Educating today's industrial workforce to embrace data-driven materials development

Enze Chen,[1,*] Oghoghosa Igbineweka,[2,†] Lenore Kubie,[2] and James S. Peerless[2]

[1] *Department of Materials Science and Engineering, University of California, Berkeley, CA 94720, USA*

[2] *Citrine Informatics, Redwood City, CA 94063, USA*

INTRODUCTION

The transformative power of artificial intelligence (AI) and machine learning (ML) has revolutionized the way people think and operate in virtually every sector of society, including the materials and chemical sciences[1]. This fact is underscored by the immense attention materials informatics (MI) has garnered across academia[2], industry[3], professional societies[4], and government[5], which heralds future innovations enabled by MI.

As with any new technology, however, our ability to execute on this ambitious vision will depend upon our ability to develop our workforce to have the necessary mindsets and skill sets. In academia, rapid developments are underway to enhance MI education with the introduction of new curricula, workshops, and programs[6]. While these initiatives satisfy critical needs in the education of undergraduate and graduate students, most do not address a related challenge: MI training for the *current* workforce in the materials and chemicals industries. We observe a similar hunger for MI knowledge in industry, not only from the scientists who directly interface with these tools, but also from managers and executives who wish to understand if MI is valuable for their organization. The desire to obtain a competitive edge with MI follows from its demonstrated impact in dramatically reducing development time and costs in sectors such as manufacturing, formulations, and energy storage[3]. With the paucity of on-the-job training opportunities in MI, however, it can be challenging for the current workforce to find quality professional development opportunities to advance fluency in informatics skills. Therefore, we see industry workforce development as a critical step on the path toward broadening the impact of informatics-driven materials development that complements existing academic initiatives.

In support of this goal, targeted MI tutorials[7], training courses[8], and workshops[9,10] are starting to be offered to industry. At Citrine, we have also led numerous workshops and training programs for diverse industrial audiences in a wide range of materials domains. As the MI community continues to expand, we believe the community can benefit from sharing best practices for teaching MI concepts, just as recent articles have shared best practices for performing MI research[11]. Herein, we hope to contribute to this important discussion by summarizing several of the lessons we have learned which will be informative for a wide audience, including:

* chenze@berkeley.edu

† oigbineweka@citrine.io

- *Educators and curriculum developers*, who can adapt these suggestions to guide the design of training curricula.

- *Executives and managers*, who can understand the challenges associated with MI and assess whether training is necessary to ensure a smooth implementation.

- *Researchers and scientists*, who can understand the challenges associated with MI and self-assess whether training will increase their fluency in informatics skills.

We begin by identifying three topics that training programs should cover to facilitate adoption of MI in today's workforce, namely: data management, AI/ML capabilities, and change management (Fig. 1). We then outline our vision for a robust training curriculum that applies pedagogical best practices to capitalize on these opportunities to grow our informatics-skilled workforce. We find that these considerations—which are not meant to be exhaustive—help create optimal learning experiences that acknowledge audience needs and empower beginners to adopt MI strategies.

THREE BROAD TOPICS FOR MI TRAINING

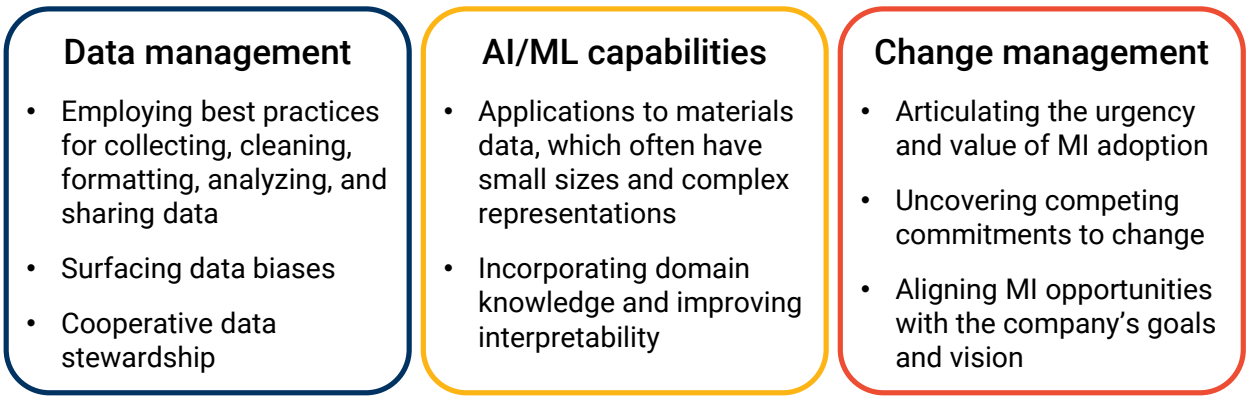

FIG. 1. Three topic areas that should be addressed during training to facilitate MI adoption. For each one, we list a few sub-topics that are specific challenges we see in the materials domain that could lead to greater success if overcome. "AI" is artificial intelligence and "ML" is machine learning.

