An Accelerated, Data-Driven, Materials Discovery Future

Challenges and opportunities in two dimensional, interfacial, and layered materials to advance science and engineering and impact society

ENVISIONING ACCELERATED MATERIALS DISCOVERY

Co-Chairs

Adri van Duin, Pennsylvania State University and 2DCC
David Elbert, Johns Hopkins University and PARADIM
Tyrel M. McQueen, Johns Hopkins University and PARADIM

Mentors

Laura Bartolo, Senior Research Associate, Center for Hierarchical Materials Design (CHiMaD), Northwestern University
Gretchen Greene, Data Sciences Group, National Institute of Standards and Technology
Richard Hennig, Professor, Alumni Professor of Materials Science & Engineering, University of Florida
June Lau, Office of Data and Informatics, National Institute of Standards and Technology
Jeffrey Neaton, Former Director Molecular Foundry and Professor of Physics, UC Berkeley

Special Assistance

Toby Scott, KnowInnovation
Emma Skipper, KnowInnovation
Sara Bradley, KnowInnovation
Jasmine Harris, Johns Hopkins University and PARADIM

Steering Committee

Joan Redwing, Professor Pennsylvania State University and Director NSF 2DCC
Darrell Schlom, Professor Cornell University and Co-Director NSF PARADIM
Stephen D. Wilson, Professor UC Santa Barbara and Co-Director NSF Quantum Foundry

Forum Participants

Given in Appendix B

Acknowledgements

We acknowledge support of the forum through a supplement to the National Science Foundation (Platform for the Accelerated Realization, Analysis, and Discovery of Interface Materials (PARADIM)) under Cooperative Agreement No. DMR-2039380, and thank all for their insights and feedback.

Please cite as: Ed. A.C.T. Van Duin, D. Elbert, T.M. McQueen, “An Accelerated, Data-Driven, Materials Discovery, Future,” doi:10.34863/g88h-kq06 (2022).

Executive Summary

Materials research output, as judged by number of publications, has been doubling every 40 months. There is recognition that regardless of the precise physical form a material (e.g. nanoparticle, 2D layer, thin film, bulk ceramic, single crystal), it is the interfaces – the surfaces, the connections between grains, the internal boundaries between a structure and its point or extended defects – that conspire with the intrinsic ("bulk") properties to determine materials behavior. This has led to the burgeoning field of 2D layered and interfacial materials (2DILM), the tailored creation and study of such interfaces. A cornucopia of new data science techniques, such as those popularized as artificial intelligence / machine learning (AI/ML), are succeeding in tackling problems once considered almost impossible – whether it is superiority over humans in games such as Chess or Go, or computer-driven object image recognition, or speech recognition and understanding of the written word.

What new materials advances are possible through the convergence of the 2DILM and AI/ML fields? This report seeks to answer this question based on community input gathered in multiple stages, culminating in a pair of forums in January and March 2022. The community identified significant societal benefits from the convergence of materials with AI/ML, particularly in the ability to advance predictability of synthetic reactions and to enable preparation of materials not possible without automated control, leading to advancing materials research and fundamental knowledge, and scholarship, with applications to energy, health care, infrastructure and communications, manufacturing, and space technology and exploration (section I.2). Within the community, we have identified four areas in which immediate research and actions can lead to realization of these future visions. Specifically, we highlight focused efforts to:

1. Predict Outcomes of Actual Synthesis Conditions, by
    a. Establishing a regular, critical assessment of phase diagram predictions vs experimental reality, analogous to the assessments in protein folding that culminated in AlphaFold.[1]
    b. Establishing the science to more rapidly produce experimental phase diagrams and include kinetics on an equal footing to thermodynamics within these diagrams.
2. Achieve AI-assisted and Autonomous Synthesis, by
    a. Identification of specific materials discovery and synthesis problems where ML and AI can aid to inform human intelligence to accelerate synthesis research – e.g. to prepare compounds on the verge of instability (where phenomenal properties usually emerge).
    b. Developing the techniques for "AI and human in the loop" synthesis and processing.
    c. Setting the expectation of broad dissemination of such tooling across the nation.
    d. Encouraging development of methods to extract "how" and "why" AI/ML models reach the conclusions they do.
3. Develop the Data Materials Science Workforce, by
    a. Convening practitioners and educators in a transdisciplinary way to lay out best practices and develop community resources enabling seamless incorporation of data.
    b. Encouraging hybrid team science across chemistry/physics/materials science/materials engineering/computer science.
4. Enable Data Curation and Community Use, by
    a. Establishing the community expectation that the data supporting figures and tables of publications be shared publicly with an explicit license or usage policy.
    b. Encouraging development of the science behind effective and interoperable materials data models.

Table of Contents

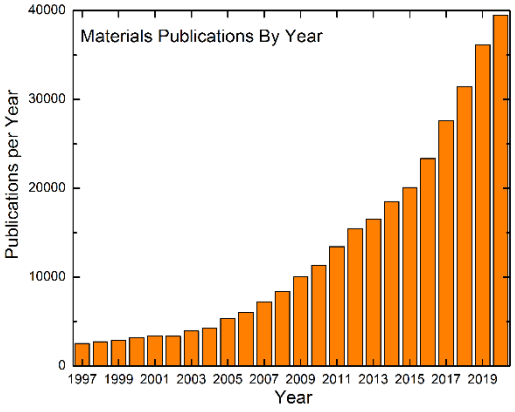
I. Introduction

I.1. Landscape and Context

"Data-intensive Scientific Discovery is a Fourth Paradigm for Science" – Jim Gray (quoted by Hey, 2007)

Recent reviews have highlighted novel and unusual optical, electronic, magnetic, and mechanical properties of two-dimensional, interfacial, and layered materials (2DILM).[2-9] Harnessing these properties is widely believed to hold promise for wide-ranging, transformative impacts for society by creating innovative technological opportunities and applications across sectors as diverse as energy generation and storage; healthcare and biomedicine development; cyberinfrastructure and national security; and environmental sustainability.[7-17] This promise and the fundamental scientific challenges encompassed in understanding and creating 2DILM materials has driven international investment in research and research infrastructure to accelerate their realization including development of artificial intelligence and machine learning methods to leverage the burgeoning materials data landscape of the Materials Genome Initiative.[18-27] The timing, scope, and future directions of such investment will be driven not only by the potential impact of 2DILM materials, but by the maturity of resources and community readiness needed to meet goals that require collective and coordinated initiatives/approaches.

Acceleration of research through coordinated community pursuit of grand-scale challenges is a goal that spans scientific fields. Perhaps the most well-known exemplar of such community collaboration for strategic goals is the International Technology Roadmap for Semiconductors and its successor.[28] The hallmark of such efforts is the recognition of the community that the goals are too big to be accomplished by traditional research groups or collaborations; by sharing a common vision and clear waypoint goals along the path to that vision the community has created broad opportunity and high-impact results that are synonymous with Moore's Law and the scaling of processor advancement over decades. The advent of the Materials Genome Initiative (MGI)[25] in 2011 ushered in the beginning of broad recognition that materials research could sustain a similar period of accelerated advance, in this case by communal sharing and endorsement of data and data driven methods.

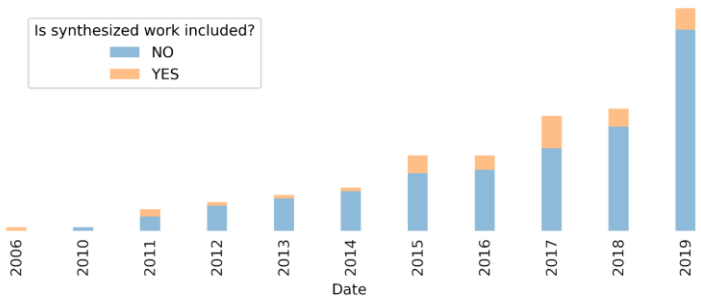


Materials research is rapidly doubling every 40 months.

The 2017 TMS report Building a Materials Data Infrastructure: Opening new Pathways to Discovery and Innovation in Science and Engineering[29] highlights common challenges and provides a focused review of the importance of data in science and engineering. Their review speaks to aspects of the data lifecycle, success stories from other fields (notably geosciences and life sciences), and specific prior efforts from the materials community. Important common challenges include a lack of data sharing incentives; incomplete or inappropriate data formats; intellectual property and data ownership restrictions; uncertainty about valuation

of data lifecycles; and a lack of data-aware performance metrics.[29-30] These challenges are ones we routinely hear from colleagues at data workshops and events when engaging more technical problems related to the diversity and enormous volume of materials research data. The 2021 revision of the MGI Strategic Plan,[25] MGI 2.0, emphasizes the importance of community coordination to drive implementation of FAIR data principles and development of innovative materials data infrastructure. The Materials Research Data Alliance (MaRDA), a major outcome of the 2019 Summit of Big Data and Cyberinfrastructure in Materials Research co-organized by PARADIM and NanoHub, is an instantiation of just such a materials research community network which provides a framework to connect focused efforts.[31]

Materials *Predictions* Dominate

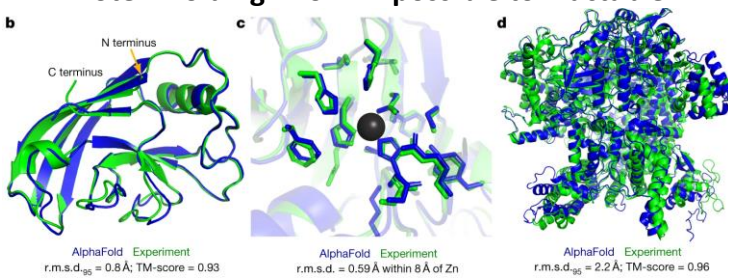


The impact of the MGI on the *prediction* of new materials that improve or produce new functionality is profound. The above shows the high throughput materials predictions in energy materials from 2006 to 2019, and the fraction that contain synthetic work, showing a rapid increase in predictions and that experimental validation of predictions is not keeping pace. How do we improve predictions and accelerate synthetic validation? Courtesy I. Dabo.[32]

The importance of advancing the materials data infrastructure arises from a revolution in our ability to capture, store, and process data with the transformative new ways science is proceeding by focusing on data. Leveraging terabytes, even petabytes of data is now routine in the business world and has sparked a revolution in data-driven science (sidebar). The 2018 National Academies report Open Science by Design: Realizing a Vision for 21st Century Research[19] affirms this importance and the integral need for open science and data across disciplines, as does a more recent 2022 National Academies report on Automated Research Workflows for Accelerated Discovery.[33] NASA's 2018 report, Vision 2040: A Roadmap for Integrated, Multiscale Modeling and Simulation of Materials and Systems[34] also emphasizes the opportunity materials data provide to enable development of novel materials while realizing MGI goals to accelerate the pace of discovery and deployment of new materials. Each of these reviews provide valuable information on current materials data infrastructure calling out gaps and challenges to be addressed.

In addition to opportunity and need, timing is important. The impetus to push ahead now is clear given the breadth of tools available to make open data practical as well as the strategic pressures of commitments already bearing fruit in Europe and Asia. In 2000, China launched two centralized materials databases involving 18 research institutes with data collected and entered in a standardized

Protein Folding: From Impossible to Tractable



The problem of predicting the structure of a protein from its sequence was a 50 year grand challenge.[37] After years of slow but steady improvements, the field was transformed from 2018 to 2022 with the introduction of AlphaFold, a structure prediction approach combining AI/ML techniques with domain specific knowledge and a large corpus of experimentally determined protein structures. It has demonstrated atomic level accuracy even in the absence of comparable structures in the training set,[1] and took what was once considered an impossible task, and made it tractable. This was possible due to a confluence of available experimental data, a community driven biannual competition to fairly evaluate techniques (CASP - Critical Assessment of protein Structure Prediction), and the development of domain-specific AI/ML techniques. What similar advances are the 2DILM community ready to tackle?

format.[26,35] Their 2016 Materials Genome Engineering Project[35] has moved quickly with substantial government investment. These and other efforts are paying dividends as they help transform China into a high-tech power.[35] The European Commission Joint Research Centre has organized over 5000 datasets into the MatDB enabling system integration, data interoperability, and minting of digital object identifiers.[36] Organizations such as OpenAIRE and EUDAT provide support mechanisms and technical infrastructure for researchers to support data stewardship and open science.[22,23]

2DILM research is at a crossroads created by the confluence of rapid advances in high-quality data production, broad availability of innovative cyberinfrastructure, a growing community focused on sharing data resources, the maturation of NSF's first two Materials Innovation Platforms (MIPs), PARADIM and 2DCC providing synthesis of 2DILM materials and knowledge sharing of data and know-how through development of infrastructure and a research ecosystem to drive advances in the field, the founding of the HDR institutes including the Institute for Data Driven Dynamical Design, the creation of the AI Institutes and AI Planning efforts including those for 4D Materials Science and AI-enabled Materials Discovery, Design, and Synthesis, and the creation of the DOE Artificial Intelligence and Technology Office and the Artificial Intelligence Advancement Council. The combination of these elements has set the stage for directed, rapid data-driven advances to accelerate discovery and deployment of high-value 2DILM materials, especially when combined with parallel advances in algorithms and methods to infer meaning from such data. In addition to its own inherent value to technological development, a clear path for data-driven 2DILM research can provide a roadmap/exemplar for other areas of materials research, translational methods for next-generation manufacturing, and fundamental challenges for AI development with broad impact. Materials research sits at the intersection of physics, chemistry, materials science, and engineering. Convergent opportunities to transform materials research arise from changes in these fields. Generational opportunities arise from changes across all these disciplines. The rise of data-driven, AI/ML empowered methods and a community ready to embrace them puts materials research on the cusp of generational change. The potential impacts of data-rich 2DILM exist (section I.2). The community is ready and has converged on a common vision and goals (section I.3). The structures to encourage synergies in this area are just now forming (section I.4).

