approach would be like using collective wisdom to discover the successful path toward freedom out of the near-infinite ocean of failing paths, irrespective of the arrangement of the boxes and slips of paper in their particular scenario.

Finally, the 100 prisoners might decide that they can discover the most effective solution to the Problem by employing an informational sleight-of-hand. In this case, the correct solution maybe be too time-consuming or complex to 'discover' and communicate to every other prisoner. An

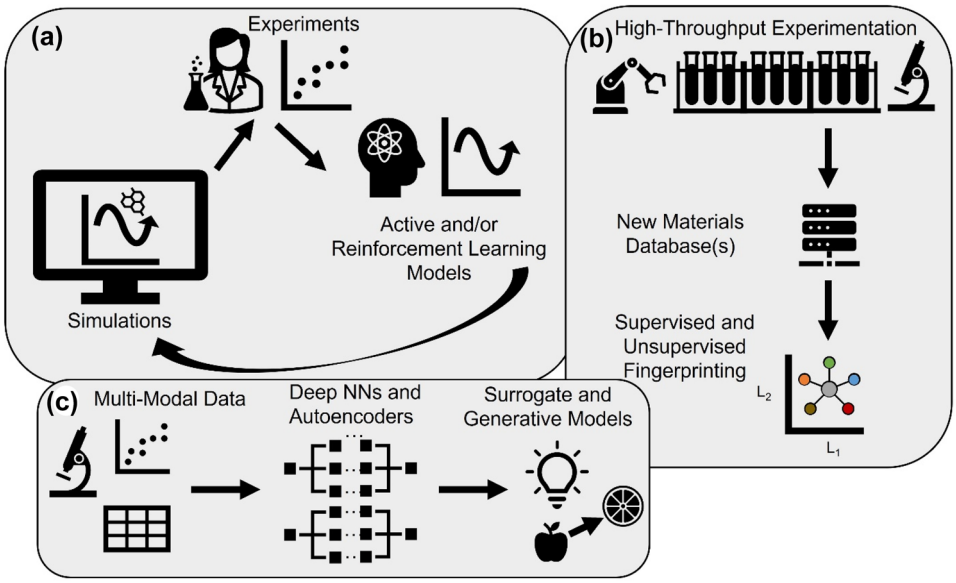

Fig. 2 Three examples of existing frameworks for AI/ML-guided materials discovery: **a** chemical simulations, such as DFT and based upon human intuition or algorithmic optimization protocols, are run to identify new materials of interest, which are then synthesized, characterized and used to identify new classes of materials to simulate; **b** A large variety of materials are synthesized and characterized in a high throughput manner in order to build new materials databases which are explored in disentangled latent spaces to discover new materials relationships; **c** Multi-modal data are fed to deep neural networks or autoencoders which develop surrogate or generative models between two disparate types of materials information

emerging solution to this challenge for the prisoners or the materials community is to utilize models which can act as surrogates for or generate predictions of one type of information from another, thus cutting down the complexity of the search space significantly (Fig. 2c). It has been shown that multi-modal data types can be fused together in deep neural networks (DNN) and variational autoencoders to discover novel connections between them that enable replacement of one (more complex) kind of information for another (less complex) type [29, 30]. For example, it has been shown that a DNN model could replace a computationally intensive finite-element model to accurately predict structural material properties [30].

Without doubt, AI/ML has already been shown as a proven strategy to reduce the experimental burden of solving the world's most important materials science challenges. However, recognizing the solution is only a small part of the task. *The next 'big challenge' in the integration of materials science and AI/ML, to accelerate materials discovery and utility, is addressing the social and engineering challenges that hinder the implementation of more efficient, AI/ML-enabled materials discovery platforms.* In the following, we identify and propose solutions for the biggest social and engineering challenges standing in the way of AI/ML-enabled materials discovery.

Social Challenges: Communication, Conflict and Communities of Practice (C^3)

The social challenges facing the field of AI/ML-enabled materials discovery can largely be divided into those regarding efficient *communication*, prevention or mitigation of *conflict*, and the establishment of diverse and effective *communities of practice*. While each of these subjects has been addressed to varying extents by others, it is important to consider them together as a social strategy for materials discovery.