I. Data management

While audiences are perhaps most excited about building advanced ML models, it is important to remind them that the quality of their models depends on the quality of their data. This relation underscores the importance of *data management*, which we use to refer to the holistic set of practices that includes collecting, storing, cleaning, formatting, analyzing, and sharing data. A strong data foundation relies not only on gathering the raw data, but also recording the associated metadata (e.g., experimental conditions), provenance (e.g., origin and processing history), and uncertainty (e.g., random and systematic variability), which are new practices for many trainees. It can be frustrating for them to learn that they might need to change their data management practices in order to leverage MI. Consequently, it is important to teach the value of this transformation along with best practices to overcome the distinctive challenges when handling data for MI applications (Table I).

Challenge	**Traditional ML**	**MI applications**
Data type	Often standardized (e.g., file types, formatting)	Rarely standardized; often heterogeneous (e.g., TXT, PNG, PDF)
Data volume	Big, dense (up to $\sim 10^8$ examples)	Small, sparse ($\sim 10^2$ examples)
Data bias	Often present (e.g., racial, selection, noise)	Experiments correlated; negative results stigmatized (e.g., unpublished)
Data representation	Straightforward or optimized by algorithms	Requires deep domain knowledge to create useful inputs for ML tasks

TABLE I. Some common data challenges faced in materials informatics (MI) and how they differ from traditional machine learning (ML) applications where "big data" and powerful models are more readily available (e.g., marketing, image classification, language models). Table adapted from ref.[12].

For many industrial teams that embark on MI projects, their data are rarely standardized, leading to inconsistent data formats and file types (often proprietary) across different individuals and teams. Even for something as ubiquitous as spreadsheets the same property might be listed in a single file across several sheets with different names and ambiguous units. These issues are easily overlooked when one teaches with cleaned, demo-ready datasets, but

they present some of the toughest bottlenecks when carrying out MI in practice. Thus, a unified language and standardized practices must be established early on with the help of training. First, it is imperative to communicate to trainees the core principles of data management, which could include the FAIR guiding principles (Findable, Accessible, Interoperable, and Reusable) and key infrastructure capabilities[2]. Second, trainers can choose from several different programmatic tools (e.g., SQL, MongoDB, pandas) to illustrate practical applications, all depending on the audience and desired learning outcomes. There are also nuances in the data that merit consideration: for example, reported data are typically biased toward "positive," high-performing cases, which leaves out a large set of "negative" data (e.g., failed experiments) that can still be valuable for training ML models[12]. It is important during training sessions to communicate the impacts of these biases and share programmatic techniques such as exploratory data analysis and data visualization to examine data quality.

II. AI/ML capabilities

Once the data are organized, it is appropriate to begin exploring *AI/ML capabilities*, which are the machinery for building sophisticated models and extracting deeper insights from data. The outsize impact of AI/ML on accelerating materials development while reducing costs and promoting sustainability has led to a spike in the demand for AI skills and associated upskilling opportunities. While traditional ML courses are useful for learning fundamental concepts, it is essential for trainers to explicitly demonstrate the connection to modeling materials data to help the audience appreciate the nuances of MI and avoid common pitfalls.

To illustrate how training can better calibrate audience expectations, we will briefly discuss challenges regarding data volume and data representation (Table I) and how they affect AI/ML capabilities. In traditional ML tasks, data points can number in the millions, which facilitates training complex models, enables accurate predictions, and mitigates the effects of a few outliers. MI rarely has this luxury, where even a few hundreds of data points might be considered "big," and greater care is required when operating in this "small data" regime. A smaller dataset makes it harder to identify the latent relationships in the data, particularly given the complexities of data representation. The user must choose an

appropriate set of model inputs (e.g., material properties, processing parameters) as model inputs that, ideally, are easy to obtain and to interpret. In this regard, audiences will benefit from learning about software packages that automate the generation of physically-meaningful inputs that are effective for MI tasks[11].

