Data-Centric Engineering (2021), 2: e19
doi:10.1017/dce.2021.19

CAMBRIDGE UNIVERSITY PRESS

PERSPECTIVE

Materials informatics and sustainability—The case for urgency

Hannah R. Melia*, Eric S. Muckley and James E. Saal

Citrine Informatics, Redwood City, California 94063, USA
*Corresponding author. E-mail: h.r.melia.96@cantab.net

E.S.M. and J.E.S. contributed equally to this work.

Received: 30 June 2021; **Revised:** 15 October 2021; **Accepted:** 20 October 2021

Keywords: Artificial intelligence; data; materials informatics; sustainability; technology adoption

Abstract

The development of transformative technologies for mitigating our global environmental and technological challenges will require significant innovation in the design, development, and manufacturing of advanced materials and chemicals. To achieve this innovation faster than what is possible by traditional human intuition-guided scientific methods, we must transition to a materials informatics-centered paradigm, in which synergies between data science, materials science, and artificial intelligence are leveraged to enable transformative, data-driven discoveries faster than ever before through the use of predictive models and digital twins. While materials informatics is experiencing rapidly increasing use across the materials and chemicals industries, broad adoption is hindered by barriers such as skill gaps, cultural resistance, and data sparsity. We discuss the importance of materials informatics for accelerating technological innovation, describe current barriers and examples of good practices, and offer suggestions for how researchers, funding agencies, and educational institutions can help accelerate the adoption of urgently needed informatics-based toolsets for science in the 21st century.

Impact Statement

Over the coming decades, human societies will face unprecedented challenges related to climate, energy, and water security. A key component to address these is the rapid development of a broad range of transformative technologies. Historically, these technologies were developed using the traditional scientific method, in which a human scientist performed experiments by trial and error. However, the short timeframe with which technologies must be developed requires significantly faster materials discovery and commercialisation methods than what is commonly being practiced today. This article describes the importance of developing advanced materials and chemicals to enable breakthrough technologies, why we must utilise cutting-edge data-driven tools to accelerate this process, and how researchers, funding agencies, and educational institutions can help improve adoption of this necessary data-centric scientific paradigm.

1. Introduction

A 2018 report by the UN's Intergovernmental Panel on Climate Change suggests that humanity must rapidly solve a broad range of global technological challenges by 2030 to avoid unprecedented environmental, social, and economic risks (IPCC, 2018). Already in 2015, the UN defined 17 sustainable

development goals (United Nations, 2015) which encompass a diverse set of important themes ranging from social justice and public health to environmental stewardship. Many of these goals, particularly responsible consumption and production, water security, clean energy generation, and better healthcare, may be directly addressed by innovation in the materials and chemicals industries. Advances in materials are critical for, among other things, improving the performance of battery materials, photovoltaics, and carbon capture techniques; the use of new feedstocks for molecular recycling of plastics and bioplastics; reduced reliance on critical elements; and making the production of materials more energy- and carbon-efficient using catalysts for electrochemical processes, hydrogen reduction of iron ore, clinker substitutes in cement, and decreased use of high-temperature heat treatments of metals.

Production of materials accounts for nearly a quarter of total global greenhouse gas emissions (Hertwich, 2021). For many companies in the materials and chemicals industries, sustainability themes have historically been driven by public relations and marketing narratives. Recently, however, due to strains on the availability of critical materials, proposed EU regulation (European Commission, 2020), and customer demand for environmental responsibility, sustainability initiatives have become critical to the survival of the industries, and as such, ambitious goals have been set (BASF, 2020; ArcelorMittal, 2021). These goals generally aim to address one of four primary themes: (a) reduction in greenhouse gas emissions, (b) increases in energy efficiency, (c) sale of more sustainable products, and (d) responsible sourcing of raw materials.

While each of these goals depends on the development of novel high-performance materials and chemicals, historically it has taken an average of 20 years to bring new material to market after its discovery in the lab (Materials Genome Initiative, 2021). That timescale prevents rapid deployment of new technologies which are essential for enabling significant mitigation of climate change before 2030. To accelerate the process of materials discovery and deployment, we must move beyond the historical paradigm of human-centered trial-and-error approaches in the lab and employ cutting-edge tools, high-performance computing resources, and data-driven methodologies to harness the full power of human ingenuity and innovation.

The authors of this perspective piece work for a company that has engaged in 60+ materials informatics projects over the last 8 years with commercial, academic, and government partners. Seeing the success of these projects, and the acceleration that material informatics has catalysed for materials development, motivates this paper.