I.2. Motivation and Potential Societal Outcomes

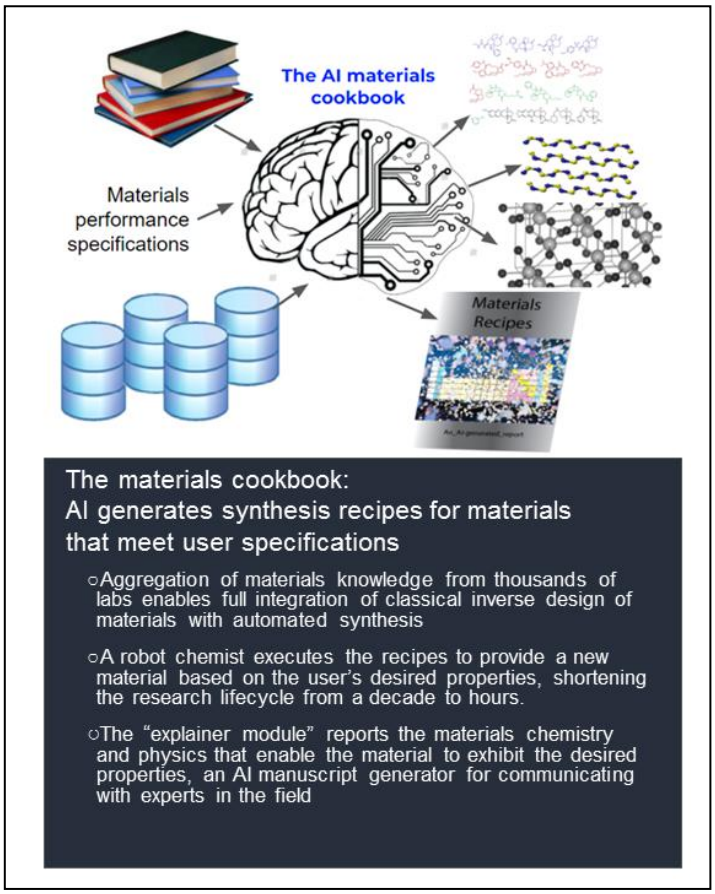
"We will not rest until the periodic table is exhausted," - Intel CEO Pat Gelsinger (2022)

Ever since the creation of the first tools, the development of materials has gone hand-in-hand with improvements in communities and society. Indeed, advanced materials are so integral, diverse, and ubiquitous in the fabric of everyday life that their extreme performance and utility are often and paradoxically underappreciated and taken for granted. Yet almost every material advance has the potential to lead to improvements in society.

Given this enormous scope, it is important to identify those domains in which advances in data-driven materials discovery for two dimensional, interfacial, and layered materials would be particularly impactful. Beyond the obvious deep connection to semiconductor devices alluded to by the quote above,[10] the forum identified significant impacts in the following sectors:

- **Energy.** Any device doing useful work requires power. 2D materials might enable a closed cycle photocatalytic hydrogen generation and fuel cell power source to replace or augment batteries in many devices.
- **Fundamental Knowledge and Scholarship.** Access to instrumentation and expertise bottlenecks participation in science and engineering. A 2D materials virtual synthesis laboratory would democratize discovery, enabling more equitable access to state-of-the-art techniques and expertise, e.g. by empowering those with certain physical disabilities to carry out research that would have been impossible before.
- **Health Care.** Treatment and prevention of disease is necessarily personalized, depending on each person's genetic and environmental history. Atomically thin biocompatible materials with embedded functionality may allow for identification and elimination of cancer cells in vivo.
- **Infrastructure and Communications.** The internet of things has popularized the opportunities brought by seamless integration of computing and networking with human activity. Extremely low power neuromorphic chips built of 2D materials enable this reality, operating without external power sources (much like one can build a radio that runs on the power of the radio waves[38]) and with on-device processing to minimize communication overhead.
- **Manufacturing.** The semiconductor revolution was enabled by the ability to build intricate structures at the micron and nanometer scale. If the ability to controllably layer atomically thin 2D materials is added, then bottom-up, precise, atomic-scale 3D manufacturing becomes a reality, enabling creation of not just new semiconductor platforms but also the creation of materials with superior combinations of thermal, electrical, and mechanical properties.
- **Space Technology and Exploration.** Space, and neighboring planets, are harsh environments. Atomically-thin materials might enable the fabrication of large solar arrays or catalysts for food or energy production that can be rolled up and efficiently transported off world, with sufficient chemical durability to survive.
- **Transforming Materials Research.** Knowing what to make and how to make it is hard. Combining AI techniques with aggregation of materials knowledge and automated materials synthesis platforms may enable automatic – or at least guided – generation of synthesis recipes for materials meeting user specifications.

These impacts were identified and framed in the context of “future headlines” that might result from investment in 2DILM research, and have associated specific steps that must be taken or challenges overcome to realize those impacts.



A prominent aspect of the areas and mechanisms of impact identified is the commonality of underlying themes. The potential of 2D interfacial and layered materials to change the energy landscape – be it through replacements for lithium ion batteries, better energy conversion devices (catalysis) or cutting power consumption for useful tasks by orders of magnitude – is notable. The importance of advancing both synthesis and materials design – whether through virtual laboratories, new ways to put every atom in its place, or the aggregation of data across domains – is striking. Multiple impacts also identified the importance of access – and in particular the democratization of access to all components of the materials pipeline.

Considered as a whole, within materials research, a particularly compelling future headline was: “The Materials Cookbook: AI generates synthesis recipes for materials that meets user specifications,” The vision being that synthesis is often the key bottleneck in materials development, and that data-rich techniques could enable a level of control over general materials synthesis that has not previously existed [except, possibly, within the narrow domain of organic retrosynthesis]. Such developments could then couple to emerging efforts in 3D printing (“Intelligent Manufacturing: Worldwide "learning" 3D Printer Network”) and enable unprecedented crowdsourcing/engagement in the materials discovery process (“NSF virtual synthesis laboratory goes online”).

Such commoditizing of materials synthesis and discovery would then also enable discoveries in nearby fields (e.g. “Axionic Dark Matter Detected!”), as well as enabling broad participation in developing materials to improve batteries, chemical transformations, and more.

I.3. Community Readiness

“Over the last decade, we’ve seen major advancements in using big data to predict new materials." - Eric Toberer, Director, HDR Institute for Data Driven Dynamical Design (2021)

Accelerating materials discovery requires building connections between communities. Arguably, the four key communities involved in this process are: (1) materials synthesis; (2) materials characterization; (3) materials simulation; and (4) data science. Connections between these communities are essential, given the increasing demand for new materials - and for dramatically decreasing the timeline between their initial discovery and for bringing these materials to the marketplace. The material synthesis/characterization and materials simulation communities have seen significantly increased integration levels through the Material Genome Initiative and associated endeavors - and advancements in experimentally accessible time/size resolution, together with improved algorithm development[39-45] and increased computational facilities is now enabling almost side-by-size communication between experimental and modeling,[46-47] leading to improved communication between these two communities. With these advancements of experimental and simulation capabilities, we now have the ability to generate unprecedented amounts of material synthesis and material simulation data, demanding a connection with the fourth - and arguably the most recently established - essential community, data science.[30,48] Data science - and the development of associated Artificial Intelligence (AI)[49-50] based tools like Machine Learning (ML)[51-53] - thrives on big data, allowing its unbiased analysis, enabling us to find trends and use these to make predictions - for example for material/property relations and material synthesizability. The data science community has now developed a series of well-developed ML tools that are already beginning to impact material synthesis.[53-54] However, communication channels between data science and material synthesis and/or material simulation efforts are at a relatively immature stage - and broadening and deepening these channels is essential for reaching our future material discovery goals.

I.4. Synergistic Readiness

"A machine is not a genie, it does not work by magic... [t]o believe otherwise is either to believe in magic or to believe that the existence of man's will is an illusion and that man's actions are as mechanical as the machine's." – Arthur L. Samuel (1960)

Materials research has always thrived by drawing upon the expertise and advances in many domains. In considering: "is now the right time to make large steps towards and investment in data-rich materials discovery," the readiness of the neighboring communities required must also be considered.

AI/ML Frontiers/Questions

- How do we detect, control, and minimize bias in AI/ML models?
- How do we make AI/ML model outputs robust, especially in the face of adversarial inputs?
- How do we encourage sharing of codes and data to enable reproducibility and transparency?
- How do we extract underlying scientific principles and fundamental truths from AI/ML models?
- How do we make AI/ML techniques work well in the presence of sparse, error rich, and biased data?

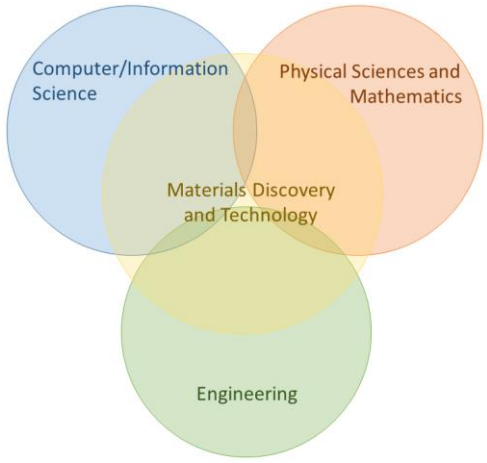
Many current challenges and frontiers in fundamental AI/ML science and engineering, such as those above, are especially relevant for the application of AI/ML to 2DILM materials research, and there are ample opportunities to make advances in both simultaneously.

Driven in no small part by the relentless climb of computing power and resources, data science has seen a meteoric rise and expansion over the past

two decades.[19,33,55-56] This has enabled the practical implementation of data analysis and learning techniques first developed (theoretically) more than sixty years ago,[57-59] now popularized under the handle "AI/ML techniques".[60-62] Driven primarily by the computer science, computing, and mathematics communities, these techniques have rapidly solved challenges once considered to be intractable – such as identifying objects in images ("computer vision"), and comprehending the meaning of the written word ("natural language processing").[63-64]

The tools and techniques behind these accomplishments are reaching maturation, and there is great excitement in the computer science, computing, and mathematics communities about the potential for these to impact new problem domains. Indeed, there are multiple national and international efforts (e.g. the National AI initiative, the National AI Research Institutes, DARPA's Explainable AI and AI Next Campaign, NISTs standards process for reliable and trustworthy AI, the Global Partnership on AI, etc.) to bring these tools to new challenges spanning the physical sciences and engineering.

There are particular synergies in bringing this community specifically to materials discovery and 2DILM. Compared to the examples mentioned above, data in materials is sparse, diverse, and error rich. At the same time, there are one hundred years of knowledge and models of how materials behave, making it a great playground for the development of new AI/ML techniques capable of handling such realities. Another strength of ML techniques is the ability to pull patterns out of data – and many of the impacts described above rely on this feature. Further, if an understanding of how and why an ML model reaches the conclusion it does can be developed (this is an open problem in data science, generally), then that understanding can be mapped to improved models of materials and materials discovery.



The juncture of materials science with engineering, technology, and computer science identified in this report has the potential to enable new synergies across fields and domains.

The potential synergies go further. All fields of science, including basic AI/ML research, struggle in timely publication of data and codes underlying studies. Much foundational materials discovery is driven by the need to find a new class of materials with a specific property – be it superconductivity, high catalytic activity for water splitting, or a myriad of others. And there are many examples of successes in this regard. But the translation to useful technology is often hampered by the fact that it is not "one property" that matters – what matters is how it fits into the whole system driving the useful functionality. At a minimum, this means that multiple properties must be considered simultaneously (e.g. dopability and thermal conductivity and lattice matching to silicon), and in many cases means that the desired property is not an attribute of a single component or single material, but an emergent property of the collection of materials and components – ie, the system. This requirement to consider multiple factors simultaneously has recently been popularized as "co-design",[65-66] but is extremely difficult due to the long feedback loops and feedback times between how discoveries happen and how technologies are developed (For example, as noted in the 2014 Materials genome initiative strategic plan, it currently takes 20 years to transfer materials innovation

to semiconductor fabrication of new devices[67]). The inclusion of AI/ML techniques, even as coarse screens in a high dimensional multi-property landscape, has the potential to greatly reduce the latency of such a feedback loop and accelerate the rate at which discoveries and technologies are iterated.

