Matter

CellPress

MATTER OF OPINION

The Materials Research Platform: Defining the Requirements from User Stories

Muratahan Aykol,[1,*] Jens S. Hummelshøj,[1] Abraham Anapolsky,[1] Koutarou Aoyagi,[2] Martin Z. Bazant,[3] Thomas Bligaard,[4,5] Richard D. Braatz,[3] Scott Broderick,[6] Daniel Cogswell,[3] John Dagdelen,[7,8] Walter Drisdell,[7] Edwin Garcia,[9] Krishna Garikipati,[10] Vikram Gavini,[10] William E. Gent,[11] Livia Giordano,[3] Carla P. Gomes,[12] Rafael Gomez-Bombarelli,[3] Chirranjeevi Balaji Gopal,[1] John M. Gregoire,[13] Jeffrey C. Grossman,[3] Patrick Herring,[1] Linda Hung,[1] Thomas F. Jaramillo,[4,11] Laurie King,[11,18] Ha-Kyung Kwon,[1] Ryosuke Maekawa,[2] Andrew M. Minor,[7,8] Joseph H. Montoya,[1] Tim Mueller,[14] Colin Ophus,[7] Krishna Rajan,[6] Rampi Ramprasad,[15] Brian Rohr,[11] Daniel Schweigert,[1] Yang Shao-Horn,[3] Yoshinori Suga,[2] Santosh K. Suram,[1] Venkatasubramanian Viswanathan,[16] Jay F. Whitacre,[16] Adam P. Willard,[3] Olga Wodo,[6] Chris Wolverton,[17] and Brian D. Storey[1,*]

A recent meeting focused on accelerated materials design and discovery examined user requirements for a general, collaborative, integrative, and on-demand materials research platform.

What common elements are necessary to create a general framework for materials innovation? Here, we provide a retrospective analysis of high-level themes that emerged from a focused discussion on the requirements for a future materials research platform. These discussions occurred during the annual Toyota Research Institute—Accelerated Materials Design and Discovery meeting (May 29, 2019, in Boston, MA) with more than 40 field experts from universities, US national laboratories, Toyota Research Institute in the United States, and Toyota Motor Corporation in Japan. These researchers contributed ideas toward what capabilities such a platform should have to accelerate the design and discovery of materials.

To maximize the information captured from these field experts' ideas, we followed a strategy comprised of exploitation and exploration inspired by knowledge acquisition strategies in machine learning. In a first session, we matched researchers with themes that we expect them to be knowledgeable about (hence exploitation). In a follow-up session, we did a quasi-random assignment of researchers to themes (hence exploration), with the goal of capturing unique ideas that might elude experts embedded deep in their work. Sixteen pre-selected themes relevant for a materials research platform were discussed twice in two unique groups, one formed as an exploitation and the other formed as an exploration team. The selected themes were: Adaptive systems—active-learning and beyond; Automation of experiments; Automation of simulations; Collaboration; Data ingestion and sharing; Integration; Knowledge discovery; Machine learning for experiments; Machine learning for simulations; Multi-fidelity and uncertainty quantification; Reproducibility and provenance; Scale bridging; Simulation tools; Software infrastructure; Text mining and natural language processing; and Visualization. We digitally recorded the ideas in the form of "user stories" from agile software development practices. The meeting captured many stories that would be considered obvious, but several unexpected ideas emerged. An initial parsing revealed three interrelated themes for the design of a useful platform: (i) data and knowledge assets, (ii) automation of science, and (iii) integrative approaches, as outlined in Figure 1.

The ideas related to data and knowledge assets rely on the FAIR principles[1] but with added distinguishing capabilities relevant for a materials research platform. For example, the platform should enable sharing and collaboration—not just around data, but also knowledge assets such as machine-learning models or scientific workflows. Baselines to gauge new findings are critical but often overlooked. Artificial intelligence (AI)-assisted search and visualization could amplify the scientific abilities of human researchers. As part of automation of science, experimental workflows are envisioned to be run "on-demand", where tasks are picked up by relevant laboratories with fully or partially automated experimental capabilities, forming collaborative networks via the platform. Given