2. Adoption of Informatics in the Materials and Chemicals Industries

Large-scale digital transformation is occurring across a broad range of industries, fueled by cheap computing power, proliferation of cloud-based database hosting infrastructure, ubiquitous data collection, and powerful artificial intelligence (AI). Materials and chemicals companies are also following digitalisation trends, and industry leaders have begun adopting systematic data-driven R&D practices to optimize materials and formulations through tuning of composition and processing conditions. Many manufacturers are creating "digital twins" by leveraging complex computer simulations to test the design and performance of their products in a cost-efficient manner. Physics-based models from the nano- to mesoscales can faithfully reproduce physical behavior, while AI and other data-driven tools can be used to interpolate results at intermediate length scales. AccelorMittal, currently the top-ranked producer of steel in the world by volume, announced that "global R&D is focusing on launching digital transformation projects throughout all aspects and segments of the business" (ArcelorMittal, 2021). Similarly, BASF, the largest global chemical company, stated that "we integrate digital technologies into everyday operations and make them an integral part of any R&D project workflow to boost effectiveness of research, increase efficiency and open up new innovation opportunities" (BASF, 2021). These vignettes highlight a new paradigm that is transforming the materials and chemicals industry: materials informatics.

Materials informatics (MI), the application of data science, materials science, and AI to the materials and chemicals space, has enabled researchers to leverage complex, data-driven insights for the discovery of novel materials faster than ever before by reducing the number of experiments required during the

https://doi.org/10.1017/dce.2021.19 Published online by Cambridge University Press

materials development process by 50–70% (Ling et al., 2017; Saal et al., 2020). AI is an ideal tool for this challenge, as it excels at high-dimensional, multiobjective optimisation, enabling simultaneous refinement of processing and composition parameters, which helps push products toward desired property targets at both lab and production scales. The speed at which data-driven models can output new predictions enables researchers to cast a wider net, exploring compositions that would be lower down the priority list for physical experiments, which may lead to the discovery of novel, highly differentiated materials. Furthermore, AI models are not constrained to conventional modes of thinking like typical human researchers, which enables them to highlight new or unexpected results that a traditional researcher may ignore.

Before deployment in real-world commercial or industrial applications, novel materials must meet a myriad of constraints including processing requirements, cost, sustainability, durability, esthetics, safety, and functionality. This development process involves multidimensional optimisation and extensive testing, a costly and time-consuming process that is heavily reliant on human input and domain expertise. However, with the cutting-edge tools of materials informatics at their disposal, researchers could accelerate this process by identifying important patterns across datasets that are too large or complex to be understood by traditional means (Cao et al., 2018; Mosavi et al., 2018), reducing the number of experiments necessary to mature a technology from lab bench to market. While adoption of materials informatics does not come without challenges (Citrine Informatics, 2020), great strides have been made in the last 5 years toward the development of open-source materials-specific data models (Citrine Informatics, 2021), incorporation of expert domain knowledge into AI models (Childs & Washburn, 2019), transfer-learning architectures that can accommodate small, sparse datasets (Hutchinson et al., 2017), and uncertainty quantification methods for enabling targeted iterative AI and Bayesian optimisation (Mosavi et al., 2018). As these powerful tools gain traction in the materials and chemicals communities, newly developed and existing software platforms will continue to enable broad adoption of MI across the industry.

Common MI applications encountered in materials and chemicals industry today include

- the discovery of new materials that have specific target properties,
- optimisation of the composition or processing parameters of existing materials,
- identification of formulations that simultaneously meet performance, cost, and sustainability criteria, and
- identification of the most informative experiments to perform under budgetary requirements or other constraints.

As widespread adoption of MI continues across materials and chemicals industries and advances in computing enable more powerful modeling techniques, it is expected that the core challenges addressed by today's materials informatics tools will expand to encompass new domains and materials classes.

The following two case studies illustrate innovative uses of MI, in the first to replace animal testing and in the second to guide strategic decision making.

2.1. Case study: Screening out toxic formulations

Safety, including toxicity, is a primary design consideration during chemicals and materials development. Animal testing is commonly used as a final safety check for a chemical in development before bulk production can occur, and this process often requires many months of testing. Failing at this last hurdle can be expensive, requiring a return to an earlier point of the chemical design process. Therefore, development of accurate model-based prediction of toxicity can significantly reduce materials and chemicals screening costs by ensuring success of toxicity tests. More importantly, alternative screening methods can reduce or even eliminate the use of controversial animal testing practices.