Communication

In materials science, unexpected dead ends happen more often than researchers would like to admit. For example, months of research into a novel materials chemistry may yield a conclusion that the synthesis pathway will not feasibly yield a material at relevant quantities, and a research team must pivot to a new approach. The daunting nature of discovery and its high failure rate, compounded by funding agency expectations to achieve returns of research investments, have led modern discovery science to be exceptionally risk averse [31, 32]. However, risk aversion is antithetical to discovery [33]. Ultimately, funding agencies, universities, senior research leadership and principal investigators must engage in more open and honest dialog on the expectations of discovery science, and failure must be embraced as an information gain rather than a monetary loss.

Siloing is another monumental communication challenge. For one, discovery of most classes of materials is rarely—if ever—siloed into a single research lab and can be quite spread out not only geographically but also in desired outcomes. Generally, discovery benefits from borrowing and permutating on successful paths taken from other groups pursuing similar goals. Contrary to this, the current paradigm for data sharing is through after-the-fact publishing of research findings in peer reviewed publications. There are incentives for some communities to withhold portions of data or methodology which may contain potential intellectual property or work intended for future publication. The reality of commercial and academic scientific structures in the real world are in direct conflict with an optimal paradigm for materials discovery.

Even without the constraints on communication that exist because of real-world incentives, Brooks Law informs us that communication will almost always become a bottleneck as the number of collaborators grows [34, 35]. Therefore, the democratization of professional materials science must be pursued together with the emergence of low barrier to entry sharing platforms. The Materials Data Facility is an excellent early example of the kind of platform needed; *however, peer reviewed journals—particularly open access journals—must continue adding requirements for sharing data as part of the publishing process and principle investigators must make data sharing an innate part of scientific communication* [36, 37].

Conflict

Issues with communication naturally lead to conflict. Conflict between optimal and realistic discovery environments has a direct impact on the spin-up costs of AI/ML-enabled autonomous materials discovery. For one, the reality of sparsely communicated collective knowledge in the materials science community implies that optimal training data sets may almost never exist a priori to train a desired AI/ML model. A few notable exceptions exist, for example NIST's mass spectrometry libraries, Argonne National Lab's Modular Constitutive Modeling Library for Structural Materials, and the Materials Genome Initiative [38–40]. The ethical debate on equal access to scientific domain knowledge has led to notable policy changes by the United States government in 2022 to enforce open access on publicly financed research [41]. Overall, these are exceptions that demonstrate the enormous undertaking required to retrieve and compile collective knowledge at scale.

In some ways, limited access to domain knowledge has led to a recent surge in interest in the use of Deep Transfer

Learning and Generative Adversarial Networks which can train from large unrelated data sets or small initial data sets where a generative model can be leveraged to enhance the network [42–46]. Additionally, many materials science problems can leverage the possibility of ab initio and even multi-scale predicted data sets using first principles calculations. However, the trade-off for these solutions is a generalized need to increase the scope and resource costs for expertise and computing power included on a project aimed at any materials discovery. *To mitigate the natural challenges of conflict in materials discovering, building open, valuable, voluminous, and diverse training data sets must be built into the cost of discovery.*

Communities of Practice

For the 100 prisoners, building communities of practice would be a bit like consulting a statistician and a clairvoyant during the planning stage—unlikely to be approved by the prison warden. However, materials scientists are not prisoners and can build communities of expertise which could optimize multiple, converging approaches toward a successful outcome(s) in materials discovery. As recent examples, Berlinguette et al.'s development of self-driving laboratories for the accelerated discovery of adhesive and thin film materials involved a community of subject matter experts from multiple universities, institutes, and industry and was supported by an equally diverse set of funding sources [47, 48]. The concept is now commonly referred to as 'co-design'—involving stakeholders such as engineers and industry in the material development process. *Co-design illustrates the need for investment areas and materials science leaders to accept that high-consequence materials discovery (i.e., low probability of success) will require sufficiently diverse expertise and funding to support broad coalitions of collaborators to accomplish truly aspirational goals.*

Engineering Challenges: Stewardship, Standard Data Architectures, and Science and Technology Innovation Centers (S^3)

In addition to social challenges, there exist opportunities to engineer systems and structures that empower AI/ML-enabled materials discovery platforms. Engineering opportunities include *stewardship* of training and mentoring programs at all levels of education, establishment of *standard data architectures* for improved organization, handling and preprocessing of data files, and foundation of *science and technology innovation centers* which dedicate facilities, labor, equipment, and computing resources to support virtual-hardware-virtual ecosystems to integrate AI/ML and materials science in an autonomous and/or artificially intelligent way.

