Joule

CellPress

Perspective

Accelerating Materials Development via Automation, Machine Learning, and High-Performance Computing

Juan-Pablo Correa-Baena,[1] Kedar Hippalgaonkar,[2] Jeroen van Duren,[3] Shaffiq Jaffer,[4] Vijay R. Chandrasekhar,[5] Vladan Stevanovic,[6] Cyrus Wadia,[7] Supratik Guha,[8,9] and Tonio Buonassisi[1,*]

Successful materials innovations can transform society. However, materials research often involves long timelines and low success probabilities, dissuading investors who have expectations of shorter times from bench to business. A combination of emergent technologies could accelerate the pace of novel materials development by ten times or more, aligning the timelines of stakeholders (investors and researchers), markets, and the environment, while increasing return on investment. First, tool automation enables rapid experimental testing of candidate materials. Second, high-performance computing concentrates experimental bandwidth on promising compounds by predicting and inferring bulk, interface, and defect-related properties. Third, machine learning connects the former two, where experimental outputs automatically refine theory and help define next experiments. We describe state-of-the-art attempts to realize this vision and identify resource gaps. We posit that over the coming decade, this combination of tools will transform the way we perform materials research, with considerable first-mover advantages at stake.

Context & Scale

The convergence of high-performance computing, automation, and machine learning promises to accelerate the rate of materials discovery by ≥10 times, better aligning investor and stakeholder timelines. Infrastructure and human-capital investments are discussed, including equipment capabilities, data management, education, and incentives. As our field transitions from thinking "data poor" to thinking "data rich," we envision a scientific laboratory where the process of materials discovery continues without disruptions, aided by computational power augmenting the human mind, and freeing the latter to perform research closer to the speed of imagination, addressing societal challenges in market-relevant timeframes.

The development of novel materials has long been stymied by a mismatch of time constants (Figure 1). Materials development typically occurs over a 15- to 25-year time horizon, sometimes requiring synthesis and characterization of millions of samples. However, corporate and government funders desire tangible results within the residency time of their leadership, typically 2–5 years. The residency time for postdocs and students in a research laboratory is usually 2–5 years; when a project outlasts the residency of a single individual, seamless continuity of motivation and intellectual property is often the exception, not the rule. Market drivers of novel materials development, informed by business competition and environmental considerations, often demand solutions within a shorter time horizon. This mismatch in time constants results in a historically poor return on investment of energy-materials (cleantech) research relative to comparable investments in medical or software development.[1]

To bridge this mismatch in time horizons and increase the success rate of materials research, both public- and private-sector actors endeavor to develop new paradigms for materials development.[2] The U.S. Materials Genome Initiative focused on three "missing links": computational tools to focus experimental efforts in the most promising directions, data repositories to aggregate learnings and identify trends, and higher-throughput experimental tools.[3] This call to action was mirrored in industry and by university- and laboratory-led consortia, many focused on simulation-based inverse design and discovery and properties databases. As

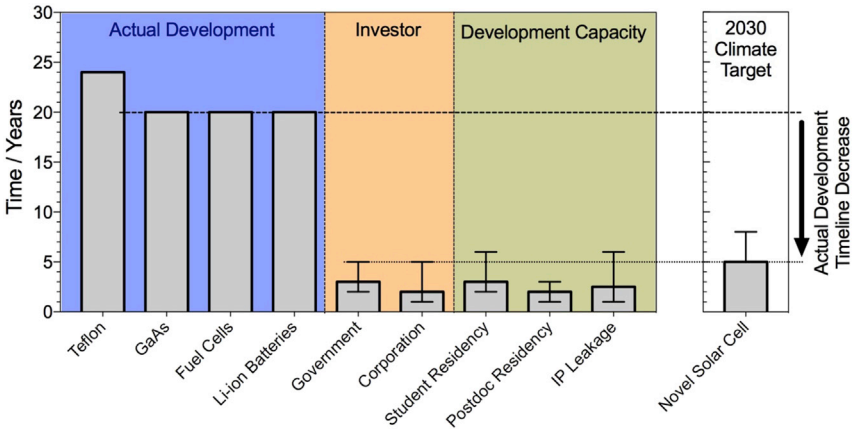

Figure 1. Timelines for Materials Discovery and Development

Timelines of examples of certain technologies (blue area), typical academic funding grants (orange), development capacity (green), and deployment of sustainable energy (i.e., via solar cells) to fulfill the 2030 climate targets. Error bars indicate typical ranges.

these tools matured, the throughput of materials prediction often vastly outstripped experimentalists' ability to screen for materials with low rates of false negatives.

Today, a new paradigm is emerging for experimental materials research, which promises to enable more rapid discovery of novel materials.[4,5] Figure 2 illustrates one such prototypical vision, entitled "accelerated materials development and manufacturing." Rapid, automated feedback loops are guided by machine learning, and an emphasis on value creation through end-product and industry transfer. There is a unique opportunity today to develop these capabilities in testbed fashion, with considerable improvements in research productivity and first-mover advantages at stake.