In a similar vein, one solution that we find effective—for training both people and ML models—is the incorporation of materials science *domain knowledge*[12]. Using domain knowledge to guide data collection, featurization, and modeling is critical for MI, especially for small datasets, as it can improve model performance and interpretability. Moreover, we find domain knowledge integration to be effective at engaging audiences as the topics are now in their areas of expertise, making the solutions more relevant and impactful. From these discussions, they learn how to infuse their existing scientific knowledge into the AI workflow and understand the synergy between scientists and AI—that AI is a tool designed to help them, not replace them.

III. Change management

The third topic, which is less technical but equally important, is *change management*, which describes the processes and techniques to manage employees in order to achieve a desired business outcome. The adoption of MI will lead to significant changes at the individual and organizational levels, emphasizing how successful change management is both top-down (promoted by leadership) and bottom-up (initiated by contributors). While change management is largely shaped by company practices and culture, trainers can also encourage this change by honing their messaging around the adoption of MI tools.

For example, before delving into the technical details of how to use MI, it may be helpful to start higher-level and set the stage for this change. First, it is important to impart a sense of urgency around the adoption of new MI methods by describing opportunities for success and how MI impacts the bottom line. Next, present the vision for how MI will drive their company toward the future. Alignment on a vision will both make it easier for employees to understand why they are spending time on training and also make them feel connected to something bigger than themselves. For these discussions, we like to lay out time-boxed goals by leveraging existing company goal-setting and tracking frameworks (e.g., Objectives and Key Results[13]) to engage participants using a familiar terminology and clarify how their

efforts help advance the strategic vision for MI.

Of course, despite an instructor's best intentions, change management can still be met with resistance. It is important to remember that resistant trainees are not being difficult for the sake of being difficult, but rather there may exist a *competing commitment*, which describes an ingrained mindset that prevents an individual from making an inspired change[14]. Competing commitments can materialize in many forms, such as a person's fear of automation taking over their job, feelings of embarrassment when learning a new skill, or disdain for doing more work when "the status quo has always worked." While such conflicts are often surfaced and resolved in conversations between managers and employees, it is equally important for trainers to recognize the competing commitments of their audience in order to maximize the efficacy of MI training. These pernicious barriers may exist for multiple topics presented during training, including data management and AI/ML capabilities, suggesting that change management is best weaved into the curriculum rather than existing as a standalone topic. The repeated emphasis of change management will help shape the proper mindset for data-driven materials development and steer the audience toward successful adoption.

OUTLINE OF AN MI TRAINING CURRICULUM

Cognizant of the aforementioned issues, we now outline a framework for constructing effective training curricula to facilitate adoption of MI. Whether a training coordinator is developing this curriculum for external audiences or internal team members, four major components should be considered: audience, content, assessment, and evaluation.

I. Define the audience

MI requires the entire organization to be on board, but different stakeholders will have different backgrounds and goals. Can both the "materials researcher" and "IT specialist" be trained so the former knows enough programming to be effective contributors and the latter learns enough materials science to build the proper infrastructure? How these individuals will be trained will depend on the answers to questions such as:

- Are the concepts of MI already embedded into their job or will these be entirely new?

- What will they be expected to do with this technology/knowledge?
- How will the implementation of this technology change the way they do their job?

The answers to these questions will serve as the foundations for the subsequent development of training goals and materials. This approach follows the principles of backward design[15], where the curricular activities are developed based on the desired learning outcomes for the target audience. We advise instructors to poll trainees before solidifying a training plan to assess their familiarity with MI concepts and their programming proficiency. This step allows one to better tailor their material to the audience, anticipate user pain points, and maximize the transfer of skills into their audience's daily practice.

II. Select the appropriate content

The delivery of training can come in various forms and the selection of the content type is highly dependent on the defined audience. Below is a description of a few different content types we've used, how they can be applied, and their impact on MI adoption.

- *On-demand video recordings.* Video recordings can be useful because they significantly decrease resource capacity in the long run (e.g., ref[8]). The obvious downside to this method is that if the information being shared is dynamic and will update frequently, the video becomes outdated. We recommended that video recordings are best used for the delivery of conceptual information, technology overview, or anything that is unlikely to change even as the technical implementation evolves over time. On the upside, the ability for users to access this material on demand makes this content type preferred by many trainees.

- *Instructor-led training.* Most people are familiar with this format where one to two trainers demonstrate the material live to the audience. This method is well-suited for complex information because instructors can immediately address follow-up questions to resolve any confusion (e.g., the data management challenges in Table I). An evident disadvantage with this method is scheduling a time that all participants are available, so recording the training session can be a viable alternative. If the training plan calls for an all-/half-day training session, it is important to schedule breaks, lunch, and

energizer activities to keep everyone's focus high throughout the day. The training structure can be chunked into 45- to 60-minute sessions focused on distinct topics, and incorporating interactive elements such as ice breakers, knowledge checks, and small-group discussions will help promote audience engagement and deeper learning.