I.5. Downselecting from broad topics to the threads

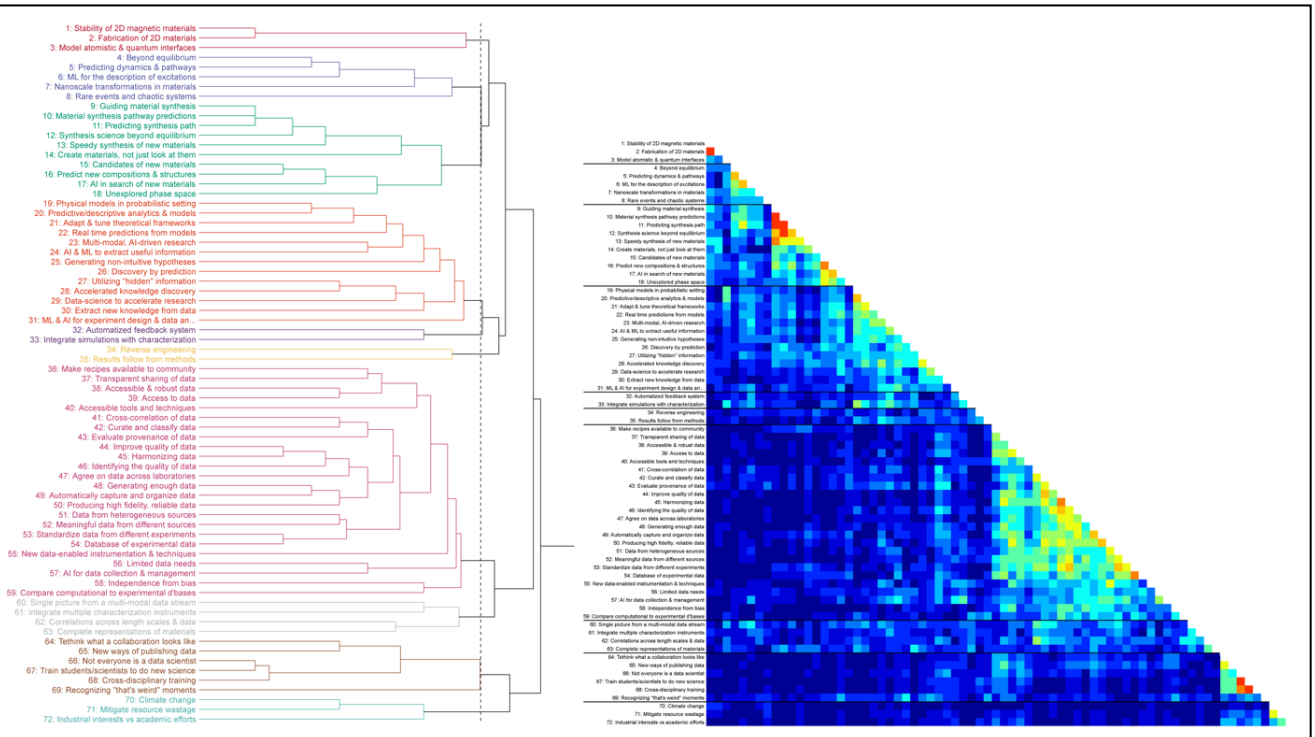
"Deciding What Not To Do Is As Important As Deciding What To Do." – Former Apple CEO Steve Jobs

Community identification of the fundamental vision and associated goals that will drive the next decade and more of 2DILM research has been in progress for several years.[5,8,10] The 2DILM forums on Envisioning Pathways to Accelerated Materials Discovery in January and March 2022 centered on leveraging exploratory phases of idea conception and community growth. The forums themselves focused on winnowing and connecting community-wide experience by identifying driving questions, gaps, and central challenges to reveal a focused, shared mission. An important component was learning from efforts already in this area. For example, in Fall 2017, an NSF-supported meeting of grantees associated with the Emerging Frontiers in Research and Innovation in 2D Atomic-Layer Research (EFRI 2-DARE), Designing Materials to Revolutionize and Engineer our Future (DMREF), and Materials Innovation Platforms (MIP) advanced communal research concepts centering on leveraging data resources in the 2D materials community.[18] This community effort led to supplemental funding of nearly 40 PIs and 25 universities and the start of broad understanding of the need for coordinated work centered on optimizing data to fuel burgeoning techniques.

A 2020 NSF Request for Information (RFI) aggregated PI experiences from the 15 supplements spanning experimental, computational, and data-driven approaches across the processing-structure-properties spectrum of 2DLIM research. Collectively, respondents reported on: Computational efforts including improved tools to predict structure and electronic/optical/magnetic properties and interpret and process experimental data; Experimental efforts including generation of material sample libraries and validation/guidance of computational efforts; and data archiving and machine learning activities including screening of candidate materials and guided materials processing. The activities funded by these supplements provided experience with close coupling of experimental and computational teams to create outcomes beyond the parent EFRI, DMREF, MIP, and PREM activities. RFI responses emphasize perspectives on maximizing student training in data-science knowledge and skills; barriers to carrying out data-intensive research; characteristics of successful 2D data infrastructure; and activities to promote continued development of 2D data infrastructure. Recurring themes in the responses highlight the need for sustained motivation and investment in:

- infrastructure to curate physical samples
- incentives for curation of experimental and computational data
- closing critical gaps in functional metadata through use cases, reference datasets, and in coordination with major materials data repositories
- connecting and collaborating from data creation through data use and reuse to bridge disparities in experimental and computational data and clarify metadata development
- broadening community input to shared data resources and repositories
- interoperability between the large and growing number of data resources at universities, national laboratories, and in industry
- fundamental advances in AI/ML that incorporates fundamental physics to meet the challenges of small or specialized datasets and yield statistically significant results in materials science
- diverse workforce development and cultural change including broad adoption of advances in training and research methods/workflows

These themes are not unique, of course, to 2DILM research, but can provide important benefit to the wider materials research community in both substance and as an exemplar of methods and best practices. This constellation of funded projects and facilities that kickstarted data-intensive work also



Preliminary dendrogram (left) and heatmap (right) visualizations of questions provided by forum applicants used for sorting and to guide organization of working groups at the January forum. Note that the initial sorting identified 10 groupings as shown in the color coding of the dendrogram. Refinement of these groups by the core organizing team combined smaller groups to create the six, central themes of the forum.

created a shared understanding of critical gaps and prepared a foundation to build readiness to lead in a sustained community effort to close them.

To capitalize on this foundation, the PARADIM and 2DCC MIPs invited 490 researchers from the broader 2DILM community to apply to participate in the January 2022 forum, for which ~96 responses were received. The scope of this community included researchers with active or past awards in 2DILM research; prior 2DCC and PARADIM users and summer school/workshop participants; and data science/AI experts with collaborations or other ties to materials research. As defined, this community ranged from early-to-senior career stages; positions in national laboratories, industry, and academia (R1, non-R1, MSI); and deriving research support from the NSF, DOE, DOD, and non-profit foundations. The application process was used to help select a diverse cross section of the community as well as to identify the scope and core challenges of interest across the community. The applicant questionnaire specifically elicited views of the most interesting questions that might be addressed in 2DILM materials in the next 10-15 years.

Responses to the applicant questionnaire generated 168 questions/challenges (Table C1) which the core team refined to 90 based on topical scope and uniqueness. These 90 questions were sorted using dendrogram, tree map and heatmap visualizations (above) to assist review and refinement by the core team and steering committee. This process identified six groupings of the community-created questions and provided a basis for working groups at the forum. Importantly, all questions were used whether the author was selected for participation in the forum and the team refrained from naming or evaluating the groupings beyond that done in the information provided by applicants and thereby avoiding biasing community discussions during forum working groups. The identified groups covered (Table C3):

- Teaching, training, and trusting data science for materials discovery
- Democratization of data, data tools access, and archiving
- Interconnection between data and knowledge
- AI-enabled digital twins created through unification of experimental and theoretical data
- Establishment of autonomous inorganic synthesis science with secure synthesis and discovery
- Translation of materials and data from discovery to technology and industry

While not named as such, these themes revealed concern and interest in the widely recognized, foundational challenges found in most visions of data- and AI-driven futures in science and engineering. These themes are also aligned with those reported in the 2020 RFI and serve to highlight the breadth of community consensus and persistent nature of these foci. Specifically, the identified themes provided focus on: issues of understanding and implementing FAIR data principles; the opportunities and value of leveraging shared data resources; development of materials-specific applications and extensions of AI; automation and high-throughput approaches to accelerate the translation of data to knowledge to industrial applications; creating and delivering sustained workforce development across all age, experience, and privilege sectors; and the broader impacts of the work within and beyond the 2DILM domain. Over the second half of the January forum, participants worked in small groups to find conceptual connections and create four focused, high-priority challenges encompassing the central themes of the springboard questions as well as new questions that arose during discussions. These community-scale challenges encompass innovation and novelty that is critical to meeting the community's shared future vision.

Specifically, how might we (Table C5):

- Better predict outcomes from actual synthesis conditions?
- Achieve AI-assisted and autonomous synthesis?
- Promote knowledge sharing?
- Prepare practitioners to better use data to accelerate discovery?

In March, a second forum convened a focused group of participants with expertise to dissect these challenges and set specific goals to create a roadmap. While specific, these are *goals* and *not solutions*. Each goal will require community-scale work that benefits from coordinated, but diverse approaches. This process of engaging, winnowing, and refining the community's views has created the elements of a roadmap. This roadmap has a shared destination, agreed upon challenges, and specific goals to act as milestones that define the path; it does not profess clairvoyance to specify how to implement roadmap.

II. Priority Research Opportunities

II.1. Predict Outcomes of Actual Synthesis Conditions

"How might we use data to guide and predict materials synthesis?" – Forum Participant

II.1.1. Background

Synthesis is often the rate limiting step of materials discovery. Phase diagrams provide a road map for synthesis efforts by describing the range of conditions within which a desired compound is thermodynamically stable. Incorrect phase diagrams pollute and inhibit new materials discovery; there is a need for internally consistent free energy data. Most large-scale first principles databases (AFLOW,[68] OQMD,[69] Materials Project[70]) provide only phase diagrams (convex hull) at T=0 K with no strain. This is a mismatch with the needs for the community which performs experiments at high temperature and typically in the presence of interfaces. Also, there are numerous experimental techniques – including Molecular Beam Epitaxy, Chemical Vapor Deposition, Atomic Layer Deposition, for vacuum-based techniques and other methods in liquid environments – which each present different challenges and environments. Accurate free energies—at the temperatures and other thermodynamic conditions at which synthesis is performed—are needed. From the free energies, synthesis and processing conditions providing thermodynamic stability can be mapped out, phase diagrams can be calculated, reactions at interfaces or with crucibles or containers can be assessed. Unfortunately, critically assessed and internally consistent experimental free energies exist for only a limited number of inorganic materials (the approximately 4,300 entries in the SGSUB database[71]) exist. This number pales in comparison to the number of phases that are known, e.g., the over 250,000 crystal structures in the ICSD database.[72] In addition, there is limited information on kinetic processes such as diffusion and reaction which are needed to describe non-equilibrium conditions.

Part of the substantial complexity arises from the sheer variety and number of different techniques available for the synthesis of materials, that together span vast ranges of thermodynamic and kinetic

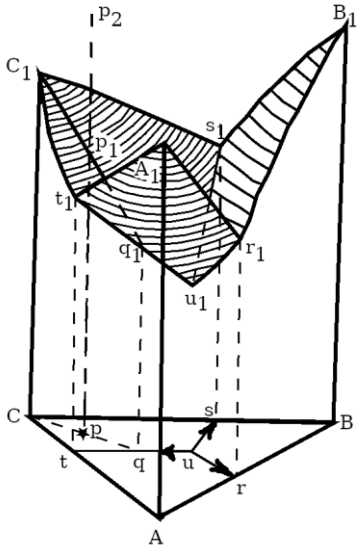
parameter space – and often with only indirect control over relevant thermodynamic parameters. Numerous books and reviews on this plethora of experimental approaches – and current theoretical models used to simulate them – exist.[73-81] What this means in practice is that even with 250,000 crystal structures in ICSD, the amount of available experimental data is extremely limited and sparse compared to the size of the parameter space.

On top of these general issues related to predicting material synthesis, specific additional challenges exist for 2D-materials. Phase diagrams to guide synthesis of complex oxide, nitride, carbide and chalcogenide materials are often less developed that for alloys and the calculations are more complex because of constraints of several sublattices and ionic substitutions. Many such materials are made by relatively low temperature precursor routes, for which the interplay of thermodynamics and kinetics is important. 2D and layered materials often are stabilized or even made possible by the simultaneous existence of both order and disorder - ordered structural domains separated by cation disordered regions. Structure, thermodynamics, and synthesis may then be very temperature and precursor dependent. Synthesis often follows an energy landscape of intermediates of increasing stability. These can be treated both in terms of thermodynamic pathways and kinetic constraints. An additional challenge for theory – in particular Quantum Mechanics (QM) based methods - includes the large unit cells and complex compositions of layered and 2D materials. The complexity of these materials essentially necessitates theory tools that can scale up – both size and time-scale wise – from QM-based methods, for example pseudo-ab initio methods like Tight-Binding DFT[82-83] or physics based or machine-learning based interatomic potentials (IPs).[84-89]

Traditional phase diagrams also do not consider defect structures – including vacancies, grain boundaries and surfaces, which can play key roles in material synthesis as either intermediate phases or even as kinetically stable end structures. Fortunately, theory methods at various scales, including quantum mechanics-based methods and IPs, exist at the atomistic scale and phase field[90] and finite-element[91] based methods on the meso- to macroscale, have now reached a level of maturity and accuracy that they can, in principle, provide reaction barriers for material growth and defect structure energies and migration barriers. Communication across the time- scales covered by these theory methods is still not trivial, but the dramatic improvement of experimental size- and time resolution combined with the ever-increasing size- and timescales accessible by theory methods[44,92] continues to decrease the gap between theory and experiment, thus allowing more direct validation of theory predictions and more direct testing of theory improvements. Machine learning (ML) methods can play a key role in further closing this theory/experiment gap – for example by analyzing experimental data for grain boundary movement and vacancy clustering and their relation to temperature and gas-phase composition- thus providing clear targets for theory validation. Furthermore, with theory methods expanding to ever-larger systems (>> 1,000,000 atoms for IPs) human-based analysis of IP based molecular dynamics trajectories becomes virtually impossible; here advanced pattern-recognition ML tools[93-94] can be essential to harvest key observables.