As is often the case with convergent technologies, one observes significant advances in individual "silos" before the leveraged ensemble effect bears its full impact. A historical example is three-dimensional (3D) printing, wherein 3D computer-aided design, computer-to-hardware interface protocols, and ink-jet printing technologies evolved individually, before being combined by Prof. Ely Sachs and his MIT team into the first 3D printer. The ability to observe emergent technologies within individual silos, and assemble them into an ensemble that is greater than the sum of its parts, mirrors the challenge in novel materials development today. The following paragraphs describe the discrete, emergent innovations in "siloed" domains that are presently converging, and promise to enable this paradigm shift within the next decade.

Theory

Today, the rate of theoretical prediction vastly outstrips the rate of experimental synthesis, characterization, and validation.[7] This emergence is enabled by three trends: faster computation, more efficient and accurate theoretical approaches and simulation tools, and the ability to screen large databases quickly, such as MaterialsProject.org. To better focus limited experimental bandwidth, there is increasing interest to simulate the "how" of synthesis, not just the "what"—capturing in computer models the full complexity of environmental factors (e.g., humidity), reaction energy barriers, and kinetic limitations (so-called "non-equilibrium" synthesis).[8] In parallel,

[1]Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[2]Institute of Materials Research and Engineering (IMRE), A*STAR (Agency for Science, Technology and Research), Innovis, Singapore

[3]Intermolecular Inc., San Jose, CA 95134, USA

[4]TOTAL American Services, Inc., 82 South Street, Hopkington, MA 01748, USA

[5]Institute for Infocomm Research (I^2R), A*STAR (Agency for Science, Technology and Research), #21-01 Connexis (South Tower), Singapore

[6]Colorado School of Mines, Golden, CO 80401, USA

[7]Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[8]Center for Nanoscale Materials, Argonne National Laboratory, Argonne, IL 60439, USA

[9]University of Chicago, Chicago, IL 60615, USA

*Correspondence: buonassisi@mac.com

https://doi.org/10.1016/j.joule.2018.05.009

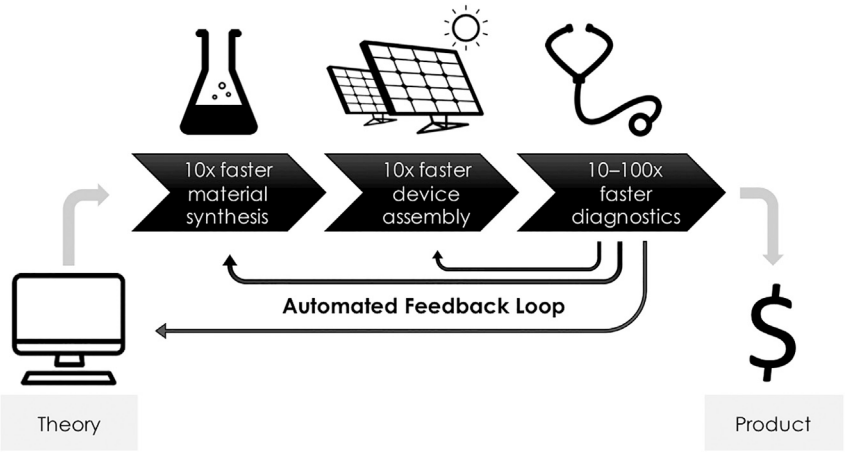

Figure 2. Schematic of the Accelerated Materials Discovery Process
The automated feedback loop, driven by machine learning, drives process improvement. The theory, synthesis, and device processes take advantage of high-performance computing and materials databases. For many materials systems today, an ~10 times multiplier is a minimum necessary to bridge the mismatch in actual and aspirational materials development timelines shown in Figure 1. Note that the cycle of learning is limited by the slowest step (bottleneck). Icons from Freepik.[6]

theorists seek to rationally design materials with combinations of properties; first, by predicting combinations of properties (e.g., chemical, microstructural, interface, surface) in one simulation framework and/or database, then connecting material predictions with device performance and reliability predictions, then extending this framework to both known and not-yet-discovered compounds, and ultimately, solving the inverse problem.[9–11]

High-Throughput Materials, Device, and Systems Synthesis

Historically, slow vacuum-based deposition methods inhibit materials development. Modern vacuum-based tools, including combinatorial approaches and large-scale, fast serial deposition/reactions, enable meaningful rate increases for materials and device synthesis.[12,13] Variants of existing deposition methods (e.g., close-space sublimation) offer higher growth rates, point-defect control, and precise stoichiometry and impurity control for process-compatible materials. Solution synthesis has gained acceptance with the emergence of higher-quality precursors and materials, including CdS quantum dots, polymer solar cells, and lead-halide perovskites.[7,14] The growing diversity of precursors (from molecular to nanoparticle), synthesis control (including solvent engineering), and thin-film synthesis methods (lab-based spin coating to industrially compatible large-area printing) makes this a powerful and flexible platform to deposit a range of new materials. Emergence of 3D printed materials provides another ubiquitous alternative. At laboratory scale, throughputs for such rapid synthesis routes[7,15] can be up to an order of magnitude greater than vacuum-based techniques, and remain to be explored for multinary materials with novel microstructures. With declining component costs and greater adoption of standards, the ability to rapidly combine discrete devices into components and systems in a modular and flexible manner is emerging.