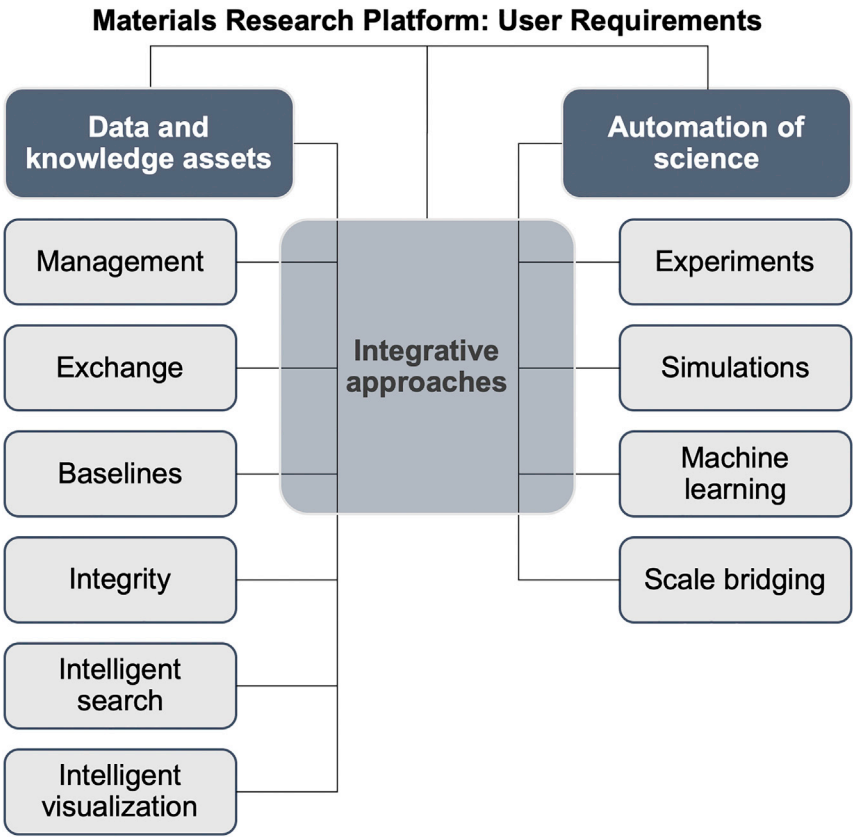

Figure 1. Requirement Categories Identified by the Researchers as the Main Pillars of a Useful Materials Research Platform: Data and Knowledge Assets, Automation of Science on the Platform, and Integrative Approaches

the state-of-the-art for automation in modeling and simulation, a similar but more "productized" automated capability with web-based user interfaces is envisioned to assist researchers to run on-demand simulations complementary to experiments, or train or use on-demand machine-learning models without deep knowledge in any particular method. Automation of scale bridging, which would entail at a minimum designing workflows for codes beyond density functional theory (DFT) and multi-scale application programing interfaces (APIs) for linking the codes, emerged as a game-changing capability to bridge the gap between the computer-design or laboratory and device level properties. The design of integrative approaches on the platform that leverage diverse datasets, physical, empirical, computational, or statistical models and experiments would require a modularized software framework, cost estimation functions, uncertainty quantification and fidelity assessments of new data points, and subsequently, discovery "engines" like optimal experiment design (OED), active-learning, or optimization can be built for diverse material-device domains.

Materials science of the future is expected to be interwoven with data, automation, machine learning, and other emerging information technologies. Many aspects of these paradigms are being actively reviewed, debated, and discussed by the materials community.[2–4] In this article, we focus on the requirements for a useful, general, next-generation materials research platform that would combine and expand on these data-driven paradigms to enable discovery and innovation. A number of academic and industry teams are actively engaged in efforts relevant for this vision (see Web Resources).[2,5–8] We expect that the entire materials community will benefit from the distilled summary of ideas about the requirements of the envisioned future system that we present in this article. The following sections expand the concepts underlying the platform related themes of data and knowledge assets, automation of science, and integrative approaches for materials research.

The Core of the Platform: Data and Knowledge Assets

A research platform inherits, generates, stores, and serves data and knowledge as part of its mission. Such a platform's utility for the scientist, therefore, is tied with how these key bits of information are assimilated and managed; how their quality, reliability, and integrity are judged; how their exchange and dissemination are enabled; and ultimately, whether the informatics aspect of the platform is meeting the needs of the scientist and helping them innovate.