II.1.2. Opportunities

Phase Diagrams: An Opportunity?



As a representation of the stable species as a function of thermodynamic parameters, phase diagrams are indispensable both for designing reactions to make desired materials, and to put boundary conditions on what compounds are possible to produce. Experimentally tabulated[95-96] diagrams are essential in designing reactions for materials synthesis and discovery, but do not exist for some binary, most ternary, and virtually all quaternary or higher combinations of elements. The fundamental principles are known, and computationally predicted[70,97-98] phase diagrams of multinary mixtures can be produced on demand, but are often found to lack the necessary precision to be useful in a practical synthetic sense. A regular, community-driven, impartial and critical assessment of phase diagram prediction has the potential to produce transformational improvements in predictive power and acceleration of materials synthesis. This requires significant and sustained investment in the creation of reproducible and curated phase diagrams, community consensus on impartial assessment metrics and approaches, and the sharing of data, tools, and techniques.

Reliable phase diagrams and expanded kinetic data will accelerate materials discovery. Text mining and collection of phase stability data from the literature in concert with an automated method of critically assessing this data by using it in combination with the entries of a starting database (e.g., SGSUB) and experimental or theory based phase stability data to build a larger critically assessed database of free energies is essential to making progress. A community driven, periodic, critical and impartial assessment of phase diagram predictions has the potential to lead to transformative improvements in quality, as has happened over the past decade in the challenge problem of protein folding. Similar fusions of reproducible, curated experimental data with unbiased and critical assessments of predictive tools are anticipated to be useful in advancing our ability to make successful structure predictions – for a given set of chemical formulas, what structures exist?

II.1.3. Needed Research

Systematic creation of new experimental phase diagrams for immediate use in guiding synthesis and for critically evaluating theoretical model performance is essential. This cannot be the job of a single institution, but must be driven by a community of practitioners. Improving theoretical tools to obtain chemical accuracy in calculated phase diagrams – as assessed by impartial, periodic, community-driven measurements, is critical. As with the recent advances demonstrated by AlphaFold in protein structure prediction, it is the fusion of these components – with new advances in how to incorporate the extensive body of existing domain specific knowledge with AI/ML training techniques to improve prediction accuracy. The materials community should strive to develop a similar impartial evaluation of the many theoretical tools that have been developed, are currently in development, and will be developed in the future.

Equally important is to develop the conceptual formalisms and frameworks by which to include reaction kinetics on an equal footing to thermodynamic variables in the prediction of reaction products and reaction pathways. This must be tightly coupled to experimental data probing such kinetics, in a way that is transparent, reproducible, and useable by the entire community. This includes the development of approaches to study reaction dynamics *in situ*, and to identify which variables are important in dictating reaction outcomes. Given there must exist a mapping between synthetic conditions and the result, the development of AI/ML techniques leveraging such emerging database to make such reactivity predictions is invaluable. The scientific impact would be further enhanced if the guiding principles / fundamental truths of the universe picked up by accurate AI/ML models can be extracted from them (this being a grand challenge in the AI/ML space more generally).

To truly realize predictive synthesis, it is also necessary to be able to determine the structures that will exist – or might exist if approached via the right reaction pathway. As with phase diagrams and the inclusion of kinetics, it is our view that having transparent and reproducible data, combined with sustained efforts in critically evaluating model predictions by comparison to high quality experimental results, is key.

In addition, we need to develop new publication and credit mechanisms associated with phase diagram generation – either from experiment or from theory. Given the substantial effort involved in evaluating a complete phase diagram – especially for multicomponent materials, we need to identify mechanisms that allow for the addition of partial phase diagram data to existing databases and develop better tools for data sharing – both for positive results (e.g. theory/experiment match) or negative results – since both these results may be scientifically equally valid and as such are equally valuable from a data science perspective. This also includes making sure appropriate recognition and credit goes to those who invest the time and resources to do so. Here, the approach taken by the protein folding community – encouraging experimental determination of new structures to provide unbiased, critical tests of predictive tools, as part of a regular whole-community critical assessment, is one pathway to explore.

II.2 Achieve AI-assisted and autonomous synthesis

"How might we discover metastable and transient materials and states with remarkable properties that we don't even know about, nor find in databases?" – Forum Participant

II.2.1. Background

Synthesis is notorious for its irreproducibility. Changes in the personnel performing the work, the synthesis chamber used, or the location of the synthesis often yields distinct results. This extends to scaled manufacturing, where reproducibility is only obtained due to extremely tight controls on all aspects of the materials production pipeline and the investment of significant resources to optimize the synthesis process for the given location, tools, suppliers, etc., though even there seemingly inconsequential variations can cause catastrophic disruptions.[99] This is particularly detrimental to progress in materials research, as follow up work on promising findings is often bottlenecked on the ability to obtain reproducible specimens. Or, put another way, the general lack of reproducibility in

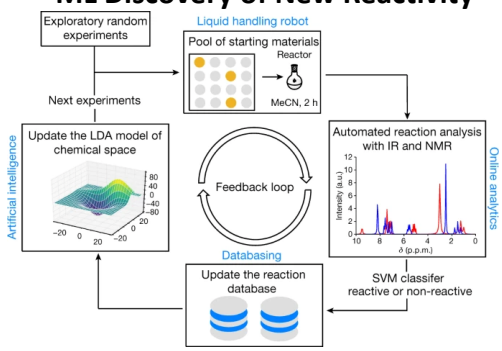
synthesis without significant per-individual and per-location investment in parameter changes and tuning, limits the rate of technologically useful materials discovery.

At the same time, the last twenty years have seen significant advances in in-situ characterization and synthesis tooling, and in digital control and partial automation of process variables with in situ feedback. In many respects, then, the pieces are ready to use AI/ML to advance synthesis and synthesis science. What is missing is the "glue," the methods to close the data-to-growth loop within the tool, matching timescales of growth and modeling plus computation/ML to interpret and act upon the data in real time. This also sets the stage for a transistor-like explosion in available materials synthesis data, as it lowers the cost per reproducible experiment (much like the computing industry lowers cost per bit/transistor).

II.2.2. Opportunities

There are many parameters that enter into synthesis, from the provenance and chemical details of reagents and their processing and use, to the nominal versus actual thermodynamic and kinetics conditions of growth, and their variability with time and space. It is the combination of this large number of known parameters (plus unknown ones) combined with a lack of knowledge of which are most important and how to control them with sufficient precision that leads to the irreproducibility. Translated into the language of data science, this means that synthesis operates in a highly multidimensional space in which the mapping between inputs and outputs is non-obvious, and in which even the best in situ measurements generally do not directly map onto a single or small number of process control variables. Determining such mappings (both between inputs and outputs, and between in situ measurements and outputs) is exactly the type of problem at which AI/ML techniques excel! And, if combined with advances in understanding how ML models are reaching predictions (see I.4), also open up the ability to use these models to reach a chemical understanding of what variables are important and why.

There is thus substantial opportunity in accomplishing the merger of AI/ML techniques and synthesis in an automated (or semi-automated) way. The development of autonomous synthesis would accelerate materials research by orders of magnitude, similar to the impact of the steam engine on factories or the effect of computers on mathematics. Research will be more dynamic as robotic and AI-based processing can facilitate quick changes in direction. Continuity and reproducibility will be dramatically increased as protocols and approaches for materials synthesis and characterization will be institutionalized rather than residing with specific individuals within an organization. Finally, students and researchers will be freed from menial and potentially hazardous tasks enabling them to focus their synthesis talent on the crux of the synthesis bottleneck. Open access to autonomous synthesis tools via national facilities will increase the number of researchers who can contribute to advancements and engage citizen scientists. All of the above may sound too good to be true – but the events of the COVID-19 pandemic provide a window into what is possible. National user facilities with robust automation and remote access mechanisms in place (e.g. NSLS II, the Advanced Photon Source, PARADIM, etc.), saw greater sustained productivity than those without such access mechanisms (e.g. NIST NanoFab). Similar anecdotal stories exist for work within individual laboratories.



Granda, et al. reported in 2018 the use of an ML guided organic synthesis robot to discover four new chemical transformations.[100] This achievement was enabled by combining robotic synthesis with domain appropriate in situ characterization and a novel twist on ML to enable predictive accuracy with relatively few input data (by leveraging known retrosynthetic tools), and has spawned whole new directions in organic synthesis research.[101] Robotic synthesis tools have begun to be developed in the 2DILM space, and in situ characterization approaches are rapidly developing. Does sufficient domain specific knowledge of materials synthesis exist to enable a similar watershed achievement in 2DILM?

II.2.3. Needed Research

Historically, the costs associated with equipment, sensors, computing hardware and software, and engineering effort to automate processes has limited investment in automation to production lines and other situations in which a limited number of well-defined products are synthesized in large volumes. In contrast, materials discovery relies on accessing new regions of parameter space, and many impactful discoveries are the result of serendipity (i.e. something changing in a fortuitous way *and* someone noticing the change). AI/ML guided discovery thus requires the redesigning/engineering of synthesis and characterization tooling to enable human/digital/robotic interaction while retaining the flexibility to access (and, importantly, know you have accessed) new or tailored synthetic conditions. Further, it is crucial that the time scale of the in situ real-time feedback from AI algorithms be on the same time scale as the synthesis task.[102] While fully autonomous AI-based materials discovery likely requires yet-to-be-made advances in data/computer science (see section I.4), success with the more tractable (but still hard) task of harnessing AI and ML to capture the actual conditions of fortuitous reactions and enable their reproducibility is a natural steppingstone.

Achieving AI-assisted and autonomous synthesis can be enabled through:

- Development of protocols and procedures for automatic centralized gathering of data from existing synthesis and characterization tools into integrated databases with metadata and formats that are amenable to ML and other data science techniques (see community opportunity III.2), as is being done in some limited cases already (e.g. BlueSky from NSLS-II, NIST's NexusLIMS).
- Identification of specific materials discovery problems and synthesis processes where ML and AI can aid to inform human intelligence to accelerate synthesis research – i.e. identify specific "challenge" problems on which to develop the techniques.
- Collaboration between vendors and researchers to design plug-and-play equipment to enable synthesis and characterization tools that exchange physical samples as well as data with minimal human intervention, and lower the cost to that of existing synthesis equipment today.
- Further development of in situ characterization tools that provide rapid feedback on material properties – this means not just providing the data, but having the data and analysis sufficiently integrated and automated to extract meaningful, actionable information – i.e. what does the data mean should be tweaked?
- Development of knowledge extraction approaches that illuminate "why" ML models are making the predictions/decisions they are, so we are learning "why" and not just "how to"

A combination of steps across multiple timescales and lengthscales is needed to meet these needs. A central theme among these is the need to collaborate closely with tool designers and vendors to provide the capabilities needed, but to do so in a way that ensures open access to data and APIs. Here, lessons from aligned fields that have already largely made this leap are useful. For example, automated protein and DNA synthesis machines exist because the modular chemistry enabling them was developed, and it was recognized by the individuals (and their institutions) that selling such tools not only made economic sense, but would enable a much larger community to adopt the techniques than would otherwise have been possible. Following much the same flow, automated stacking robots exist in the laboratory and are being commercialized and will lead to radical improvements in both reproducibility and accessibility of tailored 2D heterostructures. Older examples also exist – e.g. from laboratory to commercial optical floating zone furnaces. Two ingredients appear key: (i) that there be a market for the tools, and (ii) that the incentives be aligned with broad dissemination of the technology rather than reserving as a competitive advantage.

The size of the market is more or less set by the generality of the tool (more general = larger market) and the willingness of stakeholders (across academia, government, and industry) to support acquisition of tooling. Alignment of incentives comes from the gain from decentralization being larger than that from holding onto the technology. Addressing both of these requires hard conversations amongst all stakeholders, and hard decisions on directions in which to focus effort and resources, with preference given to those that are simultaneously unique and generalizable. For example, we believe that vendors who do not provide easy and open mechanisms to workflow data in open and known formats should not be selected for new equipment designs and installations. Further, there should be an expectation of broad dissemination go hand in hand with resources enabling the development of automated synthesis tooling.

III. Community Opportunities

III.1 Develop the Data Materials Science Workforce

"How might we better train students/scientists to do new science in the context of all the large data processing tools and techniques?" – Forum Participant

III.1.1. Background

Training beyond the high school level is traditionally closely tied to specific areas/fields of study – be it economics, chemistry, or materials science. This specialization is important to have a suitably deep body of knowledge on which to draw in a career. It does, however, result in a natural barrier to the incorporation of new and important areas of knowledge in existing curricula as the new area is often viewed as augmentative to, rather than a core part of, the discipline. Even in cases where the gain is obvious – see for example the permeation of computers into nearly every area over the last 50 years – residual vestiges of this barrier persist, e.g. many chemistry departments do not require any computer/computing training at all and instead expect the knowledge to be gained independently or

by osmosis. The priority research opportunities described in this document require a materials science and engineering and computer science workforce that share a common language in order to execute on these opportunities and turn them into real impacts. It is thus imperative that methods and approaches towards growing a data materials science workforce will be developed.