Defect Tolerance and Engineering

Often, theoretical predictions are made for "ideal" materials systems. However, real samples contain defects (e.g., impurities, structural defects) that can harm (or,

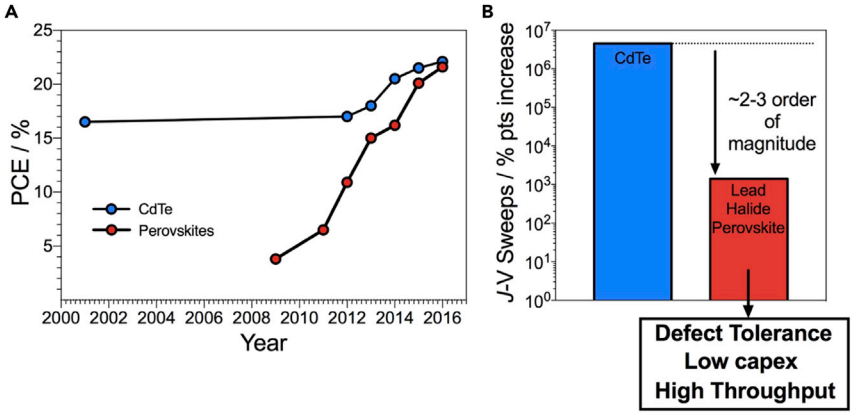

Figure 3. A Case Study of Fast Materials Development Based on Photovoltaic Applications

(A) Certified power conversion efficiency (PCE) over time for CdTe and perovskite solar cells.

(B) Number of *J-V* sweeps measured divided by the increase in percentage point achieved during the device development of CdTe and perovskite solar cells. Three orders of magnitude fewer *J-V* sweeps per percentage efficiency improvement were needed to advance perovskite efficiencies relative to traditional thin-film solar cell materials. We hypothesize that this difference is partially due to greater "defect tolerance" of perovskites, enabling a faster and more economical materials development process. Data from Correa-Baena et al.[18]

occasionally, benefit) bulk and interface properties. To mitigate the risk of defect-induced false negatives during high-throughput materials screening, it is desirable to identify classes of materials less adversely affected by defects (so-called "defect tolerant"[16,17]), and rapidly diagnose and decouple the effects of defects on material performance. A notable recent example is the serendipitous discovery of lead-halide perovskites for optoelectronic applications.[14,15] In addition to being amenable to high-throughput solution-phase deposition, lead-halide perovskites also required orders of magnitude less research effort to achieve similar performance improvements to traditional inorganic thin-film materials (Figure 3). It is suspected that part of the facility to improve performance is owed to increased defect tolerance of lead-halide perovskites, resulting in improved bulk-transport properties. Determining the underlying physics of and developing design rules for defect tolerance may inform screening criteria for new materials, especially with new computational tools such as General Adversarial Networks that are state-of-the-art in anomaly detection.[19,20] The next step lies in focusing experimental effort on candidates capable of rapid performance improvements during early screening and development, and wider process tolerance in manufacturing. In relation to the beneficial aspects of defects and impurities, recent theory advancements,[21] in combination with computational tools to rapidly assess and predict solubility and electrical properties of defects,[22] allows high-throughput screening of materials for applications where the desired functionality is enabled by the defects and/or dopants (e.g., thermoelectrics, transparent electronics).

High-Throughput Diagnosis

Characterization tools have also benefited from high-performance computing, automation, and machine learning. For instance, one high-resolution X-ray photoelectron spectroscopy spectrum could take an entire day with technology from the 1970s, while the same measurement today requires less than an hour. Today, advanced statistics and machine learning promise to further accelerate the rate of learning.[23,24] Tools now exist that can acquire multiple XPS spectra on a single sample (e.g., with composition gradients), and automated spectral analysis of large

datasets is now possible, enabling estimation of unknown materials in a compositional map. Others seek to replace spectroscopy with rapid non-destructive testing; several bulk and interface properties can be simultaneously diagnosed by using Bayesian inference in combination with non-destructive device testing, enabling $\geq$10 times faster (and in certain cases, more precise) diagnosis than traditional characterization tools.[25] This kind of parameter estimation can be applied to finished components, devices, and systems, and has the potential to not only enable faster troubleshooting, but also to accurately estimate intrinsic material properties[26,27] as well as ultimate performance potential, thus informing the decision to pursue or abandon further investment in a given candidate material even at early stages of materials screening.

Machine Learning

Machine learning comprises a broad class of approaches, which may play several different roles in the future materials development cycle. First, a common application of machine learning is for materials selection, in which historical experimental observations are used to inform predictions of future properties (attributes) of unknown compounds, or discover new ones.[28] Such an approach has been realized to help discover novel active layers in organic solar cells[29] and light-emitting diodes,[30] and metal alloys,[31,32] among many others.[33] Second, machine-learning tools can help extract greater and more accurate information from diagnosis, as detailed in the previous section. Third, machine-learning tools may help close the automation loop between diagnosis and synthesis, shown in Figure 2, by reducing the degree of human intervention and reliance on heuristics. For example, when relationships between experimental inputs and diagnosis outputs can be inferred by neural networks, detailed process and device models may no longer be needed to predict outcomes and optimize processes. All three applications of machine learning to the materials development cycle benefit from the availability of more data, to train and sharpen the predictive capacity of such tools.