Data Management

The requirements regarding data management are a subset of the now well-accepted FAIR principles: findable, accessible, interoperable, reusable.[1] As a basic requirement, the platform should contain standardized datasets. Standardization of all data imported to, created on, and disseminated by the platform would entail adoption of established ingestion procedures, data formats, capturing of metadata (e.g., experimental conditions), provenance and instrument logs, for instance to differentiate human versus machine generated data. In addition, the platform should deliver not only well-known computational databases, but also diverse, large, high-quality *experimental* datasets, where inclusion of

"dark data" (i.e., data that is considered a "negative" result and not publishable) is essential. Inclusion of all data is key to removing human bias from datasets. Interaction with the data system needs to be easy and intuitive, programmatically or via a web-based user interface (UI) that allows easy or automated upload, instant visualization, search, and sharing and is connected to the other components of the platform. Importance of a simple yet powerful UI for all components of the platform constantly came up throughout the discussions: we will not repeat that requirement, and it should be assumed by default to be a core component for every module hereafter.

Collaboration and Knowledge Exchange

Data sharing is a core component of today's data-driven research. The storage and sharing of analytical tools, machine-learning models, workflows and other knowledge (overall, knowledge assets) as well as experimental resources (for instance, see section labeled "Automation of Experiments") via the platform would enable a more complete, collaborative research experience. The system is envisioned to enable citations for all such shareables and provide utilization logs, allowing the apportioning of separate credit to datasets, models, and scientific results.[5] Furthermore, the platform may allow rapid user feedback and community review, which will motivate developers to maintain quality components on the platform (e.g., clearly licensed, well-documented, and maintained code repositories). Given the diversity of the materials research community, the platform is expected to balance certain users' and institutions' desire for privacy while incentivizing the sharing of methods, data, and scientific results. As such, some interactions may be open collaborations or crowdsourcing, others might require the platform to act as a marketplace or exchange, while still others can have a training or educational component.

Baselines

Baselines help gauge where a new scientific finding stands but are often lacking. The data system on the platform can host curated baseline materials and device components with pertinent properties and operating conditions so that new material discoveries can be compared to the relevant state of the art. The same is true for analytical methods and tools, which would require having standardized, benchmark datasets (e.g., curated DFT datasets) and baseline models trained on them.

Data Integrity Tools

Integrity of data is critical and can be enabled in part by standardization and ingestion procedures. More advanced systems, which are often not part of existing scientific workflows (experiments in particular), such as data validation and anomaly detection, can reinforce data integrity. Such capabilities, if they operate on the fly, can increase the value of the data and can make the platform attractive for research groups that produce live data streams. Since discoveries, especially of high-performance materials, may appear as outliers, it is critical to not only detect but also verify, via reproducibility studies, all apparent outliers.

Intelligent Search

Reusability and discoverability of data on the platform are expected by default. A feature that emerged as part of multiple themes is an "intelligent search" system for materials research. The system is envisioned to operate beyond chemical formula or material labels, and can search over properties, models (e.g., machine-learning models), methods (e.g., synthesis recipes, characterization methods), tools, and other metadata.[9] A search capability that is based on data itself (e.g., similarity match with a user supplied XPS spectrum) would be a game changer. More context aware, interactive search capability beyond simple keywords should be implemented (e.g., "find alternatives to sol-gel synthesis of Cr_2O_3"). In general, the system could recommend new research directions, including materials, techniques, keywords, collaborators, or publications. Specialized recommender systems, as described above for materials, keywords, collaborators, publications, or simulation tools came up in distinct groups.

Intelligent Visualization

A powerful, scientifically focused visualization technology can be considered a distinguishing feature of a next-generation platform. A useful capability would include standard, automated exploratory data analysis and visualization for data on the platform. Visualizations should extend beyond current workflows (e.g., plotting experimental or computational data, and data derived from those) to visualization of high-dimensional parameter spaces (e.g., embeddings in machine-learning models), visualizing relationships between codes, models, and simulations (e.g., see Scale bridging) or mapping data provenance and property relationships, where graph- or network-based representations may be useful.

Automation of Science on the Platform

Generation of high-quality, large-volume, and consistent data streams is often enabled by automation of manual tasks, making processes less prone to human error and increasing throughput. Automation of experiments and simulations are two fundamental paradigms that were considered, where both converged to a desired on-demand capability on the platform. Automation of other components, such as machine learning and scale bridging independently emerged.

Automation of Experiments

Automation of experiments is expected to provide critical functionalities for materials discovery such as reduction of human bias and enabling rapid access to multiple material design axes such as composition, reproducibility, or processing. As a basic functionality, the platform is expected to be integrated with a distributed system of experimental facilities to connect to their data streams and to enable experiment requests.[6] The platform should provide capabilities for creation, execution, and moderation of workflows that span one or more experimental facilities and also should recommend such workflows for specific applications, experimental cost estimates, and fidelity. The platform can provide "on-demand experiments", a marketplace for experiments or a collaborative closed-network, where a user can, for example, request a synthesis of a target material or characterization of a sample at the participating facilities or labs. Importance of automation of low-throughput, repetitive experiments with the aid of robotics was also highlighted.