Stewardship

As it relates to the 100 prisoners' problem and stewardship, there's a question of the expected number of drawers that'll need to be investigated before a solution is found, far from an easy thing to know beforehand. Below a certain number, which will likely be different for every project, there's a certain up-front cost to generating large data sets and selecting and training an algorithm that will surpass that of a more traditional approach. Beyond that point the ML assistance is justified. Developing a familiarity among data generators (materials scientists) with constraints grappled with by data scientists in developing AI/ML-guided solutions—so that all problems aren't viewed as a nail and AI/ML the hammer—is an overlooked, critical step.

Imagine a near future in which programming language instruction is as normalized in education as learning a foreign language. Interacting effectively with structured databases and AI/ML algorithms requires a baseline knowledge of programming languages that may be addressed at all levels of education. The availability of foundational computer science courses has increased in primary and secondary educational systems, though opportunities exist to address disparities in offerings and enrollment of groups underrepresented in STEM fields [49, 50]. Once in higher education, students should be given the opportunity to continue their education in more advanced topics and apply those skills in other courses outside the confines of a computer science program. Educators are encouraged to bring a foundational materials science concept—stochasticity—to all fields of STEM, which may ultimately encourage greater bridge-building between the sciences and applied mathematics. Finally, effective mentor–mentee relationships in the workplace will help ensure that the barriers to entry for data science at all stages of learning remain reasonably low. *Addressing stewardship requires material scientists to actively engage the data sciences at all levels of hypothesis generation, experimental design and data analysis.*

Standard Data Architectures

The prisoners can't read each other's minds to discover which drawers they've previously investigated. Likewise, algorithms can't read details that've only been scribbled in a lab notebook and raw, digital data rarely starts in a consumable form for input into AI/ML algorithms. Materials characterization tools are generally designed for the niche fields of science in which they are most applied. Those niche fields carry with them nuanced data treatment practices supported by decades of literature precedence. As technology—particularly software and computational speeds—has advanced it has become increasingly likely that exported data has already undergone some form of normalization,

pre-processing or subject matter expert-facilitated analysis by the time it is shared with a broader community [51]. In many cases the native data files from an instrument intentionally obscure the raw (i.e., columnated for *x*–*y* data, unfiltered or uncompressed images) data behind proprietary file types or layers of compression algorithms [52]. Materials scientists commonly generate non-homogenous (i.e., different native resolutions, variable ranges, or test conditions) data sets, particularly in pursuit of higher data fidelity for publications.

Figure 3a and b illustrates two examples of how data can flow from source to AI/ML algorithm. Figure 3a is a classic example where many research groups contribute similar data types on materials to published literature, which can be captured in public or private databases. In this example, the project team must determine how the different sources of data were pre-processed and design a scheme to homogenize it for AI/ML. Figure 3b represents a specific case where data generation and AI/ML are intended to happen concurrently (or on similar time scales). In this case it is extremely important to define two separate databases—the materials science *repository* where raw data can be efficiently dumped and the data science *library* where curated data sets can be queried seamlessly. The data management infrastructure underlies these databases and facilitates the task of meeting the materials and data scientists in the middle, reducing the intrusion of Brooks' Law. *Developing standard data architectures will ensure that materials scientists capture the processed data and associated metadata in a format which can be easily queried and utilized by ML algorithms in a manner which does not burden data scientists with wondering if they have violated underlying physical or chemical laws.*

Science and Technology Innovation Centers

Training a sustainable cohort of AI/ML ready materials scientists and engineers loses significant value without the presence of facilities designed to optimally translate their skills into discovery. Science and engineering laboratories have become much more compartmentalized in the last 100 years, in part due to a desire to increase control over the local environment as the pursuit of more sensitive equipment has led decision making [53, 54]. However, the social challenges previously addressed mandate consideration of laboratories that are communally minded.