Achieving predictability without losing physical insights is an emergent challenge and research opportunity. Such methods may also increase learning from diagnosis, by consolidating research output in singular databases, drawing automated inferences from the data, and in the future perhaps aggregating the experience and knowledge base via natural language processing of existing research papers and materials property databases.[34]

Envisioning the "Hardware Cloud"

Materials synthesis equipment today is becoming increasingly remotely operable—enabling research and operation by an investigator who is not in proximal presence to the deposition equipment. This opens up two related opportunities with far-reaching consequences. Large, expensive, synthesis equipment can be grouped together with massively parallel characterization equipment to form synthesis centers of the future, which are operated by remote users and researchers and managed by an on-site professional staff. Akin in concept to the Software Cloud concept, where one's computing and data are stored across machines worldwide in a seamless manner, a Hardware Cloud would enable a user to deposit, measure, and carry out research (with real-time feedback through *in situ* characterization tools) across a number of networked materials-processing systems distributed nationally or internationally in a seamless manner. This also leads to the second opportunity: to be able to store, curate, access, process, and diagnose all data gathered in these networked experiments in Public or Private Clouds.[3,35–37] (protocols and formats for such science data collectives are discussed in the following paragraphs.) This

will greatly facilitate two emerging issues: (1) increasing the efficient availability of data across a wide number of experiments and experimental platforms for post-analysis; and (2) making available for analysis data that indicate "what did not work"; this is not easily available but is instrumental in the learning process, and has its own value in increasing the collective efficiency of research progress.

Infrastructure Investments toward Accelerated Materials Development and Manufacturing

Realizing the vision shown in Figure 2 requires a sustained commitment over several years to develop software, hardware, and human resources, and to connect these new capabilities in testbed fashion.

Investments in Applied Machine Learning

Supported by ample investments into machine-learning methods development, a pressing challenge is how to down-select and apply the most appropriate machine-learning methods to enable the "automated feedback loop" shown in Figure 2. Compared with other widely recognized applications of machine learning today (e.g., vision recognition, natural language processing, and board gaming), materials research often involves sparse datasets (e.g., small sample sizes and number of experimental inputs and outputs, for training and fitting) and less well-constrained "rules" (e.g., complex physics and chemistry, non-binary inputs and outputs, large experimental errors, uncontrolled input variables, and incomplete characterization of outputs, to name a few). These realities make the typical materials science problem (e.g., layer-by-layer atomic assembly of a thin film) decidedly more complex and less well defined than a match of "Go," where the rules and playing board are constrained. Deep machine learning (DML) appears well poised to address this complexity.[38] Computation speed can be improved by developing "pre-trained" neural networks that incorporate the underlying physics and chemistry common to materials synthesis, performance, and defects, bringing DML within reach of commonly available hardware and software.

A balance must be found between achieving actionable results and inferring physical insight from "black-box" computational methods, to advance both engineering and scientific objectives, and minimize unintended consequences. There is a need to apply "white-box" (i.e., opposite of black-box) machine-learning methods to materials science problems. One possible approach may be application of semi-supervised deep learning algorithms, which learn with lots of unlabeled data and very little labeled data.[39]

Lastly, the ability of machine-learning tools to adapt to uncontrolled and changing experimental conditions is essential. Promising developments include online deep learning, which builds neural networks on the fly, gradually adding neurons (e.g., as baseline experimental conditions change, or as new physics becomes dominant).[40]

Investment in Standards Governing Data Formatting and Storage

Investment in standards governing data formatting and storage would facilitate data entry into machine-learning software. Standards embed contextual know-how, hierarchy, and rational thought. Some communities have implemented standards governing raw and processed data, e.g., crystallography, genetics, and geography. However, in most materials research communities, there are no universally accepted and implemented data standards. Several materials databases have been created, often specialized by material class or application, and with varying protocols for

updating information and enforcing hygiene. Furthermore, these databases often lack ability to quickly and accurately predict device-relevant combinations of properties (e.g., chemical, mechanical, optoelectronic, microstructural, surface, interface). Several data standards have been proposed[41–43]; widespread adoption may hinge on pervasive adoption of data-management systems described in the next paragraph, with FAIR guiding principles in mind.[44] In the absence of data standards, it is possible that the burden of data aggregation will shift onto natural language processors,[34] i.e., computer programs designed to extract relevant data from available media (e.g., publications, reports, presentations, and theses).

Investment in Data-Management Tools

Investment in data-management tools (e.g., informatics systems) is needed to manage data obtained from lab equipment and store records, coordinate tasks, and enforce protocols. On one hand, such systems have been shown to be of high value for well-defined research problems and tool sets. For early-stage materials research, data-management tools require a deft balance between flexibility and standardization, and the ability to accommodate non-standard workflows, multiple participants, and equipment spread across multiple sites, including shared-use facilities, in an elegant and seamless manner. When implemented well, data-management systems can increase the quality, uniformity, and accessibility of data serving as inputs into machine-learning tools; when implemented too inflexibly, data-management systems can cause frictions to researcher workflow and stimulate their resistance. It is possible that, as suggested by Rafael Jaramillo (MIT), metadata-based distributed data-management systems may warrant strong consideration for early-stage materials research; a challenge will be, how to capture metadata in an automated, accurate, thorough, and comprehensive manner.