Automation of Simulations

Computer simulations are, by their nature, more amenable to automation than experiments. Automation of DFT formed the seed for the present era of data-driven materials science by providing large, reliable material datasets.[7] Thus, the stories related to automation of simulation focused on capabilities beyond automation of DFT itself, such as molecular dynamics (MD), coarse-graining methods, phase field, and beyond to predict macroscopic and device level properties. In addition, in analogy with on-demand experiments, a paradigm of "on-demand simulations" emerged, where the platform can provide an easy to use interface for users (simple enough to be useful to non-specialists) to request new simulations complementary to their ongoing experiments. A recommender system for types of simulations, parameters, and ready-to-use license arrangements would add value. As mentioned before, the platform should display relevant benchmarks for simulation tools (e.g., accuracy, performance, cost) and document use cases.

Automation of Machine Learning

An easy-to-use machine-learning and analytics module on the platform, backed by a powerful UI that requires no deep expertise, developed as a common feature desired among multiple discussion groups. To create the necessary input for training predictive models, automated featurization of materials (or other entities) should be a part of this module. In addition, for more advanced practitioners and more complex predictive problems, a comprehensive machine-learning arsenal can be provided: e.g., for image processing, spectroscopy, natural language processing (NLP), deep learning, machine learning for rare events, failures, stochastic events, time-series, material-processing relationships, microstructure-property relationships, physically informed machine-learning models, noisy data, and generative and evolutionary models. The UI should display convenient visual information, such as performance metrics and feature importance in models, and the system should alert the user when there are concerns about the data integrity or quality (bias, anomalies, etc., see section "Data Integrity Tools"). In addition, unsupervised methods that identify relationships in the data and/or capture low dimensional representations should be available.

Automation of Scale Bridging

Scale bridging is required for a more complete assessment of device-level properties of material systems. Often, a small change in material properties used as part of a device requires redesign or re-evaluation of many other components of the same device. Incorporation of new materials in devices has traditionally been a long, slow, and costly process. Today, scale bridging is still a major roadblock in materials research and is mostly performed on an *ad hoc* basis. The need for automated scale bridging was strongly emphasized and also acknowledged as a major scientific challenge. As the most basic functionality, a visual relationship between simulation techniques that can operate at multiple scales has the potential to guide the users toward a hand-crafted scale-bridging study. It was noted that the data transfer between different scales is not sequential or one-directional, and there can often be data transferred from all scales to the others (e.g., DFT to device level, DFT to MD, MD to device level, DFT to finite element, and so on), and the transfer can be bi-directional (e.g., device design informs phase-field or phase field informs device design).

For effective scale bridging, one should parameterize and automate simulations beyond DFT and describe the "contracts" and dependencies between inputs and outputs of such simulations. Such contracts can be framed as "scale-bridging APIs", where input property requirements for methods are documented and codified, to enable programmatic integration between simulation software that operates at different length and timescales. As mentioned above, a graph or network of data transfers and dependencies of simulation tools can be constructed.

Integrative Approaches on the Platform

The research envisioned to be enabled on the platform requires blending or integration of many components, tools, and/or datasets. Several such paradigms emerged from the user stories that are centered on integration, where the ability to be automated and

modularized was an expected quality for all relevant tools. Data fusion, where multiple datasets are combined to enrich the existing data, and to create new datasets, is the most basic example. Scale bridging is a fundamental integration challenge highlighted in the previous section.

An emerging paradigm for discovery of materials and processes is the application of cyclic, active-learning, optimization, or OED based feedback-loop systems, where the science (and the underlying decision making) itself is partially automated.[4] The materials research platform should provide modular, plug-and-play, automatable closed-loop capability to enable this form of research. This capability needs to be easy to integrate with both experimental and computational data streams. In addition to the software and analytical infrastructure required for the process, cost-estimating functions for experimental and simulation processes emerged as a key feature to have on the platform.