There is room for reasonable debate on what the optimal design of a communal science and engineering laboratory might be. However, early exemplars of self-driving laboratories give insights into a few of the key factors to consider [55–59]. For one, there are decisions about the balance of human or expert-driven versus machine-driven decisions [57, 60]. A simple example is that equipment experiences drift over time and must be calibrated, and while there has been some success in automating this task, successes have been generally limited to a single piece of equipment in the loop [61]. Of larger consequence, perfectly autonomous decision making implies the encoding of extensive expert knowledge into the algorithm; however, this knowledge is

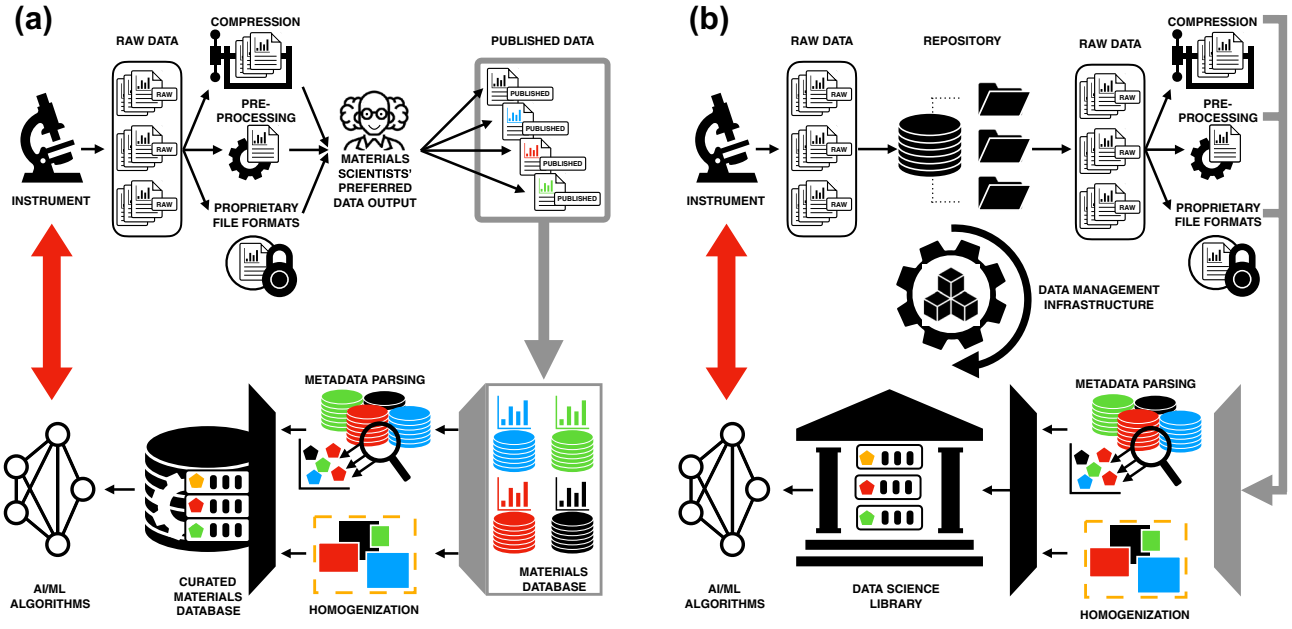

Fig. 3 **a** Traditional, inefficient materials science data management paradigm for building data sets sufficient for machine learning. **b** Proposed data management infrastructure which allows material science data to be collected with AI/ML in mind, enabling paradigm-shifting decreases in the time required to build a sufficient data set

built on an expectation of what has already been discovered. How an AI agent will recognize and differentiate useful discovery from unhelpful discovery without expert input is still an open question supporting the near-term need for spaces that support hybrid (human/machine) science and engineering [62, 63].

Modularity is another impactful factor. Imagine if our 100 prisoners had the capability to reconfigure the drawers such that the numbers on the slips of paper were matched to that of the drawer. Without access to infinite resources, AI/ML-enabled autonomous laboratories must be able to efficiently reconfigure virtual and physical layouts to support re-prioritization of materials discovery. Modularity in all aspects of science and technology innovation centers appears likely to be a blossoming topic in the coming years [64, 65]. *One can envision that next-generation materials innovation centers will house not only autonomous workflows consisting of many modular blocks, but also flexible workspaces supporting the development of scientific and engineering solutions to wholly new modular blocks.*

Applications: Past, Present, and Future

We now provide 3 examples—past, present, and future—for which properly addressing these social and engineering challenges could have, is, or will enable (-ed,-ing,-e) a more rapid exploration/exploitation of the application space.