While there is robust literature demonstrating prediction of toxicity for single chemicals, prediction of toxicity for multiphase complex formulations is more difficult. A chemical manufacturer recently used

https://doi.org/10.1017/dce.2021.19 Published online by Cambridge University Press

our company's AI platform to develop an AI model that can predict toxicity for such complex formulations based on ingredient composition and recipe. It took them 4 weeks to develop a toxicity model with 82% accuracy.

Incorporating the domain knowledge of human formulation experts was key. The AI platform did this in two ways. First, the ingredients were labeled by type (active, adjuvant, etc.) and the fractions of different types of ingredients were used as inputs to the model. Second, the platform used SMILES (Weininger et al., 1989) strings representing chemical formulas for the active ingredients and generated 30+ descriptors (e.g., molecular weight and the number of hydrogen bond donors), which were critical for improving model performance when the quantity of input data and availability of descriptors is limited.

2.2. Case study: Materials informatics used for developing novel battery materials

The transition to a carbon-neutral economy will require significant advances in battery materials (Nitta et al., 2015) which increase specific energy while maintaining high voltage. One of the primary challenges in battery development is the trade-off between battery performance and availability of raw materials (Peerless et al., 2020), which often varies by manufacturer (Vikström et al., 2013).

In a prior work (Peerless et al., 2020), three existing databases of battery materials were combined and then sorted into two groups depending on their composition and the availability of their constituent elements in the earth's crust, scarce or abundant. An AI model was trained and used to predict the specific energy and voltage of different material combinations. Uncertainty quantification (UQ), the calculation of the expected accuracy of individual model predictions, was used to visualize the likelihood of achieving target properties in the two design spaces. By calculating the likelihood of achieving target material properties for every candidate in the design space, potential performance of designs can be assessed, and the probability of achieving battery design goals with and without scarce elements can be estimated and used to inform research strategy. This analysis (Figure 1) showed that a higher specific energy is easier to

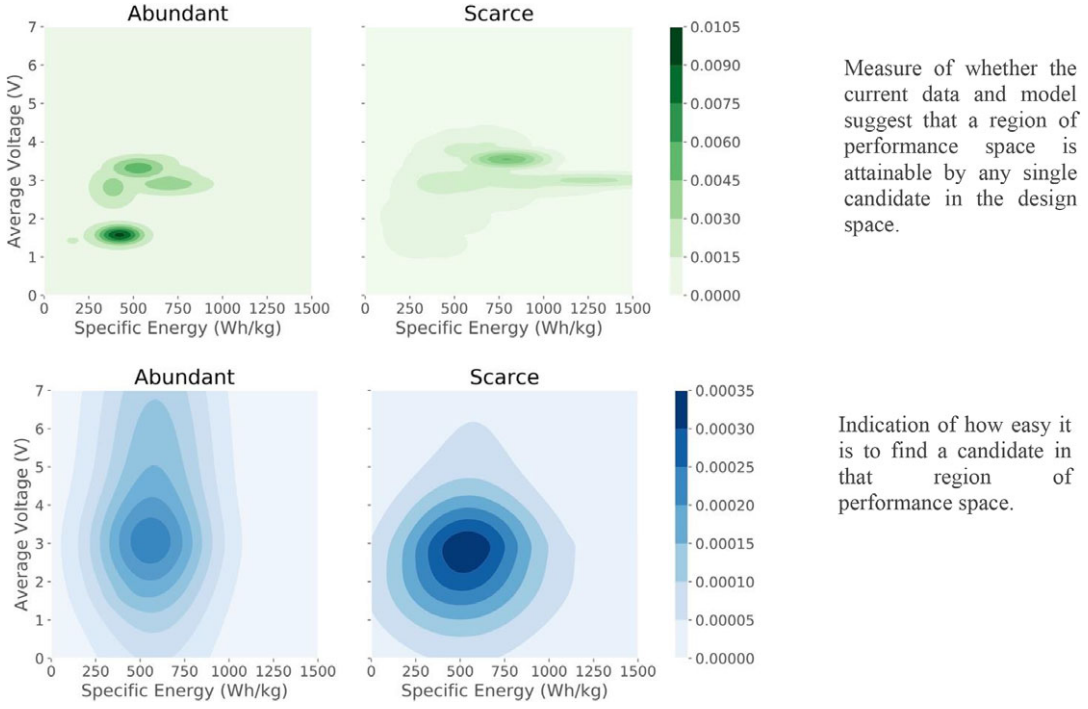

Figure 1. *Design space visualisations for battery materials (Peerless et al., 2020). Reproduced under creative commons license.*

https://doi.org/10.1017/dce.2021.19 Published online by Cambridge University Press

achieve with cathodes that use rare elements, which provided researchers information they used to make data-driven decisions about the likelihood of success when performing physical tests with different materials.