One perceived challenge is that all of these fields are very rapidly evolving. As such, what is most important then in the workforce is adaptability and ability to keep up with advances. This is not a unique problem for materials, and really requires that the training provide grounding strong foundation in the things that don't change (the fundamentals) – be it models of computing or the laws of physics.

III.1.2. Opportunities

As a naturally interdisciplinary field, materials science and engineering enjoys lower barriers than many to the incorporation of advances from related fields. This extends to education and related approaches to prepare a data materials science workforce (see, e.g., the European Union EDISON project[103]). Further, there is presently significant demand within the diverse student bodies of chemistry, physics, and materials science for data science training. Additionally, the challenge of finding, recruiting, and retaining materials scientists with data science capability has been well noted as a key limiter of process in various industries. As such, there is demand on all sides, and thus opportunity to be successful.

Equally important to individual education are the benefits that come from communal agreements on how to frame materials science and data science problems. Standardization of data, metadata, and interfaces will spur innovations in equipment design allowing their wider adoption. Platforms that provide access to data, tools and computational resources will lower the barriers to their use by practitioners. Improved educational frameworks will increase the knowledge transfer from basic science and engineering innovations to practice. A diverse cadre of next generation material scientists and practitioners will be developed who are able to leverage the power and potential of data and AI to accelerate science. Since innovations in materials are core to technological advances in many areas of national and societal priority, accelerating material discovery will positively impact virtually all areas of human endeavor.

III.1.3. Needed Community Actions

Achieving these goals will require the development of new education and training opportunities and standardization of data/metadata and interfaces so they can be combined to form reusable shareable workflows. Material scientist and engineers must be trained and encouraged to communicate with data scientists (and vice versa) to collaborate effectively to address previously unsolved material science problems. This includes not just at a technical level, but at a "human" level – i.e. communication, collaboration, and teaming skills. Platforms for collaboration must be created that include data sets, analysis tools and computational resources along with the development of benchmarks and challenge problems to focus community efforts. Curricula and training materials must be developed for different audiences (faculty, users, developers...) and skill levels. Best practices should be shared regarding running, deployment, and sustainability of cyber-physical infrastructures

for education and research. Broader engagement with the AI/data science/cyberinfrastructure communities will be required along with engagement with other communities that have successfully integrated data science within a domain (e.g. astronomy, genomics, etc.).

Steps the 2DILM community can take to make progress on this front:

- Hybrid team science should be encouraged whereby data scientists and material design engineers are embedded with materials synthesis and characterization scientists in research. This should be a long-term engagement – where the data scientist and the material scientist are expected to develop a common understanding. It has to be clear that this common understanding is the main goal – and that a material discovery target is desired, but secondary. We want to clearly understand what steps were required for these communities to develop a synergistic understanding – this does not necessarily mean that the data scientist becomes a material expert – or vice versa – but that they have enough common understanding to benefit from each other's expertise – creating a team that is more than the sum of its parts.
- Based on successes and challenges in team science, identify the training needs of various communities (graduate students, faculty, ML tool developers, etc.), and develop sample curricula for different training needs.
- Identify and develop a range of training delivery mechanisms – online training, short courses, summer institutes, badges, certificates, minor degree programs, etc.
- Identify opportunities to integrate data science into materials curricula so that it is indistinguishable from the core discipline.
- Build transdisciplinary efforts. The workforce should represent a new area that spans Chemistry, Physics, Materials Science, and Engineering disciplines. These labels should perhaps even disappear -- individual researchers bring diverse skills and expertise should not be "pidgeon-holed" or "siloed" into identifying with one or a couple of these (limited) areas.

III.2 Enable Data Curation and Community Use

"How might we move from the typical scientific publishing paradigm of papers data available on reasonable request to a truly transparent way of sharing scientific data?" – Forum Participant

III.2.1. Background

A defining issue in accelerating materials discovery and deployment is functional exchange of data to fuel critically enabling, data-driven approaches including AI/ML. Data, including experimental and computational results; codes and algorithms; and the trained ML models themselves underpins modern science and catalyzes deeper collaboration. Durable availability of meaningfully annotated data extends the value of data beyond the questions that data was collected to answer. Appropriately curated and stewarded data increases the value of data over time to create *lasting* impact and the power to address large-scale challenges required to meet the goals of the 2DILM roadmap.

The Power of Shared Data

Shared data through the FAIR data principles is often noted for the potential to transform the scientific endeavor.[104-105] Shared data is now the basis for screen potential candidates for many kinds of properties and behaviors.[106] Recently, Beck et al.[107] showed the value of reusing data computed by DFT to extend high-throughput screening of catalysts with ML to learn the most important surface features of catalysts, with less mean error than individual DFT calculations themselves.

Shared data also plays a central role in trustworthy and open science. A bittersweet example comes in the form of a recent paper retraction prompted by data curation done after publication rather than during the process of analysis and review prior to release of results.[108] In the retraction, the authors revealed "When preparing the underlying data for public release, it was discovered that some data had been inappropriately deleted or cropped when preparing the final published figures…". This example highlights both the need for open best-practices for materials data and the community challenge of potential impacts on scientific reputation.

Indeed, in all fields data is now seen as one of, if not *the most* valuable resources in the world today[109-111] specifically because it enables transformative leaps in understanding and predicting behaviors in complex systems. The ability to reuse data to address a multitude of questions hinges, however, on sharing data with high levels of interoperability and deep levels of semantic understanding. Such shared reuse is central to optimizing the value of data that are expensive or difficult to obtain; it is central to optimizing the use of shared resources that create the most critical data such as beamlines and cutting-edge instrumentation; it is central to reuse of data to provide reproducibility and transparency critical to advancing the field. Meaningful, shared reuse of data is central, therefore, to leveraging existing and future investments in instrumentation, facilities, training, and researcher effort. As such, investment in data infrastructure and systems is a priority output of this forum.

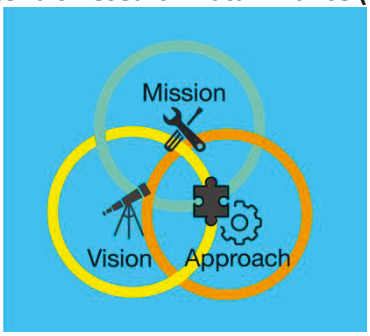
Data curation, and enabling its use by a larger group, is also seen as a foundational priority for partnerships that accelerate translation of research knowledge and insight from lab to market and society. It is clear that in the science and in the translation, data itself is becoming a primary research output and central component of knowledge sharing. In the materials domain this importance is an interagency priority called out in the MGI 2.0 strategic plan and recognized across all subfields.[34,104,112-115] Functional data policy, infrastructure, and culture is at the heart of the FAIR principles for findable, accessible, interoperable, and reusable data and why FAIR data is the key conduit to knowledge discovery and innovation.[116-119] A recurring theme in the 2DILM forums was the challenge of defining and implementing FAIR principles in the materials domain. While the concepts are simple the challenges of implementation, both technical and cultural, are not. Indeed, since the codification of FAIR principles in 2016, a multitude of science-driven initiatives in the materials domain has sought to build infrastructure

III.2.2. Opportunities

Two main conclusions emerged during forum discussions. First, there is broad recognition of the need for sustained, community work to establish and evolve metadata models and standards that allow FAIR implementations and associated infrastructure development. There were many suggestions of where such standards could arise including as dictates from publishers, funders, central standards organizations such ISO, or standard services providers such as NIST. Push-back made it clear that

centralized dictates of metadata rarely work and are unlikely to encapsulate the ever-evolving needs of a dynamic, global research community. Discussions around these competing perspectives make it clear that the path forward needs to prioritize metadata development as a community effort with sustained leadership that connects expertise and stakeholders across the data landscape from data producers to its many potential users as well as throughout the data lifecycle[120] to produce practical solutions and lower barriers to participation or compliance. The second main conclusion was that incentivization to create and use data FAIR will be complex but is central to success. Current state-of-the-art for incentivization focuses on reward systems and compliance enforcement. Reward systems include access to funding opportunities and development of data citation metrics to reflect the impact of data releases. Compliance mostly consists of requiring of data management plans (DMPs) and withholding or loss of funding for failing to meet unclear metrics or measures of FAIR. This is despite the broader data community's repudiation of FAIR metrics as useful.[121] In overview, forum discussions highlight that this rather traditional view of incentives comprising carrots and sticks remains a central problem, perhaps *the* central problem in FAIR gaining traction in materials science. Resistance to FAIR reflects that and centers mainly on the cost of broad implementation of FAIR and a cultural resistance to sharing one's data. Investigators and funding agencies both have resistance to taking on the cost of FAIR and concerns about how it cuts into their other priorities or number of projects that can be undertaken. Common materials investigator culture has viewed data a private domain where sharing is undesirable because of opportunities for inappropriate use, inadequate compensation, or simply loss of competitive advantage gained by having proprietary access to data. Development of controlled access tools and mechanisms which support embargoed, non-public collaboration are essential to opening the spigot on this untapped latent data resource as a first step towards greater data transparency.

The Materials Research Data Alliance (MaRDA)



MaRDA is community network of people and projects that are developing the open, accessible, and interoperable materials data and infrastructure needed to fuel materials discovery. The MGI strategic plan called out MaRDA as an emerging national network to accelerate implementation and adoption of materials innovation infrastructure through convergence of ideas, people, data, and tools. MaRDA has grown rapidly with over 285 people attending and participating in its 2022 Annual Meeting. By bringing together diverse materials data stakeholders, MaRDA provides a forum and framework to build the social and technological bridges needed for open sharing and reuse of data. A central finding of the forums was identification of shared vision; this focus on "why" links the community and incentivizes coordinated work. By connecting the community, MaRDA provides a framework to put the "who" into the equation. MaRDA is unique in the materials domain in democratizing opportunities for leadership and the deepening of community connection. Combining leadership and consensus around the challenging issues of data sharing, culture, and use will be central to leveraging the shared 2DILM vision and roadmap.

III.2.3. Needed Community Actions

These compounding issues of resistance and culture bring us to the conclusion that a future built on FAIR data will require moving beyond carrots and sticks to incentivize and motivate the community. Cultures don't change by mandate, but by community action driven by common purpose.[122] The MGI strategic plan recognized the importance of building community for this reason and has called for development of a National Materials Data Network such as the growing Materials Research Data Alliance.[25,31] A central outcome of the forum was emergence of a central, shared future vision for 2DILM materials. This vision provides a shared sense of purpose and opportunity. Embodied in the roadmap to reach that vision is a new incentive structure with goals that require FAIR data and community-level effort. Such aligned purpose removes resistance to culture change and positions the 2DILM community with the aligned incentives and motivations needed to inherently value the work they can define together to create data resources they need and cannot create alone.

In considering practical aspects of implementing FAIR as a community rather than as individual research groups or collaborations, several points emerged. First, experience over even the few years since the data supplements show that FAIR is not a technical problem awaiting resources, FAIR must be developed in concert with meaningful research applications from across the data lifecycle, and FAIR is not a static benchmark that can be solved and checked off. As applications and science changes demands for data discoverability, access, and interoperability will evolve. Metadata standards will evolve as uses of data change and new information is required. New methods of modeling and characterization will develop and interoperability will be challenged as real-time and autonomous concepts grow. Solutions developing FAIR data are research in their own right but occupy the unusual position of requiring development in parallel to implementation and use. It is anticipated that concepts of data stewardship will evolve and become more common either embedded in research groups or as part of the skillset of investigators throughout projects. Most importantly, data resources across projects scales and data lifecycles will require sustained efforts. On-going community efforts will be required to frame and support those efforts for the involving investigator and stakeholder landscape. Such community efforts are likely best built from focused communities such as 2DILM and provide an opportunity to focus FAIR development to avoid outsized consumption of efforts and resources.

We imagine that an immediate near-term step would be for the community to agree on a set of expectations for what data should be made available, and in what form, alongside publications. The co-chairs of this report all strongly agree that, at a minimum, the data contained in the figures and tables of a manuscript should also be provided in a machine-readable form, at a URL or DOI that does not require contacting the corresponding author to access, with an explicit license. Since these are the data upon which scientific conclusions are based, it is reasonable to expect that such data has already been curated in the process of writing up the results. Lack of facilities by which to publish these artifacts is not an excuse -- multiple platforms (including Zenodo,[123] the NIST MDF,[124] GitHub,[125] many university library services,[126-128] and 2DCCs[129] and PARADIMs[130] public data efforts) exist to enable public release of such data artifacts. Concerns over mis-use by others do not apply, since these are all data already present in the manuscript. In other cases where this has been done (e.g. crystal structures and ICSD and CSD), the quality of scientific output has increased dramatically. Publications already become public domain now in many countries. Will the community rise to the challenge and achieve the same for data?