Past: Metal Organic Frameworks

Beginning in earnest during the mid-1990s, growth in the field of metal–organic frameworks led to multiple review articles summarizing synthetic approaches, structure–property relationships, and applications by the mid-2000s [66–69]. Despite the breadth of literature at that time, Yaghi et al. highlighted a dilemma analogous to that facing the 100 prisoners, "of the almost unlimited possible networks, which can be expected to form and how can they be synthesized?" [66]. While expert and intuition-driven trial and error-based exploration of this class of materials has led to their implementation in real world applications, autonomous AI/ML-enabled research platforms likely would have accelerated the process had the ability to implement such an approach existed at the time.

Present—Sandia National Labs' Beyond Fingerprinting Grand Challenge

Currently, the Beyond Fingerprinting Grand Challenge at Sandia National Laboratories is seeking to discover new materials and manufacturing process through an AI-guided approach that integrates human-subject-matter expertise with physics-based constraints to unearth process–structure–property correlations. Algorithms are being trained on high-throughput experiments to efficiently detect key "fingerprints" in materials data, prognose materials performance, and guide effective adaptations [29]. Efforts thus far have demonstrated the application of genetic algorithms for learning time-dependent deposition protocols in thin film design, neural operators and autoencoder architectures for learning two-phase microstructure evolution, and self-supervised learning for inferring topological transitions in pattern-forming process [20, 21, 70].

Future: Microelectronics Hubs

With the passage of the CHIPS and Science Act, the United States government has authorized $280 billion to be appropriated over five years in new funding for domestic research and manufacturing of semiconductors [71, 72]. While congressional spending panels have yet to appropriate much of the money, the NSF has received a $200 million appropriation to boost workforce training programs in microelectronics. This is an important first step, but if the United States desires technological preeminence in this space, adequate funding needs to be appropriated for the development of autonomous AI/ML-enabled research platforms that address the highlighted social and engineering challenges. While the United States has a very large population of prisoners (er, scientists and engineers), these numbers are dwarfed by its geopolitical rivals throughout the world. Addressing these challenges quickly will help ensure that the United States maintains worldwide leadership in the co-design of microelectronics and will increase the utility of researcher's discoveries [73].

Conclusions

In many ways, the recent re-popularization of the 100 Prisoners Problem exemplifies the tenacity in which modern materials scientists are searching for elegant and efficient solutions to the immense challenge of discovery. Past us are the days in which purely combinatorial methods bear fruit for enormous, urgent problems such as sustainability, alternative energy, or next generation microelectronics. Here are the days of materials science integration with Artificial Intelligence and Machine Learning. While computational power, algorithms and technology will continue to advance at the cutting-edge, it is imperative that the field of materials science stays committed to solving the social (*communication, conflict, and communities of practice*) and engineering (*stewardship, standard data architectures, and science and technology innovation centers*) issues with as much enthusiasm as it gives discovery.