The strategic utility of MI is important for commercial organisations as they race to develop sustainable products and rely on smart uncertainty qualification.

3. What's Holding MI Back?

While materials informatics will be a key component of digital transformation, adoption of MI is not yet universal across the entire use-case spectrum, including fundamental research at universities, translational research at institutes, research at small- and medium-scale companies, and R&D at global materials and chemicals corporations. To achieve such widespread adoption, challenges involving skills gaps, cultural alignment, and data must be addressed.

3.1. Skills

In an environment where data is king, researchers with the data management skills needed to wrangle data into an informatics platform will become more readily available as many systems across the organisation rely on these same skills. Similarly, as materials informatics platforms continue to evolve and mature and best practices are baked in, the requisite data science skills to utilise these tools shrink and usability across a wider range of user backgrounds improves. However, a general understanding of fundamental data science concepts such as design spaces, latent variables, normalisation/dimensionality reduction, and uncertainty quantification will remain one of the most important factors determining materials informatics success. Familiarity with these concepts enables researchers to effectively communicate between different teams in an organisation, ensuring that relevant domain knowledge is integrated and leveraged by the models. A required course on key data science concepts for students in the physical sciences would help mitigate this knowledge gap and facilitate more cross-cutting research at universities and other academic institutions where projects are often siloed in a single department even when interdisciplinary collaboration is encouraged by funding agencies. There is also a need for courses that can add to the existing skillsets of experienced researchers. For example, a short crash course in advanced data science topics for materials scientists may be modeled after an analogous 1-day course designed for petroleum engineers (Society of Petroleum Engineers, 2021).

The Minerals, Metals & Materials Society (TMS) report Creating the Next-Generation Materials Genome Initiative Workforce (The Minerals, Metals & Materials Society (TMS), 2019) recommends addressing skill gaps in three key areas: data (data handling, visualisation, and software for materials workflows), computation (first-principles, microstructure, and multiscale modeling), and experiments (uncertainty-informed multiobjective decision making and automated high-throughput methods). The study found that materials researchers, including undergraduate and graduate students, should develop awareness of informatics tools and become conversant in topical areas including data management and measurement tools.

The computational materials community has increased industry-focused outreach by sponsoring short courses, hackathons, and workshops such as the NIST/University of Maryland "Machine Learning for Materials Research Bootcamp" (University of Maryland, 2021). Other examples include the Computational Materials Science Summer School hosted at Texas A&M University (2021) the Machine Learning for Materials Workshop hosted by NIST (2021) and the Workshop on Artificial Intelligence Applied to Materials Discovery and Design hosted by DOE's EERE (U.S. Department of Energy, 2021).

Ultimately, investments in workforce development are needed to ensure that materials engineers are comfortable using modern scientific analytics methods, including simulations for creating digital twins, AI for developing predictive data-driven models, and core data science and statistical techniques for applying uncertainty quantification and analysing and manipulating large datasets across multiple file formats and organisational structures.

https://doi.org/10.1017/dce.2021.19 Published online by Cambridge University Press

3.2. Culture

3.2.1. Fear of the new and redundancy

As with every new technology, early adopters help develop use-cases and drive methodology forward while others remain skeptical and only accept change once there is no alternative. Knowledge of what AI can do for materials science is in part limited by incumbent mindsets. There is also the suspicion of AI in popular culture, which has fostered its association with potential future job losses, unaccountable and biased decision making, and the potential for computers to take over the world and declare war on humankind!

On a more personal level, scientists and engineers often spend a significant fraction of their lives in higher education and justifiably place a high value on the domain knowledge they have gained over years of study and research. Their self-worth is therefore heavily tied to being the provider of that domain knowledge. Scientists are also trained not to blindly trust something that they do not understand; they abhor a "black box."

To help alleviate these concerns, careful education is needed to help materials researchers see AI as a tool in a suite of other technologies that enables them to do their jobs more efficiently, rather than a threat to their livelihood or self-worth. Emphasis should be placed on how AI models can provide insight into underlying physical mechanisms which are difficult for humans to grasp because of high-dimensionality and the interconnectedness of multiple physical phenomena. AI need not be a black box, but rather can shine a light on subtleties that are too complex for humans to easily identify. Deep domain knowledge and expertise are still needed more than ever. Rather than a "big data" problem, materials informatics often suffers from a "small data" problem of complex, sparse, and noisy datasets, and one of the most effective methods to compensate for shortcomings in datasets is the expertise and domain knowledge of a well-trained materials scientist.