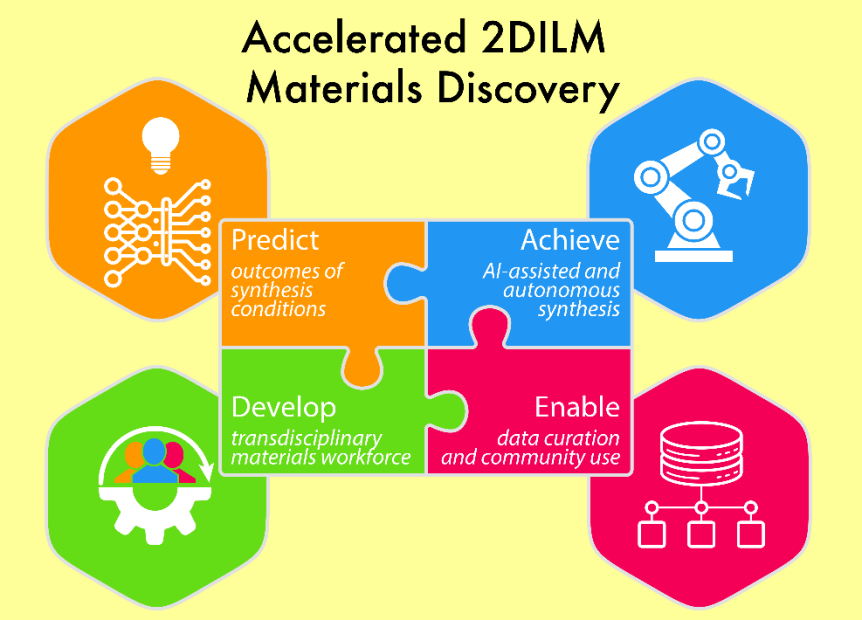
Growth of such data archives will also provide incentive and opportunity to develop interoperability required to fuel reuse to derive full value from the resource. In time, we envision standardization of formats and contextual metadata as well as new methods and programs to take advantage of the rich data contained therein. Publication of data, however, is the key first step.

IV. Future Considerations

"It is tough to make predictions, especially about the future." – Yogi Berra [attributed to many, and true]

This report focuses on concrete actions and immediate directions to realize the potential of the convergence of data science with materials discovery and 2D interfacial and layered materials. There were many interesting threads reaching beyond materials science that were identified through the forum process, a number of which are likely deserving of future fleshing out independently in their own right. It would, however, be foolish of us to predict which will resonate most strongly as the community evolves, and thus encourage all to review the questions and challenges given in Appendix C and find what resonates with them.

V. Conclusions



The extended community engagement established research towards the prediction of synthesis outcomes and the achieving of AI-assisted and autonomous synthesis as having the utmost promise and importance in advancing 2DILM synthesis. Equally important are community actions -- the need to advance broad application of best practices and principles in data curation and community use and the need for sustained workforce development that spans materials disciplines and enable both develop and application of burgeoning data-centric approaches. Indeed, the interconnection of these four identified themes of synthesis prediction, autonomy, workforce development, and data

stewardship are mileposts of a community roadmap to transform materials research and provide significant impacts on critical societal sectors including energy, health care, communications, manufacturing, and space technology. Our own perspective is that while data has always placed a central role in materials research and materials discovery, the tools of data science are resulting in a revolution similar to what occurred when computing technology became ubiquitous. Now is the time to capitalize on it and bring about a new era in materials discovery.

VI. References

1. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S.A.A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, Mi. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, "Highly accurate protein structure prediction with AlphaFold," *Nature* **596**, 583-589 (2021).
2. B. Anasori, M.R. Lukaskaya, Y. Gogotsi, "2D metal carbides and nitrides (MXenes) for energy storage," *Nature Reviews Materials* **2**, 16098 (2017).
3. B. Ryu, L. Wang, H. Pu, M.K.Y. Chan, J. Chen, "Understanding, discovery, and synthesis of 2D materials enabled by machine learning," *Chem. Soc. Rev.* **51**, 1899-1925 (2022).
4. J. Cheng, C. Wang, X. Zou, L. Liao, "Recent Advances in Optoelectronic Devices Based on 2D Materials and Their Heterostructures," *Adv. Optical Mater.* **7**, 1800441 (2019).
5. M.B. Tahir, U. Fatima, "Recent trends and emerging challenges in two-dimensional materials for energy harvesting and storage applications," *Energy Storage* **4**, e244 (2022).
6. M. Sebek, A. Elbana, A. Nemati, J. Pan, Z.X. Shen, M. Hong, X. Su, N.T.K. Thanh, J. Teng, "Hybrid Plasmonics and Two-Dimensional Materials: Theory and Applications," *Journal of Molecular and Engineering Materials* **8**, 2030001 (2020).
7. L.X. Chen, Z.W. Chen, M. Jiang, Z. Lu, C. Gao, G. Cai, C.V. Singh, "Insights on the dual role of two-dimensional materials as catalysts and supports for energy and environmental catalysis," *J. Mater. Chem. A* **9**, 2018-2042 (2021)
8. A. Jayakumar, A. Surendranath, Mohanan PV, "2D materials for next generation healthcare applications," *International Journal of Pharmaceutics* **551**, 309-321 (2018).
9. P. Snapp, J.M. Kim, C. Cho, J. Leem, M.F. Hague, S.W. Nam, "Interaction of 2D materials with liquids: wettability, electrochemical properties, friction, and emerging directions," *NPG Asia Materials* **12**, 22 (2020).
10. N. Briggs, S. Subramanian, Z. Lin, X. Li, X. Zhang, K. Zhang, K. Xiao, D. Geohegan, R. Wallace, L.-Q. Chen, M. Terrones, A. Ebrahimi, S. Das, J. Redwing, C. Hinkle, K. Momeni, A. Duin, V. Crespi, S. Kar, and J.A. Robinson, "A roadmap for electronic grade 2D materials," *2D Materials* **6**, 022001 (2019).
11. M.-M. Titirici, S. Baird, T. Sparks, S.M. Yang, A. Brandt-Talbot, O. Hosseinaei, D.P. Harper, R. Parker, S. Vignolini, L. Berglund, Y. Li, H.-L. Gao, L.-B. Mao, S. Yu, N. Diez, G. Alvarez.ferrero, M.S. Solis, P. Szilagyi, C. Stubbs, J. Worch, Y. Huang, C. Luscombe, K.-Y. Lee, H. Luo, J. Platts, D. Tiwari, D. Kovalevskiy, D. Fermin, H. Au, H. Alptekin, M.C. Ribadeneyra, V. Ting, T.-P. Fellinger, J. Barrio-hermida, O. Westhead, C. Roy, I. Stephens, S.A. Nicolae, S. Sarma, R. Oates, C.G. Wang, Z. Li, X.J. Loh, R.J. Myers, N. Heeren, A. Grégoire, Clément Périssé, X. Zhao, Y. Vodovotz, B. Earley, G. Finnveden, A. Björklund, G. Harper, A. Walton, P.A. Anderson, "The Sustainable Materials Roadmap," Accepted *JPhys Materials* (2022).

12. H. Kulik, T. Hammerschmidt, J. Schmidt, S. Botti, M.A.L. Marques, M. Boley, M. Scheffler, M. Todorović, P. Runke, C. Oses, A. Smolyanyuk, S. Curtarolo, A. Tkatchenko, A. Bartok, S. Manzhos, M. Ihara, T. Carrington, J. Behler, O. Isayev, M. Veit, A. Grisafi, J. Nigam, M. Ceriotti, K.T. Schütt, J. Westermayr, M. Gastegger, R. Maurer, B. Kalita, K. Burke, R. Nagai, R. Akashi, O. Sugino, J. Hermann, F. Noé, S. Pilati, C. Draxl, M. Kuban, S. Rigamonti, M. Scheidgen, M. Esters, D. Hicks, C. Toher, P. Balachandran, I. Tamblyn, S. Whitelam, C. Bellinger, L.M. Ghiringhelli, "Roadmap on Machine Learning in Electronic Structure," Accepted *Electronic Structure* (2022).
13. V. Nicolosi, M. Chhowalla, M.G. Kanatzidis, M.S. Strano, J.N. Coleman, "Liquid Exfoliation of Layered Materials," *Science* **340**, 1226419 (2013).
14. G. Fiori, F. Bonaccorso, G. Iannaccone, T. Palacios, D. Neumaier, A. Seabaugh, S.K. Banerjee, L. Colombo, "Electronics based on two-dimensional materials," *Nature Nanotechnology* **9**, 768-779 (2014).
15. F. Xia, H. Wang, D. Xiao, M. Dubey, A. Ramasubramaniam, "Two-dimensional material nanophotonics," *Nature Photonics* **8**, 899-907 (2014).
16. F.H.L. Koppens, T. Mueller, Ph. Avouris, A.C. Ferrari, M.S. Vitiello, M. Polini, "Photodetectors based on graphene, other two-dimensional materials and hybrid systems," *Nature Nanotechnology* **9**, 780-793 (2014).
17. D. Akinwande, N. Petrone, J. Hone, "Two-dimensional flexible nanoelectronics," *Nature Communications* **5**, 5678 (2014).
18. D. Morgan, R. Jacobs, "Opportunities and challenges for machine learning in materials science," *Annual Review of Materials Research* **50**, 71–103 (2020).
19. National Academies of Sciences, Engineering, and Medicine, "Frontiers of Materials Research: A Decadal Survey," Washington, DC: The National Academies Press (2019).
20. R. Stevens, V. Taylor, J. Nichols, A.B. Maccabe, K. Yelick, D. Brown "AI for Science," https://www.anl.gov/ai-for-science-report (2019).
21. C. Eberl, M. Neibel, T. Hickel, "National Research Data Infrastructure for Materials Science & Engineering," MatWerk (2021).
22. UK Department of Business, Energy & Industrial Strategy, "UK Innovation Strategy Report," https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1009577/uk-innovation-strategy.pdf (2021).
23 European Materials Modeling Council, "Ontology-Driven Data Documentation for Industry Commons," https://ontocommons.eu (2020).
24. Japan Council for Science and Technology Policy, "Strategic Innovations Program - Pioneering the Future: Japanese Science, Technology and Innovation," https://www8.cao.go.jp/cstp/panhu/sip_english/sip_en.html (2020).
25. Subcommittee on the Materials Genome Initiative, Committee on Technology, National Science and Technology Council, "Materials genome initiative strategic plan," https://www.mgi.gov/sites/default/files/documents/MGI-2021-Strategic-Plan.pdf (2021).
26. "MSDSN material data sharing," https://www.materdata.cn (2000).
27. S. O'Meara, "Materials science is helping to transform China into a high-tech economy," *Nature* **567**, S1-S5 (2019).
28. B. Hoefflinger, "ITRS: The international technology roadmap for semiconductors," Chapter 7, *CHIPS 2020—A Guide to the Future of Nanoelectronics*, Springer Berlin, Heidelberg (2020). ISBN 978-3-642-23096-7

29. The Minerals Metals & Materials Society (TMS), "Building a Materials Data Infrastructure: Opening New Pathways to Discovery and Innovation in Science and Engineering," TMS, Pittsburgh, PA (2017).
30. L. Himanen, A. Geurts, A.S. Foster, P. Rinke, "Data-Driven Materials Science: Status, Challenges, and Perspectives," *Advanced Science* 6, 1900808 (2019).
31. "MARDA - members working together to change the data landscape," https://www.marda-alliance.org
32. Y. Xiong, Q.T. Campbell, J. Fanghanel, C.K. Badding, H. Wang, N.E. Kirchner-Hall, M.J. Theibault, I. Timrov, J.S. Mondschein, K. Seth, R. Katz, A.M. Villarino, B. Pamuk, M.E. Penrod, M.M. Khan, T. Rivera, N.C. Smith, X. Quintana, P. Orbe, C.J. Fennie, S. Asem-Hiablie, J.L. Young, T.G. Deutsch, M. Cococcioni, V. Gopalan, H.D. Abruña, R.E. Schaak, I. Dabo, "Optimizing accuracy and efficacy in data-driven materials discovery for the solar production of hydrogen," *Energy Environ. Sci.* **14**, 2335 (2021).
33. National Academies of Sciences, Engineering, and Medicine, "Automated Research Workflows for Accelerated Discovery: Closing the Knowledge-Discovery Loop," Washington, DC: The National Academies Press (2022).
34. X. Liu, D.U. Furrer, J. Kosters, J. Holmes, "Vision 2040: a roadmap for integrated, multiscale modeling and simulation of materials and systems," Washington, D.C: National Aeronautics and Space Administration, 2018.
35. H. Wang, Y. Xiang, X. Xiang, and L. Chen, "Materials genome enables research and development revolution," *Sci Technol Rev* **33**, 13–19 (2015).
36. T. Austin, "Towards a digital infrastructure for engineering materials data," *Materials Discovery* **3**, 1-12 (2016).
37. R.F. Service, "'The game has changed.' AI triumphs at solving protein structures," *Science Magazine* News (2020).
38. "Construction and Operation of a Very Simple Radio Receiving Equipment," Bureau of Standards, Department of Commerce, Washington, D.C. (1922).
39. K. Ghosh, H. Ma, M. Onizhuk, V. Gavini, G. Galli, "Spin–spin interactions in defects in solids from mixed all-electron and pseudopotential first-principles calculations," *npj Computational Materials* **7**, 123 (2021).
40. I. Leven, and T. Head-Gordon, "C-GeM: Coarse-Grained Electron Model for Predicting the Electrostatic Potential in Molecules," *The Journal of Physical Chemistry Letters* **10**, 6820-6826 (2019).
41. O.Y. Long, G. Sai Gautam, E.A. Carter, "Evaluating optimal U for 3d transition-metal oxides within the SCAN+U framework," *Physical Review Materials* **4**, 045401 (2020).
42. K.-I., Nomura, R.K. Kalia, Y. Li, A. Nakano, P. Rajak, C. Sheng, K. Shimamura, F. Shimojo, P. Vashishta, "Nanocarbon synthesis by high temperature oxidation of nanoparticles," *Scientific Reports* **6**, 24109 (2016).
43. T. Verstraelen, P.W. Ayers, V. Van Speybroeck, M. Waroquier, "ACKS2: atom-condensed Kohn-Sham DFT approximated to second order," *Journal of Chemical Physics* **138**, 074108 (2013).
44. L.-P. Wang, A. Titov, R. McGibbon, F. Liu, V.S. Pande, T.J. Martinez, "Discovering chemistry with an ab initio nanoreactor," *Nature Chemistry* **6**, 1044-1048 (2014).
45. Y. Zhao, D.G. Truhlar, "The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals," *Theoretical Chemistry Accounts* **120**, 215-241 (2008).