Investments in Infrastructure

Investments in infrastructure are needed to increase throughput of synthesis, device fabrication, and diagnosis tools. The potential of automation must be realized, without sacrificing material quality and offsetting the advantages of higher throughput with an increase in false negatives. The emergence of multi-parameter estimation methodologies, including Bayesian inference and design of experiments (DoE) algorithms, invites the invention of new non-destructive diagnostic apparatus designed to take full advantage of these new methodologies.

There are significant challenges associated with producing and analyzing large quantities of data. New tools being developed by machine-learning specialists invite the possibility of modifying hardware design to take advantage of machine-learning tools, rather than the other way around.[45]

Revised policies at institution, funding agency, and government levels may accelerate or stymie the required ongoing investments at levels large and small, and invites considering how export control laws, import duties, grant purchasing restrictions, overhead rates, auditing, and claw-back clauses affect required equipment investments to enable this transformation.

Human-Capital Investments toward Accelerated Materials Development and Manufacturing

Investments in human capital are required to prepare researchers to leverage these new tools. The transition from being "data poor" to being "data rich" invites changes in how we think, how we incentivize, and how we teach.

How We Think

In a "data-poor" world, the time and cost of conducting each experiment is relatively large, and a risk-adverse mindset is advantageous. In a "data-rich" world, a larger number of unique experiments can be conducted per unit time, meaning that failure of any given experiment will have lesser negative impact on a researcher's milestones and publication record. This will enable researchers to experiment with greater creativity and risk-taking. This has three important implications for "how we think": First, a greater premium will be placed on experimental concept and design, as researchers who design experiments amenable to new tools will be rewarded. Second, a decreasing cost-per-experiment may result in reduced barriers for junior researchers to establish themselves, decreasing the premium of initial investment, prompting new as well as established researchers to explore new fields.

Third, an accelerated materials development framework invites a system-level perspective[46] that mirrors the new tools. Greater experimental throughput suggests that devices and systems may increasingly be analyzed holistically in lieu of isolated sub-components, test structures, and proxies. A "data-rich" world will allow us to analyze complex systems more directly, with lesser need to break into sub-components or impose *a priori* simplifications even without complete visibility into each sub-component. Wielding these new computer-based tools to greatest effect requires that researchers learn to "think" like machine-learning algorithms, appreciating the nuances and trade-offs of different approaches, requiring a mindset change providing an opportunity to identify weak links faster, focusing effort on those parameters with highest returns on investment.

Incentives

Encouraging the mindset change and transitions mentioned in the previous section will be complemented with a "constant of friction" governed in part by professional incentives of decades-old institutions. Young researchers will be encouraged to take proactive steps if they are rewarded by hiring committees, promotion committees, fellowship and awards committees, journal editors, and conference committees. Funding agencies could encourage open-source development of equipment that enables integration of high-throughput synthesis of materials with data management. Industries may see value in funding solution-driven system-level approaches to accelerate their development timelines.

Community

Realizing this future requires merging domain expertise currently resident in robotics, software, computer science, electronics, materials, and design silos, each with its own language/acronyms, and academic conferences. The learning curve to become even a generalist in these different domains remains very steep. Reducing barriers to communication and achieving percolation of ideas across domains may be facilitated via cross-cutting conferences, workshops, and creation of funded research centers. Adoption of best practices across various fields can be encouraged via these percolation pathways of ideas.

Education and Up-Skilling

Public opinion (read: support or opposition) to ML/AI is influenced by whether or not citizens can envision a hopeful future that includes their employment and empowers society. First, these transformations require individuals at all levels and employment types to be willing to up-skill. Educators at all levels have an opportunity to revamp their curricula, considering both technical and societal impacts. Online tools and courses for machine-learning/artificial intelligence are growing in availability

Box 1. Practical Resources for the Materials Researcher Interested in Machine Learning

Free materials databases and repositories	Aflow, Duke University	http://aflowlib.org/
	Citrination, by Citrine	http://citrination.com
	Globus, University of Chicago	https://www.globus.org/
	Harvard Clean Energy Project[47]	https://cepdb.molecularspace.org/
	Gateway for Accelerated Materials	http://www.acceleratedmaterials.org
	Materials Data Facility, University of Chicago[48]	https://materialsdatafacility.org
	MatNavi, NIMS	http://mits.nims.go.jp/index_en.html
	Materials Project, LBNL[3]	https://materialsproject.org/
	UCSB/MRL database[49]	www.mrl.ucsb.edu:8080/datamine/thermoelectric.jsp
	NIST Data Curator	https://mgi.nist.gov/materials-data-curation-system
	NIST Materials Data Repository	https://materialsdata.nist.gov/
	NOMAD Repository	http://nomad-repository.eu/
	NREL Materials Database, MatDB[50]	https://materials.nrel.gov
	NREL High-Throughput Experimental Materials Database[35]	https://htem.nrel.gov
	OQMD, Northwestern University	http://oqmd.org/
	Synthesis Project, MIT	http://www.synthesisproject.org/
	Thermoelectrics Design Lab[51]	http://tedesignlab.org
Self-learning resources	Andrew Ng's course, "Machine Learning"	https://www.coursera.org/learn/machine-learning
	Chinese-language resources	https://www.zhihu.com/org/ji-qi-zhi-xin-65/activities
	Machine-learning crash course, Google	https://developers.google.com/machine-learning/crash-course/ml-intro
	Written resources	Kevin Murphy, "Machine Learning: A Probabilistic Perspective," MIT Press, Cambridge, MA, USA, 2012; ISBN 0262018020
Residential learning programs	MIDDMI, Colorado School of Mines & Citrine Informatics	https://www.mines.edu/middmi/about/