Uncertainty quantification is essential for automated, closed-loop research systems. Experts highlighted the importance of having uncertainty estimates or confidence intervals available on all experimental and computational measurements, parameters, and outputs. How uncertainty propagates as data are being transformed (e.g., in machine learning or scale bridging), and how that affects the reliability of resulting predictions also remains an open question. Experts mentioned the potential benefit from incorporation of information-theoretic approaches into the system (e.g., information gain), as well as availability of classification tables for fidelities and costs of acquiring experimental or computational data points or those from surrogate models (machine-learning models, empirical models, etc.). Ultimately, multi-fidelity optimization where uncertainties, fidelities, and cost are taken into account offer a viable, general pathway for integration of computational and experimental data generation pipelines to solve complex scientific problems. These features should have a presence on the platform as modular systems.

WEB RESOURCES

Materials Project, https://materialsproject.org

AFLOW, http://www.aflowlib.org

OQMD, http://oqmd.org

JARVIS-API, https://jarvis.nist.gov

NREL MatDB, https://materials.nrel.gov

Materials Cloud, http://www.materialscloud.org

NOMAD Repository, https://nomad-repository.eu

Citrine Informatics, https://citrine.io

Schrodinger, https://www.schrodinger.com

Granta Design, https://grantadesign.com

Exabyte.io, https://exabyte.io

Kebotix, https://www.kebotix.com

DECLARATION OF INTERESTS

Authors with Toyota affiliations declare internal support. Remaining authors have or work as part of projects supported in part or full by Toyota Research Institute.

1. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data *3*, 160018, https://doi.org/10.1038/sdata.2016.18.

2. Alberi, K., Nardelli, M.B., Zakutayev, A., Mitas, L., Curtarolo, S., Jain, A., et al. (2019). The 2019 Materials by Design Roadmap. J. Phys. D Appl. Phys. *52*, 13001, https://doi.org/10.1088/1361-6463/aad926.

3. Tabor, D.P., Roch, L.M., Saikin, S.K., Kreisbeck, C., Sheberla, D., Montoya, J.H., et al. (2018). Accelerating the discovery of materials for clean energy in the era of smart automation. Nat. Rev. Mater. *3*, 5, https://doi.org/10.1038/s41578-018-0005-z.

4. Balachandran, P.V., Xue, D., Theiler, J., Hogden, J., and Lookman, T. (2016). Adaptive Strategies for Materials Design using Uncertainties. Sci. Rep. *6*, 19660, https://doi.org/10.1038/srep19660.

5. 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, https://doi.org/10.1007/s11837-016-2001-3.

6. Green, M.L., Choi, C.L., Hattrick-Simpers, J.R., Joshi, A.M., Takeuchi, I., Barron, S.C., et al. (2017). Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies. Appl. Phys. Rev. *4*, 11105, https://doi.org/10.1063/1.4977487.

7. Lin, L. (2015). Materials Databases Infrastructure Constructed by First Principles Calculations: A Review. Mater. Perform. Charact. *4*, 148, https://doi.org/10.1520/MPC20150014.

8. Kim, C., Chandrasekaran, A., Huan, T.D., Das, D., and Ramprasad, R. (2018). Polymer genome: a data-powered polymer informatics platform for property predictions. J. Phys. Chem. C *122*, 17575, https://doi.org/10.1021/acs.jpcc.8b02913.

9. Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., et al. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. Nature *571*, 95–98, https://doi.org/10.1038/s41586-019-1335-8.

[1]Toyota Research Institute, Los Altos, CA 94022, USA

[2]Toyota Motor Corporation, Toyota, Aichi, 471-8572, Japan

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

[4]SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA

[5]Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark

[6]University at Buffalo, Buffalo, NY 14260, USA

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

[8]University of California Berkeley, Berkeley, CA 94720, USA

[9]Purdue University, West Lafayette, IN 47907, USA

[10]University of Michigan, Ann Arbor, MI 48109, USA

[11]Stanford University, Stanford, CA 94305, USA

[12]Cornell University, Ithaca, NY 14853, USA

[13]California Institute of Technology, Pasadena, CA 91125, USA

[14]Johns Hopkins University, Baltimore, MD 21218, USA

[15]Georgia Institute of Technology, Atlanta, GA 30332, USA

[16]Carnegie Mellon University, Pittsburgh, PA 15213, USA

[17]Northwestern University, Evanston, IL 60208, USA

[18]Present address: Manchester Metropolitan University, Manchester M15 6BH, United Kingdom

*Correspondence: murat.aykol@tri.global (M.A.), brian.storey@tri.global (B.D.S.)

https://doi.org/10.1016/j.matt.2019.10.024