3.2.2. Is it worth the cost?

As with the adoption of any new technology, transitioning to an MI-focused culture will incur costs. However, these costs will be recouped from savings made on the reduced costs of developing materials and chemicals when compared to traditional workflows within a similar timeframe. Companies are seeing benefits from Materials Informatics in three ways: acceleration of development (fewer experiments per project), more information to direct research investments, and capture and systematic reuse of company intellectual property. While the reduction in number of experiments needed is well known, the other two benefits need a little more explanation. Using uncertainty quantification, researchers can calculate the probability of one or more material candidates in a design space hitting a target property set. This process can be carried out in different design spaces (i.e., representing different research directions) and the probability compared. Researchers therefore can choose in a data-driven way whether to go a high-risk route that might achieve remarkable results or the lower risk direction for incremental improvement. The final benefit is a side effect of codifying knowledge. Domain knowledge is captured in datasets, design spaces, and AI models which become digital assets that can be reused by future researchers on adjacent projects. Researchers can both learn from these resources and adapt and reuse them to accelerate their own research. This is particularly important where many knowledgeable researchers are nearing retirement. In commercial companies, of course, they want to make sure that company IP does not leave the building when a researcher decides to change jobs.

Open-source data science tools can enable insightful deep data analytics for individual projects and teaching purposes. For example, the popular Matminer Python package (Ward et al., 2018) provides free visualisation tools, open materials datasets, hooks for connecting to external databases, and data processing utilities to support machine learning. Pymatgen (Ong et al., 2013) provides tools for analysing material structure and thermodynamics, connecting to the Materials Project database, and generating phase diagrams and other visualisations. Materials Cloud (Talirz et al., 2020) and nanoHUB (Madhavan et al., 2013) provide free curated datasets, computation tools, educational videos, and training seminars.

https://doi.org/10.1017/dce.2021.19 Published online by Cambridge University Press

However, to get the most out of MI at the organisational level, scalable platforms to develop pipelines for acquisition, storage, and analysis of data must be utilised. As an example, the NFDI-MatWerk consortia (NFDI-MatWerk, 2021) is building distributed materials data infrastructure for facilitating interoperability and consistency between materials discovery workflows across different research institutions (Fraunhofer Institute for Mechanics of Materials, 2021).

While the extra effort and cost to implement such software tools raise an initial activation barrier to MI, the improvements in product development efficiency compound over time as knowledge and data are digitized, resulting in significant net savings. As prominent MI successes accumulate (Saal et al., 2020), the utility of MI will become more clearly evident to a wider audience.

3.3. Data

In some respects, misconceptions about data often serve as larger roadblocks to MI adoption than the suitability of the data itself. Successful materials informatics projects have started with as few as 32 data points (Antono et al., 2020). However, many organisations that decide to employ MI mistakenly feel that they need to focus on aggregation of data from historical sources or publicly available databases before they can start MI projects. By instead starting small MI projects with the data already available, organisations can start to demonstrate the value of MI and get researcher buy-in, educate the team on what it can do, and get a better understanding of the optimal data structure they need to build. That is not to say that more data is not better! Comprehensive, high-quality datasets typically perform better than small datasets when utilised by knowledgeable researchers with the right tools. Corporations have a financial incentive to protect their data, as it is often expensive to produce and secrecy may offer them a competitive advantage, but publicly funded academic research projects should be required to acquire, store, and organize data in a way that aligns with the FAIR (findability, accessibility, interoperability, and reusability) principles (GO FAIR, 2021). In this way future researchers will continue to be able to stand on the shoulders of giants.