Acknowledgements This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. B.L. Boyce is acknowledged for contributions on black swan events and for funding associated with the completion of this manuscript. This work was performed, in part, at the Center for Integrated Nanotechnologies, an Office of Science User Facility operated for the U.S. Department of Energy (DOE) Office of Science. This article has been authored by employees of National Technology & Engineering Solutions of Sandia, LLC under Contract No. DE-NA0003525 with the U.S. Department of Energy (DOE). The employees own all right, title and interest in and to the article and are solely responsible for its contents. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this article or allow others to do so, for United States Government purposes. The DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan https://www.energy.gov/downloads/doe-public-access-plan.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Keoleian GA, Sullivan JL (2012) Materials challenges and opportunities for enhancing the sustainability of automobiles. MRS Bull 37(4):365–373
2. Armand M, Endres F, MacFarlane DR, Ohno H, Scrosati B (2011) Ionic-liquid materials for the electrochemical challenges of the future. In: Dusastre V (ed) Materials for sustainable energy a collection of peer-reviewed research and review articles from Nature Publishing Group. Co-Published with Macmillan Publishers Ltd., pp 129–137
3. Chu S, Majumdar A (2012) Opportunities and challenges for a sustainable energy future. Nature 488(7411):294–303
4. Mohanty AK, Vivekanandhan S, Pin J-M, Misra M (2018) Composites from renewable and sustainable resources: challenges and innovations. Science 362(6414):536–542
5. Chen K-N, Tu K-N (2015) Materials challenges in three-dimensional integrated circuits. MRS Bull 40(3):219–222
6. Tabor DP, Roch LM, Saikin SK, Kreisbeck C, Sheberla D, Montoya JH et al (2018) Accelerating the discovery of materials for clean energy in the era of smart automation. Nat Rev Mater 3(5):5–20
7. Aspuru-Guzik A, Persson K (2018) materials acceleration platform: accelerating advanced energy materials discovery by integrating high-throughput methods and artificial intelligence. Mission Innov
8. Amis EJ, Xiang X-D, Zhao J-C (2002) Combinatorial materials science: What's new since Edison? MRS Bull 27(4):295–300
9. Rajan K (2008) Combinatorial materials sciences: experimental strategies for accelerated knowledge discovery. Annu Rev Mater Res 38:299–322
10. Veratasium (2022) The riddle that seems impossible even if you know the answer. In: Muller D (Eds) Youtube2022. https://youtu.be/iSNsgj1OCLA
11. Flajolet P, Sedgewick R (2009) Analytic combinatorics. Cambridge University Press
12. Boyce B (2022) Microstructural black swans. In: IOP conference series: materials science and engineering: IOP Publishing. p 012004
13. Taleb NN (2007) The black swan: the impact of the highly improbable. Random House
14. Samudrala S, Rajan K, Ganapathysubramanian B (2013) Data dimensionality reduction in materials science. Informat Mater Sci Eng. https://doi.org/10.1016/B978-0-12-394399-6.00006-0
15. Wagner N, Rondinelli JM (2016) Theory-guided machine learning in materials science. Front Mater 3:28
16. Cuomo S, Di Cola VS, Giampaolo F, Rozza G, Raissi M, Piccialli F (2022) Scientific machine learning through physics–informed neural networks: where we are and what's next. J Sci Comput 92(3):88
17. Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L (2021) Physics-informed machine learning. Nature Rev Phys 3(6):422–440
18. Jain M, Deleu T, Hartford J, Liu C-H, Hernandez-Garcia A, Bengio Y (2023) GFlowNets for AI-driven scientific discovery. arXiv preprint https://arxiv.org/abs/2302.00615
19. Vabalas A, Gowen E, Poliakoff E, Casson AJ (2019) Machine learning algorithm validation with a limited sample size. PLoS ONE 14(11):e0224365. https://doi.org/10.1371/journal.pone.0224365
20. Abram M, Burghardt K, Ver Steeg G, Galstyan A, Dingreville R (2022) Inferring topological transitions in pattern-forming processes with self-supervised learning. npj Comput Mater 8(1):205. https://doi.org/10.1038/s41524-022-00889-2
21. Desai S, Dingreville R (2022) Learning time-dependent deposition protocols to design thin films via genetic algorithms. Mater Des 219:110815. https://doi.org/10.1016/j.matdes.2022.110815
22. Liang Q, Gongora AE, Ren Z, Tiihonen A, Liu Z, Sun S et al (2021) Benchmarking the performance of Bayesian optimization across multiple experimental materials science domains. npj Comput Mater 7(1):188
23. Wang A, Liang H, McDannald A, Takeuchi I, Kusne AG (2022) Benchmarking active learning strategies for materials optimization and discovery. Oxford Open Mater Sci. 2(1):006
24. Khatamsaz D, Vela B, Singh P, Johnson DD, Allaire D, Arróyave R (2023) Bayesian optimization with active learning of design constraints using an entropy-based approach. npj Comput Mater 9(1):49. https://doi.org/10.1038/s41524-023-01006-7
25. Rajak P, Krishnamoorthy A, Mishra A, Kalia R, Nakano A, Vashishta P (2021) Autonomous reinforcement learning agent for chemical vapor deposition synthesis of quantum materials. npj Comput Mater 7(1):108. https://doi.org/10.1038/s41524-021-00535-3
26. Hattrick-Simpers JR, Gregoire JM, Kusne AG (2016) Perspective: composition–structure–property mapping in high-throughput experiments: turning data into knowledge. APL Mater. https://doi.org/10.1063/1.4950995
27. Ludwig A (2019) Discovery of new materials using combinatorial synthesis and high-throughput characterization of thin-film materials libraries combined with computational methods. npj Comput Mater 5(1):70
28. Huang K, Kain C, Diaz-Vallejo N, Sohn Y, Zhou L (2021) High throughput mechanical testing platform and application in metal additive manufacturing and process optimization. J Manuf Process 66:494–505
29. Trask N, Martinez C, Lee K, Boyce B (2022) Unsupervised physics-informed disentanglement of multimodal data for high-throughput scientific discovery. arXiv preprint https://arxiv.org/abs/2202.03242
30. Wang T, Shao M, Guo R, Tao F, Zhang G, Snoussi H et al (2021) Surrogate model via artificial intelligence method for accelerating