46. N. Briggs, B. Bersch, Y. Wang, J. Jiang, R. Koch, N. Nayir, K. Wang, M. Kolmer, W. Ko, A.D.L.F. Duan, S. Subramanian, C. Dong, J. Shallenberger, M. Fu, Q. Zou, Y.-W. Chuang, Z. Gai, A.-P. Li, A. Bostwick, C. Jozwiak, C.-Z. Chang, E. Rotenberg, J. Zhu, A.C.T. van Duin, V.H. Crespi, J.A. Robinson, "Atomically thin half-van der Waals metals enabled by confinement heteroepitaxy," *Nature Materials* **19**, 637-643 (2020).
47. D.R. Hickey, N. Nayir, M. Chubarov, T.H. Choudhury, S. Bachu, L. Miao, Y. Wang, C. Qian, V.H. Crespi, J.M. Redwing, A.C.T. van Duin, N. Alem, "Illuminating invisible grain boundaries in coalesced single-orientation WS_2 monolayer films," *Nano Letters* **21**, 6487-6495 (2021).
48. T. Lookman, S. Eidenbenz, F. Alexander, C. Barnes, "Materials Discovery and Design: By Means of Data Science and Optimal Learning," Springer (2018). ISBN 9783319994659.
49. R. Cioffi, M. Travaglioni, G. Piscitelli, A. Petrillo, F. De Felice, "Artificial Intelligence and Machine Learning Applications in Smart Production: Progress, Trends, and Directions," *Sustainability* **12** (2020).
50. W. Ning, E.A. Silva, "Artificial Intelligence Solutions for Urban Land Dynamics: A Review," *Journal of Planning Literature* **24**, 246-265 (2010).
51. K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh, "Machine learning for molecular and materials science," *Nature* **559**, 547-555 (2018).
52. N. Jones, "How machine learning could help to improve climate forecasts," *Nature* **548**, 379-379 (2017).
53. G. Pilania, A. Mannodi-Kanakkithodi, B.P. Uberuaga, R. Ramprasad, J.E. Gubernatis, T. Lookman, "Machine learning bandgaps of double perovskites," *Scientific Reports* **6**, 19375 (2016).
54. C. Wen, Y. Zhang, C. Wang, D. Xue, Y. Bai, S. Antonov, L. Dai, T. Lookman, Y. Su, "Machine learning assisted design of high entropy alloys with desired property," *Acta Materialia* **170**, 109-117 (2019).
55. T. Hey, S. Tansley, K. Tolle, J. Gray, "The Fourth Paradigm: Data-Intensive Scientific Discovery," Microsoft Research, Redmond, WA (2009).
56. Google Research, "Research Areas: Machine Intelligence," https://research.google/research-areas/machine-intelligence (2022).
57. A.L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development* **44**, 206-226 (1952, available online in 2000).
58. F. Rosenblatt, "Principles of Neurodynamics: Perceptions and the Theory of Brain Mechanisms," Report No. VG-1196-G-8, Armed Services Technical Information Agency (1961).
59. A.G. Ivakhnenko, V. G. Lapa, "Cybernetics and Forecasting Techniques," Elsevier: USA (1967). ISBN 0444000208
60. E. Alpaydin, "Introduction to Machine Learning, 2nd Ed.," MIT Press (2010). ISBN 978-0-262-01243-0
61. S. Russell, P. Norvig, "Artificial Intelligence: A Modern Approach, 4th Ed.," Pearson (2020). ISBN 9780137505135
62. G.F. Luger, W.A. Stubblefield, "Artificial Intelligence: Structures and Strategies for Complex Problem Solving, 2nd Ed." Benjamin/Cummings Publishing Company: California (1993). ISBN 0-8053-4780-1
63. M. Liang, X. Hu, "Recurrent Convolutional Neural Network for Object Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3367-3375 (2015).
64. T. Young, D. Hazarika, S. Poria, E. Cambria, "Recent Trends in Deep Learning Based Natural Language Processing," *IEEE Computational Intelligence Magazine* **13**, 55-75 (2018).

65. G. De Michell, R.K. Gupta, "Hardware/software co-design," Proceedings of the IEEE 85, 349-365 (1997).
66. J. Trischler, S.J. Pervan, S.J. Kelly, D.R. Scott, "The Value of Codesign: The Effect of Customer Involvement in Service Design Teams," *Journal of Service Research* **21**, 75-100 (2017).
67. Subcommittee on the Materials Genome Initiative, Committee on Technology, National Science and Technology Council, "Materials genome initiative strategic plan," https://www.mgi.gov/sites/default/files/documents/mgi_strategic_plan_-_dec_2014.pdf (2014).
68. S. Curtarolo, W. Setyawan, G.L.W. Hart, M. Jahnatek, R.V. Chepulskii, R.H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M.J. Mehl, H.T. Stokes, D.O. Demchenko, and D. Morgan, "AFLOW: An Automatic Framework for High-Throughput Materials Discovery," *Computational Materials Science* **58**, 218-226 (2012).
69. J.E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, "Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD)," *JOM* **65**, 1501-1509 (2013).
70. A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K.A. Persson, "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation," *APL Materials* **1**, 011002 (2013).
71. "SGTE Substance Database (SGSUB), Scientific Group Thermodata Europe, https://www.sgte.net/en/neu
72. G. Bergerhoff, I.D. Brown in "Crystallographic Databases," F.H. Allen et al. (Hrsg.) Chester, International Union of Crystallography, (1987).
73. J.N. Lalena, D.A. Cleary, E.E. Carpenter, N.F. Dean, "Inorganic Materials Synthesis and Fabrication," Wiley (2007). ISBN 9780470191576.
74. X.-D. Xiang, "COMBINATORIAL MATERIALS SYNTHESIS AND SCREENING: An Integrated Materials Chip Approach to Discovery and Optimization of Functional Materials," *Annual Review of Materials Science* **29**, 149-171 (1999).
75. J. Yeo, G.S. Jung, F.J. Martin-Martinez, S. Ling, G.X. Gu, Z. Qin, M.J. Buehler, "Materials-by-Design: Computation, Synthesis, and Characterization from Atoms to Structures," *Phys. Scr.* **93**, 053003 (2018).
76. J.R. Chamorro, T.M. McQueen, "Progress toward Solid State Synthesis by Design," *Acc. Chem. Res.* **51**, 2918-25 (2018).
77. R. Ramesh, D.G. Schlom, "Creating emergent phenomena in oxide superlattices," *Nature Reviews Materials* **4**, 257-268 (2019).
78. F. Zhang, K. Momeni, M.A. AlSaud, A. Azizi, M.F. Hainey Jr, J.M. Redwing, L.-Q. Chen, N. Alem, "Controlled synthesis of 2D transition metal dichalcogenides: from vertical to planar MoS_2," 2D Materials 4, 025029 (2017).
79. Ed. J.P. Paglione, E.E. Rodriguez, "Fundamentals of Quantum Materials: A Practical Guide to Synthesis and Exploration," World Scientific (2021). ISBN 978-981-121-938-2
80. National Academies of Sciences, Engineering, and Medicine, "Frontiers of Crystalline Matter: From Discovery to Technology," Washington, DC: The National Academies Press (2009).
81. X.-D. Xiang, I. Takeuchi, "Combinatorial Materials Synthesis," CRC Press (2003). ISBN 9780203912737
82. C. Bannwarth, S. Ehlert, S. Grimme, "GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions," *Journal of Chemical Theory and Computation* **15**, 1652-1671 (2019).

83. P. Krstic, L. Han, S. Irle, H. Nakai, "Simulations of the synthesis of boron-nitride nanostructures in a hot, high pressure gas volume," *Chemical Science* **9**, 3803-3819 (2018).
84. M.S. Daw, M.I. Baskes, "Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals," *Physical Review B* **29**, 6443-6453 (1984).
85. D.W. Brenner, "Empirical potential for hydrocarbons for use in simulating the chemical vapor deposition of diamond films," *Physical Review B* **42**, 9458-9471 (1990).
86. T.-R. Shan, B.D. Devine, J.M. Hawkins, A. Asthagiri, S.R. Phillpot, S.B. Sinnott, "Second Generation Charge Optimized Many-Body (COMB) Potential for Si/SiO_2 and Amorphous Silica," *Physical Review B* **82**, 235302 (2010).
87. T. Senftle, S. Hong, M. Islam, S.B. Kylasa, Y. Zheng, Y.K. Shin, C. Junkermeier, R. Engel-Herbert, M. Janik, H.M. Aktulga, T. Verstraelen, A.Y. Grama, A.C.T. van Duin, A.C.T., "The ReaxFF Reactive Force-field: Development, Applications, and Future Directions," *Nature Computational Materials* **2**, 15011 (2016).
88. J. Vandermause, S.B. Torrisi, S. Batzner, Y. Xie, L. Sun, A.M. Kolpak, B. Kozinsky, "On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events," *npj Computational Materials* **6**, 20 (2020).
89. L. Zhang, J. Han, H. Wang, R. Car, E. Weinan, "Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics," *Physical Review Letters* **120**, 143001 (2018).
90. L.-Q. Chen, "Phase-Field Method of Phase Transitions/Domain Structures in Ferroelectric Thin Films: A Review," *Journal of the American Ceramic Society* **91**, 1835-1844 (2008).
91. T. Belytschko, R. Gracie, G. Ventura, "A review of extended/generalized finite element methods for material modeling," *Modelling and Simulation in Materials Science and Engineering* **17**, 043001 (2009).
92. W. Martin, G. Sheynkman, F.C. Lightstone, R. Nussinov, F. Cheng, "Interpretable artificial intelligence and exascale molecular dynamics simulations to reveal kinetics: Applications to Alzheimer's disease," *Current Opinion in Structural Biology* **72**, 103-113 (2022).
93. T. Xie, A. France-Lanord, Y. Wang, Y. Shao-Horn, J.C. Grossman, "Graph dynamical networks for unsupervised learning of atomic scale dynamics in materials," *Nature Communications* **10**, 2667 (2019).
94. P. Rajak, A. Krishnamoorthy, A. Mishra, R. Kalia, A. Nakano, P. Vashishta, "Autonomous reinforcement learning agent for chemical vapor deposition synthesis of quantum materials," *npj Computational Materials* **7**, 108 (2021).
95. ASM International, "Phase Diagram Resources from ASM International," https://www.asminternational.org/asm-phase-diagram-resources
96. "ACerS-NIST Phase Equilibria Diagrams Database," NIST Standard Reference Database 31 (2021).
97. L. Kaufman, H. Bernstein, "Computer Calculation of Phase Diagrams," Academic Press, NY (1970). ISBN 0-12-402050-X
98. Thermo-Calc Software, https://thermocalc.com/products/thermo-calc/
99. For a recent semiconductor example, see: https://www.theregister.com/2022/02/10/kioxia_wd_chemical_contamination_flash_fabs/ (Accessed 2/10/2022).
100. J.M. Granda, L. Donina, V. Dragone, D.-L. Long, L. Cronin, "Controlling an organic synthesis robot with machine learning to search for new reactivity," *Nature* **559**, 377-381 (2018).
101. W. Gao, P. Raghavan, C.W. Coley, "Autonomous platforms for data-driven organic synthesis," *Nature Communications* **13**, 1075 (2022).