Early success in materials informatics came from work involving simulation-derived materials property databases, such as the Materials Project (Ong et al., 2013), due to their alignment with FAIR principles (particularly the highly structured and API accessible nature of the database) and the comprehensive, homogenous, and self-consistent nature of their data (Ward et al., 2018). Such datasets across the materials landscape would result in a step-change in the speed of materials research. Projects are underway to aggregate historical, published experimental data and make it widely accessible (Tetko et al., 2016; Dridi et al., 2021). This is challenging work. Data-scraping techniques, image recognition, and other high-throughput digitisation methods are improving, but poor labeling and incomplete metadata make useful datasets hard to come by. Investment in these projects is worthwhile for research funding bodies to ensure the reusability of data which will accelerate projects they fund in the future. Dissemination requirements for publicly funded research should include alignment with the FAIR principles, so that all future data is accessible to the public. There are efforts to encourage implementation of FAIR principles among scientists, including GO FAIR (GO FAIR, 2021), the Research Data Alliance (RDA, 2021), the Materials Research Data Alliance (MaRDA, 2021), and CODATA (CODATA, 2021). Best practices suggested by GO FAIR include storing rich metadata, using open protocols to access the data, and providing clear data usage licenses. There may also be areas of high-priority research where high-throughput testing techniques are now available, and historical data is unreliable, where it makes sense to fund projects to freshly acquire new data, so that reliable, FAIR data can be made available globally to seed important research that will help us meet our sustainability challenges. This is analogous to the Human Genome Project, which has spawned a tsunami of genetic research.

Rather than focus on methods for retrieving historical data, many researchers would be better off focusing on the steps they can take today for acquiring, storing, and disseminating data in a way that makes it searchable, reusable, and machine-readable in the future. Data should be stored in industry-standard machine-readable formats which can be parsed using a broad range of software tools and

https://doi.org/10.1017/dce.2021.19 Published online by Cambridge University Press

operating systems, such as the JSON and CSV formats. While recording data, researchers should pay extra attention to how the data will be interpreted in the future, as data without context becomes meaningless without the knowledge of the person who recorded it. Data context can be provided by including metadata which describes important attributes of the data, including the date of its acquisition, its author, detailed processing parameters or measurement conditions, instrument settings, relevant ambient conditions, and unique identifiers which can be used to distinguish between multiple materials or chemical samples which may have the same formula or other identical characteristics. Finally, data should be stored in a schema that makes it easy to retrieve. This often requires insertion into an established database architecture such as the SQL relational model or hierarchical Mongo database model. The proper use of an established database structure enables rapid searchability and consistent organisation of data, which are critical to success in materials informatics.

4. Summary

We must act quickly to address global challenges like pollution, water security, and climate change. Solutions to these problems will rely on innovative materials and chemicals which enable clean energy generation, highly efficient manufacturing, and novel methods for recycling and repurposing existing materials. The status quo, human intuition-guided trial-and-error methods of the past are not fast enough to solve these challenges on the timeframes in which they are needed. For this reason, new methods must be adopted to accelerate the R&D process and bring new materials and technologies to human society faster than ever before. Materials Informatics provides the suite of tools needed for this acceleration. To enable its swift, widespread adoption, we must support widespread education, cross-discipline communication, and data sharing. These challenges can be addressed by ensuring that

- all publicly funded research has a requirement for sharing data in alignment with FAIR data principles,
- public research organisations have access to modern informatics infrastructure which enables widespread dissemination and analysis of data,
- output data (e.g., raw data from characterisation equipment) is stored in standard machine-readable formats with appropriate metadata that describes context and details so that the data can be reused in the future,
- data is collected and combined according to a consistent, structured schema in established database architectures so that it can be rapidly searched, manipulated, extracted for use in appropriate analysis workflows, and transferred to other schemas and databases as necessary,
- undergraduate students in the physical sciences receive adequate training in relevant data science concepts, and
- professional societies such as ASM and IOM3 offer courses on materials informatics and modern data analysis techniques.

By following these important guidelines, researchers and institutions can improve their ability to incorporate materials informatics techniques in their established materials and R&D processes, ensuring that their products can be developed as rapidly as possible using powerful cutting-edge technologies.

Funding Statement. This work received no specific grant from any funding agency, commercial or not-for-profit sectors.

Author Contributions. All authors wrote the manuscript and approved the final submitted draft.

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

Competing Interests. All the authors are employed by or contracting to Citrine Informatics.

https://doi.org/10.1017/dce.2021.19 Published online by Cambridge University Press

References

Antono E, **Matsuzawa NN**, **Ling J**, **Saal JE**, **Arai H**, **Sasago M and Fujii E** (2020) Machine-learning guided quantum chemical and molecular dynamics calculations to design novel hole-conducting organic materials. *The Journal of Physical Chemistry A 124*(40), 8330–8340. https://doi.org/10.1021/acs.jpca.0c05769.

ArcelorMittal (2021) *10 Sustainable Development Outcomes.* Available at https://corporate.arcelormittal.com/sustainability/our-10-sd-outcomes Accessed 8/11/2021.