These resources have proven useful for the authors and members of their research teams. This is not an exhaustive list (nor an endorsement). For updated list, see http://www.acceleratedmaterials.org.

(see Box 1 for examples), but direct applications to materials science and systems engineering are a growing need. Second, we are invited to consider how we teach reflects the most suitable skills and mindsets to harness the full potential of accelerated materials development and manufacturing platforms. Domain expertise in supporting fields, including advanced statistics, will increase in utility with the mainstreaming of system-level design of experiments. Third, the scientific method will still be valid, and the premium will only increase for asking the right questions, designing good experiments, and disseminating results well.

Conclusions

The convergence of high-performance computing, automation, and machine learning promises to accelerate the rate of materials discovery, better aligning investor and stakeholder timelines. These new tools are set to become an indispensable part of the scientific process. More than ten times faster synthesis, device fabrication, and diagnostics in a (semi-)automated feedback loop are distinctly possible in the near future. Recent advances in theory, high-throughput materials synthesis (as well as components, devices, and systems), diagnostics, the understanding of defects and defect tolerance, and machine-learning methods are fueling

this transition. Further investments in infrastructure and human capital are needed to fully realize this future, including greater emphasis on appropriate applications of existing methods to materials-relevant problems, adoption of data and metadata standards, data-management tools, and laboratory infrastructure, including both decentralized and centralized facilities. To integrate these tools into the R&D ecosystems depends in part on several human elements; namely, the time needed to evolve incentive structures, community support, education and up-skilling offerings, and researcher mindsets, as our field transitions from thinking "data poor" to thinking "data rich." We envision a scientific laboratory where the process of materials discovery continues without disruptions, aided by computational power augmenting the human mind, and freeing the latter to perform research closer to the speed of imagination, addressing societal challenges in market-relevant timeframes.

ACKNOWLEDGMENTS

The ideas represented herein evolved in discussion with numerous individuals, including but not limited to Riley Brandt, Danny Ren, Felipe Oviedo, Daniil Kitchaev, Rachel Kurchin, I. Marius Peters, Shijing Sun, Rafael Jaramillo, Gang Chen, and Anantha Chandrakasan of MIT/SMART; Benjamin Gaddy of Clean Energy Trust; Rolf Stangl, Chaobin He, and Anthony Cheetham of NUS; Dirk Weiss and Raffi Garabedian of First Solar; B.J. Stanbery of Siva Power; Karthik Kumar, Sir John O'Reilly, Pavitra Krishnaswamy, Alfred Huan, and Cedric Troadec of A*STAR; Andrij Zakutayev, Stephan Lany, Dave Ginley, Greg Wilson, and William Tumas of NREL; Lydia Wong, Shuzhou Li, Tim White, and Subbu Venkatraman of NTU; and Marc Bauchet, among many others. The work carried out by one of the authors (S.G.) was at the Center for Nanoscale Materials, a U.S. Department of Energy Office of Science User Facility, and supported by the U.S. Department of Energy, Office of Science, under Contract No. DE-AC02-06CH11357.

REFERENCES

1. Gaddy, B.E., Sivaram, V., Jones, T.B., and Wayman, L. (2017). Venture capital and cleantech: the wrong model for energy innovation. Energy Policy *102*, 385–395.

2. 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. http://nrs.harvard.edu/urn-3:HUL.InstRepos:35164974.

3. Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., and Persson, K.A. (2013). Commentary: the materials project: a materials genome approach to accelerating materials innovation. APL Mater. *1*, 011002.

4. Nosengo, N. (2016). The material code: machine-learning techniques could revolutionize how materials science is done. Nature *533*, 22.

5. De Luna, P., Wei, J., Bengio, Y., Aspuru- Guzik, A., and Sargent, E. (2017). Use machine learning to find energy materials. Nat. Mater. *552*, 23.

6. Icons in Figure 2 are freeware, and were made by Freepik from www.flaticon.com.

7. Pyzer-Knapp, E.O., Suh, C., Gómez-Bombarelli, R., Aguilera-Iparraguirre, J., and Aspuru-Guzik, A. (2015). What is high-throughput virtual screening? A perspective from organic materials discovery. Annu. Rev. Mater. Res. *45*, 195–216.

8. US Department of Energy (2016) Basic Research Needs for Synthesis Science. Report of the Basic Energy Sciences Workshop on on Basic Research Needs for Synthesis Science for Energy Relevant Technology. May 2–4, 2016.

