

RESEARCH ARTICLE | JUNE 06 2023

## Materials cartography: A forward-looking perspective on materials representation and devising better maps

Steven B. Torrisi ; Martin Z. Bazant ; Alexander E. Cohen; Min Gee Cho ; Jens S. Hummelshøj ; Linda Hung ; Gaurav Kamat ; Arash Khajeh ; Adeesh Kolluru ; Xiangyun Lei ; Handong Ling ; Joseph H. Montoya ; Tim Mueller ; Aini Palizhati; Benjamin A. Paren ; Brandon Phan ; Jacob Pietryga ; Elodie Sandraz ; Daniel Schweigert ; Yang Shao-Horn ; Amalie Trewartha ; Ruijie Zhu ; Debbie Zhuang ; Shijing Sun 



*APL Machine Learning* 1, 020901 (2023)  
<https://doi.org/10.1063/5.0149804>



CrossMark

### Articles You May Be Interested In

Mathematical cartography based on georeferencing maps

*AIP Conference Proceedings* (March 2015)

Simultaneous resistance and capacitance cartography by conducting probe atomic force microscopy in contact mode

*Appl. Phys. Lett.* (March 2005)

Planimetric transformation for the overlap of the ancient Italian cadastral cartography on the current cartographic supports and evaluation of its metric accuracy

*AIP Conference Proceedings* (November 2020)

# Materials cartography: A forward-looking perspective on materials representation and devising better maps

Cite as: APL Mach. Learn. 1, 020901 (2023); doi: 10.1063/5.0149804

Submitted: 8 March 2023 • Accepted: 17 May 2023 •

Published Online: 6 June 2023



























View Online



Export Citation



CrossMark

Steven B. Torrisi,<sup>1,a)</sup>  Martin Z. Bazant,<sup>2,3</sup>  Alexander E. Cohen,<sup>3</sup>  Min Gee Cho,<sup>4</sup>  Jens S. Hummelshøj,<sup>1</sup>  Linda Hung,<sup>1</sup>  Gaurav Kamat,<sup>5</sup>  Arash Khajeh,<sup>1</sup>  Adeesh Kolluru,<sup>6</sup>  Xiangyun Lei,<sup>1</sup>  Handong Ling,<sup>7</sup>  Joseph H. Montoya,<sup>1</sup>  Tim Mueller,<sup>1</sup>  Aini Palizhati,<sup>6</sup>  Benjamin A. Paren,<sup>8</sup>  Brandon Phan,<sup>9</sup>  Jacob Pietryga,<sup>10</sup>  Elodie Sandraz,<sup>10</sup>  Daniel Schweigert,<sup>1</sup>  Yang Shao-Horn,<sup>11,12</sup>  Amalie Trewartha,<sup>1</sup>  Ruijie Zhu,<sup>10</sup>  Debbie Zhuang,<sup>2</sup>  and Shijing Sun<sup>1</sup> 

## AFFILIATIONS

<sup>1</sup>Energy and Materials Division, Toyota Research Institute, Los Altos, California 94022, USA

<sup>2</sup>Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA

<sup>3</sup>Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA

<sup>4</sup>National Center for Electron Microscopy, Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

<sup>5</sup>Department of Chemical Engineering, Stanford University, Palo Alto, California 94305, USA

<sup>6</sup>Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA

<sup>7</sup>Department of Materials Science and Engineering, University of California, Berkeley, Berkeley, California 94720, USA

<sup>8</sup>Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

<sup>9</sup>Department of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, USA

<sup>10</sup>Department of Materials Science and Engineering, Northwestern University, Evanston, Illinois 94305, USA

<sup>11</sup>Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

<sup>12</sup>Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

<sup>a)</sup> Author to whom correspondence should be addressed: [steven.torrisi@tri.global](mailto:steven.torrisi@tri.global)

## ABSTRACT

Machine learning (ML) is gaining popularity as a tool for materials scientists to accelerate computation, automate data analysis, and predict materials properties. The representation of input material features is critical to the accuracy, interpretability, and generalizability of data-driven models for scientific research. In this Perspective, we discuss a few central challenges faced by ML practitioners in developing meaningful representations, including handling the complexity of real-world industry-relevant materials, combining theory and experimental data sources, and describing scientific phenomena across timescales and length scales. We present several promising directions for future research: devising representations of varied experimental conditions and observations, the need to find ways to integrate machine learning into laboratory practices, and making multi-scale informatics toolkits to bridge the gaps between atoms, materials, and devices.