102. A.G. Kusne, H. Yu, C. Wu, H. Zhang, J. Hattrick-Simpers, B. DeCost, S. Sarker, C. Oses, C. Toher, S. Curtarolo, A.V. Davydov, R. Agarwal, L.A. Bendersky, M. Li, A. Mehta, I. Takeuchi, "On-the-fly closed-loop materials discovery via Bayesian active learning," *Nature Communications* **11**, 5966 (2020).
103. "EDISON: Building the data science profession," https://edison-project.eu/
104. The Minerals Metals & Materials Society (TMS), "Employing Artificial Intelligence to Accelerate Development and Implementation of Materials and Manufacturing Innovations," TMS, Pittsburgh, PA (2021).
105. P. Wittenburg, "Open science and data science", *Data Intelligence* **3**, 95–105 (2021).
106. A.D. Sendek, Q. Yang, E.D. Cubuk, K.-A. N. Duerloo, Y. Cui, E.J. Reed, "Holistic computational structure screening of more than 12 000 candidates for solid lithium-ion conductor materials," *Energy Environ. Sci.* **10**, 306-320 (2017).
107. S. Back, J. Yoon, N. Tian, W. Zhong, K. Tran, Z.W. Ulissi, "Convolutional Neural Network of Atomic Surface Structures To Predict Binding Energies for High-Throughput Screening of Catalysts," *J. Phys. Chem. Lett.* **10**, 4401-4408 (2019).
108. S. Gazibegovic, D. Car, H. Zhang, S.C. Balk, J.A. Logan, M.W.A. de Moor, M.C. Cassidy, R. Schmits, D. Xu, G. Wang, P. Krogstrup, R.L.M. Op het Veld, K. Zuo, Y. Vos, J. Shen, D. Bouman, B. Shojaei, D. Pennachio, J.S. Lee, P.J. van Veldhoven, S. Koelling, M.A. Verheijen, L.P. Kouwenhoven, C.J. Palmstrøm, E.P.A.M. Bakkers, "Retraction Note: Epitaxy of advanced nanowire quantum devices," *Nature* **604**, 786 (2022).
109. Interagency Working Group on Digital Data (IWGDD), "Harnessing the Power of Digital Data for Science and Society," Report to the Committee on Science of the National Science and Technology Council (2009).
110. "The world's most valuable resource is no longer oil, but data," *The Economist*, May 5th, 2017.
111. B. Mons, H. van Haagen, C. Chichester, P.-Bram't Hoen, J.T den Dunnen, G. van Ommen, E. van Mulligen, B. Singh, R. Hooft, M. Roos, J. Hammond, B. Kiesel, B. Giardine, J. Velterop, P. Groth, E. Schultes, "The value of data," *Nature Genetics* **43**, 281-283 (2011).
112. "Basic Research Needs for Transformative Manufacturing: Fundamental science to revolutionize manufacturing," DOE-BES BRN Report (2020).
113. "Producing and Managing Large Scientific Data with Artificial Intelligence and Machine Learning," DOE-BES Roundtable Report (2019).
114. "ASCR Workshop on In Situ Data Management: Enabling Scientific Discovery from Diverse Data Sources," DOE-ASCR Report (2019).
115. "Data and Models: A Framework for Advancing AI in Science," DOE-ASCR Report (2019).
116. M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci Data* **3**, 160018 (2016).
117. A. Jacobsen, R. de Miranda Azevedo, N. Juty, D. Batista, S. Coles, R. Cornet, M. Courtot, M. Crosas, M. Dumontier, C.T. Evelo, C. Goble, G. Guizzardi, K.K. Hansen, A. Hasnain, K. Hettne, J. Heringa, R.W.W. Hooft, M. Imming, K.G. Jeffery, R. Kaliyaperumal, M.G. Kersloot, C.R. Kirkpatrick, T. Kuhn, I. Labastida, B. Magagna, P. McQuilton, N. Meyers, A. Montesanti, M. van Reisen, P.

Rocca-Serra, R. Pergl, S.-A. Sansone, L.O.B. da Silva Santos, J. Schneider, G. Strawn, M. Thompson, A. Waagmeester, T. Weigel, M.D. Wilkinson, E.L. Willighagen, P. Wittenburg, M. Roos, B. Mons, E. Schultes, "FAIR Principles: Interpretations and Implementation Considerations," *Data Intelligence* **2**, 10-29 (2020).

118. M. Scheffler, M. Aeschlimann, M. Albrecht, T. Bereau, H.-J. Bungartz, C. Felser, M. Greiner, A. Groß, C.T. Koch, K. Kremer, W.E. Nagel, M. Scheidgen, C. Wöll. C. Draxl, "FAIR data enabling new horizons for materials research," *Nature* **604**, 635-642 (2022).
119. L.C. Brinson, L.M. Bartolo, B. Blaiszik, D. Elbert, I. Foster, A. Strachan, P.W. Voorhees, "FAIR Data will Fuel a Revolution in Materials Research," arXiv:2204.02881 (2022).
120. "Research Data Management Glossary," CASRAI (2020). https://casrai.org/term/data-lifecycle/
121. B, Mons, C. Neylon, J. Velterop, M. Dumontier, L.O.B. da Silva Santos, Luiz Olavo Boninob, M.D. Wilkinson, "Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud," Information Services and Use 37, 49-56 (2017).
122. S. Sinek, "Start with why: How great leaders inspire everyone to take action," Portfolio (2011). ISBN 9781591846444
123. "Zenodo: Research Shared," https://zenodo.org/
124. NIST, "Materials Data Repository," https://www.nist.gov/programs-projects/materials-data-repository
125. "GitHub," https://github.com/
126. "Illinois Data Bank," https://databank.illinois.edu/
127. "Johns Hopkins University Data Archive," https://archive.data.jhu.edu/
128. "DRYAD," https://datadryad.org/stash
129. "Lifetime Sample Tracking (LiST) Data Management System," https://www.mri.psu.edu/2d-crystal-consortium/user-facilities/data-management-list
130. "PARADIM Data Collective," https://data.paradim.org

VII. Appendices

A. Appendix A

What is this project?

This was an initiative of the National Science Foundation in collaboration with the NSF-funded materials innovation platforms PARADIM and 2DCC arising out of a sea of data activities within 2D materials community over the last 5 years.

Why did we do it?

Materials research is at a crossroads, and we felt that this was the best time to create a roadmap for synthesis and discovery within interfacial and layered materials over the next decade and to continue to build a community who will work together and share knowledge.

Why now?

There is an alignment of the planets. There is rapid advancement of data science and AI/ML methodologies and there is significant investment and recommended further investment in national midscale synthesis and characterization capabilities. The next MGI strategic plan has been released and, at the same time, a new NSF Directorate for Technology, Innovation and Partnerships (TIP) is just getting off the ground.

What is the output?

The most tangible output is a roadmap in the form of this report to capture actionable items from the Forums that can be used by the community, the Materials Innovation Platforms, and funding bodies. At the same time, we anticipate building even stronger connections between participants that will continue to grow the community.

How did we scope the problem?

We started by scoping the problem to identify core themes of interest to the community. We did this through an online questionnaire designed to elicit their views of the most interesting questions that might be addressed by (or within) 2D Interfacial Materials in the next 10-15 years.

Raw responses to the questionnaire (Table C1) were then reviewed by the core team.

The community then sorted these questions to arrive at a broad set of themes to explore at the first workshop.

Who attended the forums, and when were they held?

Attendees are listed in Appendix B. Attendees were drawn from across a variety of communities across both data and materials science, being selected to cover the breadth and depth of the community from a pool of applicants. The first forum was held virtually and focused on:

January 20th, 2022: Developing Future Headlines and exploring the challenge areas.

January 21st, 2022: Prioritizing Antecedent or Fundamental Research and exploring the obstacles.

A second forum was held virtually on March 10th, 2022 and focused on elaborating precise actions and directions to achieve the vision and overcome the challenges and obstacles identified in the January forum.

How did we develop the “Future Headlines”?

In order to explore the challenges and opportunities over the next 10 to 15 years that might respond to a solution created by 2D and Interfacial Materials we invited participants to imagine a future where that challenge or opportunity had already been addressed so that we could then step back and identify the fundamental research that will be required in the medium term to achieve those future headlines.

This was a deliberately creative approach where the output was less important than the process. It helped participants identify future opportunities and challenges beyond their own day-to-day research and avoids the old trope: "To a person with a hammer, every problem looks like a nail". The process of developing the headlines was designed to stimulate and inspire the ambition and breadth of our thinking as we moved through the workshop.

Headlines developed through this process are given in Table C2.

How did we align around the Research Challenges (the things we would need to understand if we were to achieve the “Future Headlines”)?

Having developed the “Future Headlines”, the participants explored, added to and refined the clusters of challenges (the research questions that would have to be answered if we were to achieve the Future Headlines), which, if answered, might lead to a step-change in the field of 2D Interfacial Materials. These are given in Table C3.

How did we identify the Fundamental or Antecedent Research that might be necessary to address the challenges?

For each of the high-level challenges we explored what antecedent research might be required in order to answer those questions and captured them in the form of Google Documents (one for each high-level question). We then asked participants to prioritize this research to give a clear roadmap of what needs to be done and in what order. We also asked participants to identify the obstacles that would hinder the antecedent research. The obstacles are summarized in Table C4.

How did we arrive at the themes to elaborate on and provide directions for in this report?

At the conclusion of the January forum, the core team along with the steering committee evaluated process and found substantial similarity amongst portions of the challenge clusters arising from the forum, to arrive at four theme areas, Table C5. We further felt that actionable areas had not been sufficiently fleshed out in the first forum. We thus ran a one-day forum on March 10th, 2022 to gather further community insight into elaboration of what is needed to make transformative progress. The outputs of this forum were then independently synthesized by the MIP directors and the core team.

The core team then took both of these interpretations, along with additional feedback from the steering committee, mentors, and other stakeholders to complete the first draft of this report.

How was public comment solicited?

There was a multi-month public comment period in summer 2022, with multiple avenues of feedback, including an anonymous web form at https://www.materialsarchive.org/2dilm-report2022/feedback.php . Feedback from this process was incorporated into the final version.

What are next steps?

We anticipate this report to be a beginning – a preface if you will – to a renaissance in the community about making meaningful advances in materials discovery to realize the potential of 2D ILM materials through advances in data science, and hope that it stimulates significant community and research efforts to bring the potential of the field to fruition. We hope that others use and build upon this report as we enter a new era of materials discovery.

B. Appendix B

January 2022 Forum: Future Headlines and Challenge Clusters Participants

Douglas Adamson	Professor	University of Connecticut
Mahshid Ahmadi	Assistant Professor	University of Tennessee, Knoxville
Virginia Ayres	Associate Professor	Michigan State University
Prasanna V. Balachandran	Assistant Professor	University of Virginia
Alexander Balandin	Distinguished Professor or ECE and UC Presidential Chair Professor of Materials Science	University of California Riverside
Art Counts	Engineer	Apple
Albert Davydov	Group Leader	NIST
Aida Ebrahimi	Assistant Professor	Penn State University
Panchapakesan Ganesh	Senior R&D Staff Scientist	Oak Ridge National Laboratory
Feliciano Giustino	W.A. "Tex" Moncrief, Jr. Chair of Quantum Materials Engineering	The University of Texas at Austin
Pelagia Gouma	Professor	The Ohio State University
John Gregoire	Research Professor	Caltech
Geoffroy Hautier	Associate Professor	Dartmouth College
Michael Heuken	VP Advanced Technologies	Aixtron
Pinshane Huang	Associate Professor	University of Illinois at Urbana-Champaign
Piran Kidambi	Assistant Professor	Vanderbilt University
Mercouri Knaatzidis	Professor	Northwestern University
Tillmann Kubis	Research Assistant Professor	Purdue University
Peng Li	Assistant Professor	Auburn University
Vinod Menon	Professor of Physics	City College & Grad Center of CUNY
Vincent Meunier	Professor	Rensselaer Polytechnic Institute
Rohan Mishra	Assistant Professor	Washington University at St. Louis
John MItchell	Senior Scientist	Argonne National Laboratory
Kasra Momeni	Associate Professor	The University of Alabama
Martin Mourigal	Associate Professor	Georgia Tech
Krishna Rajan	SUNY Distinguished Professor & Erich Bloch Chair	University at Buffalo
Jayakanth Ravichandran	Assistant Professor	University of Southern California
Wesley Reinhart	Assistant Professor	Penn State University
Thomas Searles	Associate Professor	University of Illinois Chicago
Daniel Shoemaker	Associate Professor	University of Illinois
Derek Stewart	Technologist	Western Digital
Christopher Stiles	Senior Research Scientist	Johns Hopkins University Applied Physics Laboratory
Eric Toberer	Professor	Colorado School of Mines
Arend van der Zande	Associate Professor	University of Illinois at Urbana-Champaign
Patrick Vora	Associate Professor	George Mason University

March 2022 Workshop: Addressing the Challenges Participants

How might we better predict outcomes from actual synthesis conditions?		
Paul Canfield	Distinguished Professor	Iowa State University and Ames Laboratory
Stefano Curtarolo	Edmund T. Pratt Jr. School Distinguished Professor of Mechanical Engineering and Materials Science	Duke University
Geoffroy Hautier	Hodgson Family Associate Professor of Engineering	Dartmouth University
Igor Levin	Lead, Materials Structure and Data Group	National Institute of Standards and Technology
Zi-Kui Liu	Professor	Pennsylvania State University
Alexandra Navrotsky	Distinguished Professor Emeritus	UC Davis
How might we achieve AI-assisted and autonomous synthesis?		
Darryl Barlett	CEO	k-space Associates Inc.
Benji Maruyama	Principal Materials Research Engineer	Air Force Research Laboratory
Peter Rosenthal	Manager, Technology and Product Development	Coherent, Inc.
Daniel Shoemaker	Associate Professor	University of Illinois at Urbana-Champaign
Rudolf M. Tromp	Scientist and Professor	IBM Research and Leiden University
How might we prepare practitioners to better use data to accelerate discovery?		
Mahshid Ahmadi	Assistant Professor	University of Tennessee - Knoxville
Vasant Honavar	Huck Chair in Biomedical Data Sciences and Artificial Intelligence	Pennsylvania State University
Bryce Meredig	Co-Founder and Chief Science Officer	Citrine Informatics
Klara Nahrstedt	Grainger Distinguished Chair in Engineering	University of Illinois Urbana-Champaign
Alex Norquist	Professor	Haverford College
Wes Reinhart	Assistant Professor	Pennsylvania State University
Olga Wodo	Associate Professor	SUNY Buffalo
How might we promote knowledge sharing?		
Vin Crespi	Distinguished Professor of Physics	Pennsylvania State University
Laura Franklin	Portfolio Manager, Materials Open Research	Taylor & Francis
Jane Greenberg	Alice B. Kroeger Professor and Director, Metadata Research Center	Drexel University
Linda Hung	Senior Research Scientist	Toyota Research Institute
Martin Mcbriarty	Scientist	EMD Electronics
Jim Warren	Director, NIST Center for Theoretical and Computational Materials Science	National Institute of Standards and Technology