screening materials and performance prediction. Adv Func Mater 31(8):2006245
31. Harrison GW, List JA, Towe C (2007) Naturally occurring preferences and exogenous laboratory experiments: a case study of risk aversion. Econometrica 75(2):433–458
32. Strevens M (2003) The role of the priority rule in science. J Philos 100(2):55–79
33. Narayanamurti V, Tsao JY (2021) The genesis of technoscientific revolutions: rethinking the nature and nurture of research. Harvard University Press
34. McCain KW, Salvucci LJ (2006) How influential is Brooks' law? A longitudinal citation context analysis of Frederick Brooks' the mythical man-month. J Inf Sci 32(3):277–295
35. Opelt K (2008) Overcoming Brooks' Law. In: Agile 2008 Conference: IEEE; p 208–11
36. Blaiszik B, Chard K, Pruyne J, Ananthakrishnan R, Tuecke S, Foster I (2016) The materials data facility: data services to advance materials science research. JOM 68(8):2045–2052
37. Blaiszik B, Ward L, Schwarting M, Gaff J, Chard R, Pike D et al (2019) A data ecosystem to support machine learning in materials science. MRS Commun 9(4):1125–1133
38. Stein S (2012) Mass spectral reference libraries: an ever-expanding resource for chemical identification. ACS Publications
39. Messner M (2016) Modular consitutive modeling library for structural materials https://github.com/Argonne-National-Laboratory/neml. Accessed 2022
40. de Pablo JJ, Jackson NE, Webb MA, Chen L-Q, Moore JE, Morgan D et al (2019) New frontiers for the materials genome initiative. npj Comput Mater 5(1):1–23
41. Brainard J, Kaiser J (2022) US to require free access to papers on all research it funds. Science 377(6610):1026–1027
42. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C (2018) A survey on deep transfer learning. International conference on artificial neural networks: Springer; p 270-9
43. Weiss K, Khoshgoftaar TM, Wang D (2016) A survey of transfer learning. J Big data 3(1):1–40
44. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H et al (2020) A comprehensive survey on transfer learning. Proc IEEE 109(1):43–76
45. Aggarwal A, Mittal M, Battineni G (2021) Generative adversarial network: an overview of theory and applications. Int J Inf Manag Data Insights 1(1):100004
46. Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA (2018) Generative adversarial networks: an overview. IEEE Signal Process Mag 35(1):53–65
47. Rooney MB, MacLeod BP, Oldford R, Thompson ZJ, White KL, Tungjunyatham J et al (2022) A self-driving laboratory designed to accelerate the discovery of adhesive materials. Digit Discov. https://doi.org/10.1039/D2DD00029F
48. MacLeod BP, Parlane FG, Morrissey TD, Häse F, Roch LM, Dettelbach KE et al (2020) Self-driving laboratory for accelerated discovery of thin-film materials. Sci Adv. https://doi.org/10.1126/sciadv.aaz8867
49. Freeman JB (2020) Measuring and resolving LGBTQ disparities in STEM. Policy Insights Behav Brain Sci 7(2):141–148
50. Saw G, Chang C-N, Chan H-Y (2018) Cross-sectional and longitudinal disparities in STEM career aspirations at the intersection of gender, race/ethnicity, and socioeconomic status. Educ Res 47(8):525–531
51. Simmons JP, Drummy LF, Bouman CA, De Graef M (2019) Statistical methods for materials science: the data science of microstructure characterization. CRC Press