9. Phillips, C.L., and Littlewood, P. (2016). Preface: special topic on materials genome. APL Mater. *4*, 053001.

10. Zunger, A. (2018). Inverse design in search of materials with target functionalities. Nat. Rev. Chem. *2*, 0121.

11. Roch, L.M., Häse, F., Kreisbeck, C., Lars, T.T-M., Yunker, P.E., Hein, J.E., Aspuru-Guzik, A. ChemOS: An Orchestration Software to Democratize Autonomous Discovery. https://chemrxiv.org/articles/ChemOS_An_Orchestration_Software_to_Democratize_Autonomous_Discovery/5953606.

12. Eid, J., Liang, H., Gereige, I., Lee, S., and Van Duren, J. (2015). Combinatorial study of NaF addition in CIGSe films for high efficiency solar cells. Prog. Photovolt. *23*, 269–280.

13. Jeon, M.K., Cooper, J.S., and McGinn, P.J. (2009). Investigation of PtCoCr/C catalysts for methanol electro-oxidation identified by a thin film combinatorial method. J. Power Sources *192*, 391–395.

14. Kojima, A., Teshima, K., Shirai, Y., and Miyasaka, T. (2009). Organometal halide perovskites as visible-light sensitizers for photovoltaic cells. J. Am. Chem. Soc. *131*, 6050–6051.

15. Graetzel, M., Janssen, R.A.J., Mitzi, D.B., and Sargent, E.H. (2012). Materials interface engineering for solution-processed photovoltaics. Nature *488*, 304–312.

16. Yin, W.-J., Shi, T., and Yan, Y. (2014). Unusual defect physics in $CH_3NH_3PbI_3$ perovskite solar cell absorber. Appl. Phys. Lett. *104*, 063903.

17. Brandt, R.E., Stevanović, V., Ginley, D.S., and Buonassisi, T. (2015). Identifying defect-tolerant semiconductors with high minority-carrier lifetimes: beyond hybrid lead halide perovskites. MRS Commun. *5*, 265–275.

18. Perovskite device raw data provided by private communication with Dr. Juan-Pablo Correa-Baena, Dr. Michael Saliba, and Dr. Antonio Abate from the laboratory at EPFL. Raw data to be included in an upcoming article currently

under review (M. Saliba, J.-P. Correa-Baena, C.M. Wolff, M. Stolterfoht, N. Phung, S. Albrecht, D. Neher, A. Abate, How to make over 20% efficient perovskite solar cells in regular (n-i-p) and inverted (p-i-n) architectures). CdTe learning data presented by Dirk Weiss (First Solar) at the Asia Clean Energy Summit PVAsia, October 24 and 25, 2017, Singapore.

19. Zenati, H., Foo, C., Lecouat, B., Manek, G., and Chandrasekhar, V. (2018). Efficient GAN-based anomaly detection. arXiv, 1802.06222.

20. Guimaraes, G.L., Sanchez-Lengeling, B., Outeiral, C., Farias, P.L.C., and Aspuru-Guzik, A. (2017). Objective-Reinforced Generative Adversarial Networks (ORGAN) for sequence generation models. arXiv, 1705.10843.

21. Freysoldt, C., Grabowski, B., Hickel, T., Neugebauer, J., Kresse, G., Janotti, A., and Van de Walle, C.G. (2014). First-principles calculations for point defects in solids. Rev. Mod. Phys. *86*, 253.

22. Goyal, A., Gorai, P., Peng, H., Lany, S., and Stevanovic, V. (2017). A computational framework for automation of point defect calculations. Comput. Mater. Sci. *130*, 1.

23. Kusne, A.G., Gao, T., Mehta, A., Ke, L., Nguyen, M.C., Ho, K.-M., Antropov, V., Wang, C.-Z., Kramer, M.J., Long, C., and Takeuchi, I. (2014). On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets. Sci. Rep. *4*, 6367.

24. Iwasaki, Y., Kusne, A.G., and Takeuchi, I. (2017). Comparison of dissimilarity measures for cluster analysis of X-ray diffraction data from combinatorial libraries. NPJ Comput. Mater. *3*, 4.

25. Brandt, R.E., Kurchin, R.C., Steinmann, V., Kitchaev, D., Roat, C., Levcenco, S., Ceder, G., Unold, T., and Buonassisi, T. (2017). Rapid photovoltaic device characterization through Bayesian parameter estimation. Joule *1*, 843–856.

26. Somnath, S., Law, K.J.H., Morozovska, A.N., Maksymovych, P., Kim, Y., Lu, X., Alexe, M., Archibald, R., Kalinin, S.V., Jesse, S., and Vasudevan, R.K. (2018). Ultrafast current imaging by Bayesian inversion. Nat. Commun. *9*, 513.

27. Li, L., Yang, Y., Zhang, D., Ye, Z.-G., Jesse, S., Kalinin, S.V., and Vasudevan, R.K. (2018). Machine learning–enabled identification of material phase transitions based on experimental data: exploring collective dynamics in ferroelectric relaxors. Sci. Adv. *4*, eaap8672.