ArcelorMittal (2021) *Case Study: Building a Business that Capitalizes on Digital Opportunities.* Available at https://corporate.arcelormittal.com/media/case-studies/building-a-business-that-capitalises-on-digital-opportunities Accessed 8/11/2021.

BASF (2020) *Corporate Strategy.* Available at https://www.basf.com/global/en/investors/calendar-and-publications/factbook/basf-group/strategy.html Accessed 8/11/2021.

BASF (2021) *Our Research and Development.* Available at https://www.basf.com/global/en/who-we-are/innovation/how-we-innovate/our-RnD.html Accessed 8/11/2021.

Cao B, **Adutwum LA**, **Oliynyk AO**, **Luber EJ**, **Olsen BC**, **Mar A and Buriak JM** (2018) How to optimize materials and devices via design of experiments and machine learning: Demonstration using organic photovoltaics. *ACS Nano 12*(8), 7434–7444. https://doi.org/10.1021/acsnano.8b04726.

Childs CM and Washburn NR (2019) Embedding domain knowledge for machine learning of complex material systems. *MRS Communications 9*, 806–820. https://doi.org/10.1557/mrc.2019.90.

Citrine Informatics (2020) *Challenges for AI in Materials Science.* Available at https://citrine.io/wp-content/uploads/2021/04/White-Paper-Challenges-for-AI-in-Materials.pdf Accessed 8/11/2021.

Citrine Informatics (2021) *General Expression of Materials Data Documentation.* Available at https://citrineinformatics.github.io/gemd-docs/ Accessed 8/11/2021.

CODATA (2021) Available at https://codata.org Accessed 8/11/2021.

Dridi A, **Gaber MM**, **Azad RMA and Bhogal J** (2021) Scholarly data mining: A systematic review of its applications. *WIREs: Data Mining Knowledge Discovery 11*, e1395. https://doi.org/10.1002/widm.1395.

European Commission (2020) *Proposal for a Regulation of the European Parliament and of the Council Establishing the Framework for Achieving Climate Neutrality and AMENDING REGULATION (EU) 2018/1999 (European Climate Law) COM/2020/80 Final.* Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1588581905912&uri=CELEX:52020PC0080 Accessed 8/11/2021.

Fraunhofer Institute for Mechanics of Materials (2021) *Materials Science and Engineering Institutions Collaborate on Implementing a Distributed Research Data Infrastructure.* Available at https://www.iwm.fraunhofer.de/en/press/press-releases/02_07_2021_materials_science_engineering_institutions_distributed_research_data_infrastructure.html Accessed 8/11/2021.

GO FAIR (2021) *Fair Principles.* Available at https://www.go-fair.org/fair-principles/ Accessed 8/11/2021.

GO FAIR (2021) *Go Fair Organisation.* Available at https://www.go-fair.org Accessed 8/11/2021.

Hertwich EG (2021) Increased carbon footprint of materials production driven by rise in investments. *Nature Geoscience 14*, 151–155. https://doi.org/10.1038/s41561-021-00690-8.

Hutchinson ML, **Antono E**, **Gibbons BM**, **Paradiso S**, **Ling J. and Meredig B** (2017) Overcoming Data Scarcity with Transfer Learning. Available at https://arxiv.org/abs/1711.05099.

IPCC (2018) *IPCC Special Report Global Warming of 1.5 ºC.* Available at https://www.ipcc.ch/sr15/, https://www.ipcc.ch/sr15/download/ Accessed 8/11/2021

Ling J, **Hutchinson M**, **Antono E**, **Paradiso S and Meredig B** (2017) High-dimensional materials and process optimization using data-driven experimental design with well-calibrated uncertainty estimates. *Integrating Materials and Manufacturing Innovation volume 6*, 207–217.

Madhavan K, **Zentner L**, **Farnsworth V**, **Shivarajapura S**, **Zentner M**, **Denny N and Klimeck G** (2013) nanoHUB.org: Cloud-based services for nanoscale modeling, simulation, and education. *Nanotechnology Reviews 2*(1), 107–117.

MaRDA (2021) *Materials Research Data Alliance.* Available at https://www.marda-alliance.org Accessed 8/11/2021.

Materials Genome Initiative (2021) *About the Materials Genome Initiative.* Available at https://www.mgi.gov/about Accessed 8/11/2021.

Mosavi A, **Rabczuk T and Varkonyi-Koczy AR** (2018) Reviewing the novel machine learning tools for materials design. In Luca D, Sirghi L and Costin C (eds), *Recent Advances in Technology Research and Education. INTER-ACADEMIA 2017*, Advances in Intelligent Systems and Computing, vol. 660. Cham: Springer. https://doi.org/10.1007/978-3-319-67459-9_7.