52. Wendelberger JG (2018) Extracting the data from the LCM vk4 formatted output file. Los Alamos National Lab. (LANL), Los Alamos, NM (United States)
53. Schlich T (2007) Surgery, science and modernity: Operating rooms and laboratories as spaces of control. Hist Sci 45(3):231–256
54. Musau F, Steemers K (2007) Space planning and energy efficiency in laboratory buildings: the role of spatial, activity and temporal diversity. Archit Sci Rev 50(3):281–292
55. Dyck O, Jesse S, Kalinin SV (2019) A self-driving microscope and the atomic forge. MRS Bull 44(9):669–670
56. Häse F, Roch LM, Aspuru-Guzik A (2019) Next-generation experimentation with self-driving laboratories. Trends Chem 1(3):282–291
57. Soldatov MA, Butova VV, Pashkov D, Butakova MA, Medvedev PV, Chernov AV et al (2021) Self-driving laboratories for development of new functional materials and optimizing known reactions. Nanomaterials 11(3):619
58. Butakova MA, Chernov AV, Kartashov OO, Soldatov AV (2021) Data-centric architecture for self-driving laboratories with autonomous discovery of new nanomaterials. Nanomaterials 12(1):12
59. MacLeod BP (2022) A self-driving laboratory for optimizing thin-film materials. University of British Columbia
60. Seifrid M, Pollice R, Aguilar-Granda A, Morgan Chan Z, Hotta K, Ser CT et al (2022) Autonomous chemical experiments: challenges and perspectives on establishing a self-driving lab. Acc Chem Res 55(17):2454–2466
61. Roch LM, Häse F, Kreisbeck C, Tamayo-Mendoza T, Yunker LP, Hein JE et al (2020) ChemOS: an orchestration software to democratize autonomous discovery. PLoS ONE 15(4):e0229862
62. Venkatasubramanian V (2019) The promise of artificial intelligence in chemical engineering: Is it here, finally? AIChE J 65(2):466–478
63. Tsao J, Ting C, Johnson C (2019) Creative outcome as implausible utility. Rev Gen Psychol 23(3):279–292
64. Rahmanian F, Flowers J, Guevarra D, Richter M, Fichtner M, Donnely P et al (2022) Enabling modular autonomous feedback-loops in materials science through hierarchical experimental laboratory automation and orchestration. Adv Mater Interfaces 9(8):2101987
65. Dennis LA, Fisher M, Aitken JM, Veres SM, Gao Y, Shaukat A et al (2014) Reconfigurable autonomy. KI-Künstl Intell 28(3):199–207
66. Yaghi OM, O'Keeffe M, Ockwig NW, Chae HK, Eddaoudi M, Kim J (2003) Reticular synthesis and the design of new materials. Nature 423(6941):705–714. https://doi.org/10.1038/nature01650
67. Rowsell JLC, Yaghi OM (2004) Metal–organic frameworks: a new class of porous materials. Microporous Mesoporous Mater 73(1):3–14. https://doi.org/10.1016/j.micromeso.2004.03.034
68. Kitagawa S, Kitaura R, Noro SI (2004) Functional porous coordination polymers. Angew Chem Int Ed 43(18):2334–2375
69. James SL (2003) Metal-organic frameworks. Chem Soc Rev 32(5):276–288. https://doi.org/10.1039/B200393G
70. Oommen V, Shukla K, Goswami S, Dingreville R, Karniadakis GE (2022) Learning two-phase microstructure evolution using neural operators and autoencoder architectures. npj Comput Mater 8(1):190. https://doi.org/10.1038/s41524-022-00876-7
71. Mervis J (2022) Innovation bill will reshape science agencies. Science (New York, NY) 377(6606):562–563
72. Mervis J (2022) New tech law offers billions for research. Science 377(6611):1133–1134
73. Kanaan M (2000) T-minus AI: humanity's countdown to artificial intelligence and the new pursuit of global power. Benbella Books