28. Ward, L., Agrawal, A., Choudhary, A., and Wolverton, C. (2016). A general-purpose machine learning framework for predicting properties of inorganic materials. NPJ Comput. Mater. *2*, 16028.

29. Lopez, S.A., Sanchez-Lengeling, B., de Goes Soares, J., and Aspuru-Guzik, A. (2017). Design principles and top non-fullerene acceptor candidates for organic photovoltaics. Joule *1*, 857–870.

30. Gómez-Bombarelli, R., Aguilera-Iparraguirre, J., Hirzel, T.D., Duvenaud, D., Maclaurin, D., Blood-Forsythe, M.A., Chae, H.S., Einzinger, M., Ha, D.-G., Wu, T., et al. (2016). Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. Nat. Mater. *15*, 1120–1127.

31. Conduit, B.D., Jones, N.G., Stone, H.J., and Conduit, G.J. (2017). Design of a nickel-base superalloy using a neural network. Mater. Des. *131*, 358–365.

32. Senkov, O.N., Miller, J.D., Miracle, D.B., and Woodward, C. (2015). Accelerated exploration of multi-principal element alloys with solid solution phases. Nat. Commun. *6*, 6529.

33. Green, M.L., Choi, C.L., Hattrick-Simpers, J.R., Joshi, A.M., Takeuchi, I., Barron, S.C., Campo, E., Chiang, T., Empedocles, S., Gregoire, J.M., et al. (2017). Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies. Appl. Phys. Rev. *4*, 011105.

34. Kim, E., Huang, K., Saunders, A., McCallum, A., Ceder, G., and Olivetti, E. (2017). Materials synthesis insights from scientific literature via text extraction and machine learning. Chem. Mater. *29*, 9436–9444.

35. Zakutayev, A., Wunder, N., Schwarting, M., Perkins, J.D., White, R., Munch, K., Tumas, W., and Phillips, C. (2018). An open experimental database for exploring inorganic materials. Sci. Data *5*, 180053.

36. White, R.R., and Munch, K. (2014). Handling large and complex data in a photovoltaic research institution using a custom laboratory information management system. MRS Proceedings *1654*. .Mrsf13-1654-nn11-04. https://doi.org/10.1557/opl.2014.31.

37. 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.

38. Lecouat, B., Foo, C., Zenati, H., and Chandrasekhar, V. (2018). Semi-supervised deep learning with GANs: revisiting manifold regularization. arXiv, 1805:08957.

39. Hutchinson, M.L., Antono, E., Gibbons, B.M., Paradiso, S., Ling, J., and Meredig, B. (2017). Overcoming data scarcity with transfer learning. arXiv, 1711:05099.

40. Ramasamy, S., Rajaraman, K., Krishnaswamy, P., and Chandrasekhar, V. (2018). Online deep learning: growing RBM on the fly. arXiv, 1803:02043.

41. Hill, J., Mannodi-Kanakkithodi, A., Ramprasad, R., and Meredig, B. (2017). Materials data infrastructure and materials informatics. In Computational Materials System Design, D. Shin and J. Saal, eds. (Springer), pp. 193–225.

42. Ananthakrishnan, R., Chard, K., Foster, I., and Tuecke, S. (2015). Globus platform-as-a-service for collaborative science applications. Concurr. Comput. *27*, 290–305.

43. Chard, K., Dart, E., Foster, I., Shifflett, D., Tuecke, S., and Williams, J. (2018). The modern research data portal: a design pattern for networked, data-intensive science. PeerJ Comput. Sci. *4*, e144.

44. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. Sci. Data *3*, 160018.

45. 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 Comput. Mater. *2*, 16031.

46. Cox, C.R., Lee, J.Z., Nocera, D.G., and Buonassisi, T. (2014). Ten-percent solar-to-fuel conversion with nonprecious materials. Proc. Natl. Acad. Sci. USA *111*, 14057–14061.

47. Hachmann, J., Olivares-Amaya, R., Atahan-Evrenk, S., Amador-Bedolla, C., Sánchez-Carrera, R.S., Gold-Parker, A., Vogt, L., Brockway, A.M., and Aspuru-Guzik, A. (2011). The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. J. Phys. Chem. Lett. *2*, 2241–2251.

48. Blaiszik, B., Chard, K., Pruyne, J., Ananthakrishnan, R., Tuecke, S., and Foster, I. (2016). The materials data facility: data services to advance materials science research. JOM *68*, 2045–2052.

49. Gaultois, M.W., Sparks, T.D., Borg, C.K.H., Seshadri, R., Bonificio, W.D., and Clarke, D.R. (2013). Data-driven review of thermoelectric materials: performance and resource considerations. Chem. Mater. *25*, 2911–2920.

50. Stevanovic, V., Lany, S., Zhang, X., and Zunger, A. (2012). Correcting density functional theory for accurate predictions of compound enthalpies of formation: fitted elemental-phase reference energies. Phys. Rev. B *85*, 115104.

51. Gorai, P., Gao, D., Ortiza, B., Millerd, S., Barnett, S.A., Mason, T., Lv, Q., Stevanović, V., and Toberer, E.S. (2016). TE design lab: a virtual laboratory for thermoelectric material design. Comput. Mater. Sci. *112A*, 368–376.