NFDI-MatWerk (2021) *NFDI-MatWerk Consortia.* Available at https://nfdi-matwerk.de/ Accessed 8/11/2021.

NIST (2021) Bootcamp: Machine Learning for Materials Research Workshop: Machine Learning Quantum Materials. Available at https://www.nist.gov/news-events/events/2018/07/bootcamp-machine-learning-materials-research-workshop-machine-learning Accessed 8/11/2021.

Nitta N, **Wu F**, **Lee JT and Yushin G** (2015) Li-ion battery materials: Present and future. *Materials Today 18*(5), 252–264.

Ong SP, **Richards WD**, **Jain A**, **Hautier G**, **Kocher M**, **Cholia S**, **Gunter D**, **Chevrier VL**, **Persson KA and Ceder G** (2013) Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science 68*, 314–319.

https://doi.org/10.1017/dce.2021.19 Published online by Cambridge University Press

Peerless J, **Sevgen E**, **Edkins S**, **Koeller J**, **Kim E**, **Kim Y**, **Garg A**, **Antono E and Ling J** (2020) Design space visualization for guiding investments in biodegradable and sustainably sourced materials. *MRS Communications 10*(1), 18–24. https://doi.org/10.1557/mrc.2020.5.

RDA (2021) *Research Data Alliance.* Available at https://www.rd-alliance.org Accessed 8/11/2021.

Saal JE, **Oliynyk AO and Meredig B** (2020) Machine learning in materials discovery: Confirmed predictions and their underlying approaches. *Annual Review of Materials Research 50*(1), 49–69.

Society of Petroleum Engineers (2021). Available at https://webevents.spe.org/applied-data-science-and-digital-engineering-curriculum Accessed 8/11/2021. https://www.spe.org/en/training/courses_old/asm

Talirz L, **Kumbhar S**, **Passaro E**, **Yakutovich AV**, **Granata V**, **Gargiulo F**, **Borelli M**, **Uhrin M**, **Huber SP**, **Zoupanos S and Adorf CS** (2020) Materials cloud, a platform for open computational science. *Scientific Data 7*(1), 1–12.

Tetko IV, **Lowe DM and Williams AJ** (2016) The development of models to predict melting and pyrolysis point data associated with several hundred thousand compounds mined from PATENTS. *Journal of Cheminformatics 8*, 2.

Texas A&M University (2021) *Computational Materials Science Summer School.* Available at https://cms3.tamu.edu/ Accessed 8/11/2021.

The Minerals, Metals & Materials Society (TMS) (2019) *Creating the Next-Generation Materials Genome Initiative Workforce*. Pittsburgh, PA: TMS.

U.S. Department of Energy (2021) *Workshop on Artificial Intelligence Applied to Materials Discovery and Design.* Available at https://www.energy.gov/eere/amo/downloads/workshop-artificial-intelligence-applied-materials-discovery-and-design Accessed 8/11/2021.

United Nations (2015) *Sustainable Development Goals.* Available at https://sdgs.un.org/goals Accessed 8/11/2021.

University of Maryland (2021) *Machine Learning for Materials Research Bootcamp.* Available at https://www.nanocenter.umd.edu/events/mlmr-2021/ Accessed 8/11/2021.

Vikström H, **Davidsson S and Höök M** (2013) Lithium availability and future production outlooks. *Applied Energy 110*, 252.

Ward L, **Aykol M**, **Blaiszik B**, **Foster I**, **Meredig B**, **Saal J and Suram S** (2018) Strategies for accelerating the adoption of materials informatics. *MRS Bulletin 43*(9), 683–689. https://doi.org/10.1557/mrs.2018.204.

Ward L, **Dunn A**, **Faghaninia A**, **Zimmermann NER**, **Bajaj S**, **Wang Q**, **Montoya JH**, **Chen J**, **Bystrom K**, **Dylla M**, **Chard K**, **Asta M**, **Persson K**, **Snyder GJ**, **Foster I and Jain A** (2018) Matminer: An open-source toolkit for materials data mining. *Computational Materials Science 152*, 60–69.

Weininger D, **Weininger A and Weininger JL** (1989) SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Sciences 29*(2), 97–101.

Cite this article: Melia HR, Muckley ES and Saal JE (2021). Materials informatics and sustainability—The case for urgency. *Data-Centric Engineering*, 2: e19. doi:10.1017/dce.2021.19