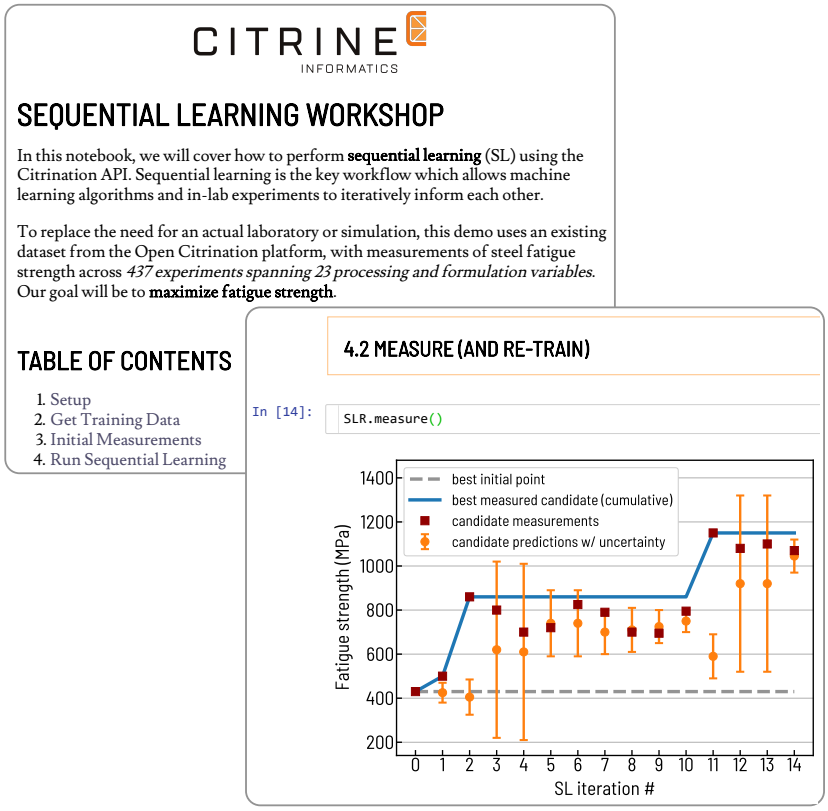

FIG. 2. Representative screenshots of the Jupyter notebook environment from Citrine's training materials.

- *Interactive programming notebooks.* A third type of content featured in Citrine's training materials (see Fig. 2) and work by others[7,10,11] is Jupyter notebooks. These digital notebooks provide interactive programming environments where trainers can design computational narratives that combine text, code, and graphics to provide hands-on experience with MI workflows. They are equally helpful for novice programmers, who will appreciate guidance with the code, and power users, who will experiment with their own ideas that often stimulate deeper discussions. To increase audience engagement, trainers can include discussion questions, missing code blocks, or follow-up exercises that give trainees an opportunity to apply their knowledge and receive immediate feedback. To help mitigate unexpected accessibility issues (e.g., installation,

firewalls, operating systems), we suggest thoroughly testing the activities beforehand and providing clear setup instructions. When thoughtfully designed, these learning activities can be a boon to accessibility as trainees can progress through the notebooks at their own pace and engage with the content in multiple ways.

III. Assess for understanding

Lectures and presentations may be necessary to introduce new concepts, but they should be supplemented with short exercises that review the key concepts and provide in-process formative assessments for learning. Exercises such as multiple-choice questions, free-response questions, or case studies are intended to be completed during the training session so that they can be reviewed, and outstanding questions can be answered. Another consideration is whether the exercises should be completed individually or in small groups. If the technology is one that would ordinarily have multiple business groups involved to execute a task, it is recommended that the exercises mirror that interaction to simulate the realities of the work environment. This alignment will enhance learning and adoption of MI methods in industrial audiences who want to see how these skills can directly transfer to their work.

After each training session is delivered, it is critical that a summative assessment is administered to participants to determine their mastery of the content. In contrast to the exercises, we recommend that these assessments are completed individually post-training in order to accurately identify the gaps in understanding. It could be advantageous to use assessment tools that allow for automatic scoring to provide timely feedback. Another consideration when developing assessments is to ensure that the questions provide signal on user adoption. If the majority of participants are incorrectly answering questions that directly assess their ability to use the new technology with ease, this may indicate the need for a targeted follow-up session or to revise the training materials.

IV. Evaluate the training experience

Every training session should be an opportunity for future improvements to the training program itself, which requires soliciting feedback. At a minimum, a survey should be sent at the conclusion of the training curriculum, but we have also experimented with sending a pre-

training survey or one after each training session. Whichever combination of frequencies is chosen, it is important to consider the length of the surveys, as to not turn off the participants from providing honest and constructive feedback, and the nature of the questions, such that actionable steps can be taken based on the responses. Many of the ideas that we share in this article are distilled from the constructive feedback we have received from our audiences, for which we are grateful.

CONCLUSION AND FUTURE OUTLOOK

As the field of MI rapidly matures, we have both a responsibility and an opportunity to develop robust training programs to cultivate a knowledgeable, modernized, and informatics-skilled workforce in the materials and chemical sciences. To achieve this goal, it is imperative that we take the lessons learned thus far and synthesize a set of best practices that will enhance the quality of future programs. We hope the considerations listed above give facilitators flexibility to adapt them to their own curricula and opportunities to contribute their own ideas toward collective best practices in MI training.

While the scope of this article focuses on industry and the current workforce, many of these ideas can also be incorporated into academic curricula to support the next-generation workforce, whose talents will be highly desired by teams wishing to scale up their pilot MI initiatives. A recent report[4] from The Mineral, Metals & Materials Society identified the need to expand MI education and highlighted collaborative opportunities between academia and industry to sustain the growth of MI. Indeed, through the concerted, complementary efforts of academia, industry, and government in workforce development, the celebrated materials data revolution may prove to be even more fruitful than previously realized.

DECLARATION OF COMPETING INTERESTS

All authors are former or current employees of Citrine Informatics.

AUTHOR CONTRIBUTIONS

All authors planned, drafted, and revised the manuscript.

ACKNOWLEDGMENTS

We thank Malcolm Davidson for assistance with Fig. 2, and Joshua Tappan, Kyle Killebrew, Erin Antono, Edward Kim, and Julia Ling for helpful discussions about training. We thank Bryce Meredig for detailed feedback on the manuscript.

---

[1] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, Machine learning for molecular and materials science, Nature **559**, 547 (2018).

[2] L. Himanen, A. Geurts, A. S. Foster, and P. Rinke, Data-driven materials science: Status, challenges, and perspectives, Advanced Science **6**, 1900808 (2019).

[3] B. Meredig, Industrial materials informatics: Analyzing large-scale data to solve applied problems in R&D, manufacturing, and supply chain, Current Opinion in Solid State and Materials Science **21**, 159 (2017).

[4] The Minerals Metals & Materials Society (TMS), *Creating the Next-Generation Materials Genome Initiative Workforce*, Tech. Rep. (TMS, Pittsburgh, PA, 2019).

[5] National Science and Technology Council, *Materials Genome Initiative Strategic Plan*, Tech. Rep. (Executive Office of the President, National Science and Technology Council, Washington, D.C., 2014) accessed December 2021.

[6] D. L. McDowell, Gaps and barriers to successful integration and adoption of practical materials informatics tools and workflows, JOM **73**, 138 (2021).

[7] A. Strachan and S. Desai, Hands-on Data Science and Machine Learning Training Series, nanoHUB (2020), accessed December 2021.

[8] S. Kalidindi, Materials Data Sciences and Informatics, Coursera (2021), accessed August 2021.

[9] Northwestern University and National Institute of Standards and Technology, Center for Hierarchical Materials Design (CHiMaD) (2020).

[10] National Institute of Standards and Technology, Machine Learning for Materials Research Bootcamp & Workshop on Machine Learning Microscopy Data (2021), accessed December 2021.

[11] A. Y.-T. Wang, R. J. Murdock, S. K. Kauwe, A. O. Oliynyk, A. Gurlo, J. Brgoch, K. A. Persson, and T. D. Sparks, Machine learning for materials scientists: An introductory guide

toward best practices, Chemistry of Materials **32**, 4954 (2020).

[12] Citrine Informatics, Challenges in machine learning for materials - and how to overcome them (2021), accessed December 2021.

[13] J. Doerr, *Measure What Matters* (Portfolio/Penguin, New York, NY, 2018).

[14] R. Kegan and L. Lahey, The real reason people won't change, Harvard Business Review (2001).

[15] G. P. Wiggins and J. McTighe, *Understanding by Design* (Association for Supervision and Curriculum Development, Alexandria, VA, 2005).