© 2023 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>). <https://doi.org/10.1063/5.0149804>

## I. INTRODUCTION

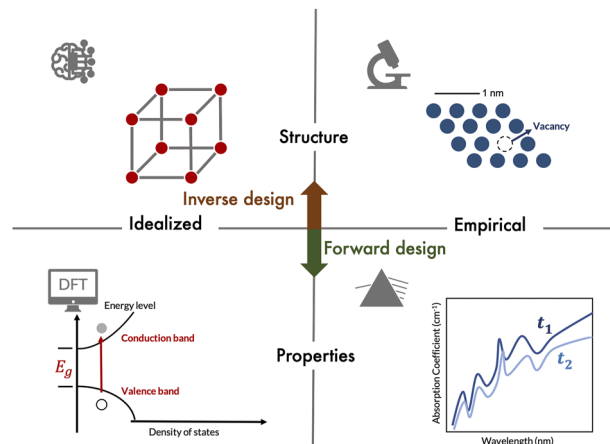
Machine learning (ML) as a tool is here to stay in materials science. Early gains have come from innovative applications of methods from the computer science literature, such as graph-based neural networks or computer vision, to inorganic materials contexts, such as accelerating molecular dynamics, predicting properties, or automating data analysis.<sup>1</sup> As ML and informatics expertise become increasingly mainstream in materials science and engineering, future progress in the field may depend on the integration of scientific domain knowledge into the fundamental building blocks of ML tools, including representations and model architecture.<sup>2</sup>

Within materials science, practitioners are concerned with modeling time and length scales that span many orders of magnitude, presenting differing challenges across the atomistic, mesoscale, and device levels. Moreover, problems where data are scarce<sup>3</sup> pose challenges for applying machine learning in many scientific fields. When designing, training, and applying a model, the representation of input features can be just as important as the target and architecture of the model itself.<sup>3,4</sup> Finding the appropriate way to represent a material of interest is not always straightforward and remains an active area of research. In this Perspective, informed by a recent workshop held within a consortium of industry and academic researchers, we set out to articulate some of the goals and challenges faced by ML practitioners in materials science and propose paths forward focused specifically on materials representation.

We identify at least two broad kinds of supervised machine learning problems in materials science: the forward problem and the inverse problem. Both critically depend on the choice of materials representation, as the representation can be both a means and an end, and a well-chosen representation can simplify demands on the model architecture.

The forward problem is to efficiently and approximately reproduce the results of an experiment (empirical measurements, for example, optical property measurements in Fig. 1) or simulation (idealized abstraction, for example, band structure calculations in Fig. 1) from some knowledge of the material (e.g., structure or composition). The material representation used as input, depending on the available data and the task at hand, can take on any form: ranging from the precise description of the local atomic environment for an interatomic potential to high-level knowledge, such as only the composition itself. This act of mapping from knowledge of the material to a resultant property encompasses the whole of composition- and structure-based property prediction,<sup>5</sup> scaling relations in heterogeneous catalysis,<sup>6</sup> interatomic potentials,<sup>7,8</sup> device lifetime prediction,<sup>9</sup> and the part of an “inverse design” loop, which predicts if a candidate material will be desirable. The inverse problem is to predict the underlying physical attributes of materials that are correlated with material characteristics, such as spectroscopic features.<sup>10,11</sup> Here, a computational representation of a material is simultaneously a means and an end, and the inversion process can map into more-or-less-physically motivated categories, though the problem can be made more challenging when the relationship between the underlying structure and measured output is not 1:1.

Better representations may help to bridge the gap between benchmarks and routine applications of ML in experimental contexts.<sup>4,12</sup> Furthermore, in improving materials representations, the end goals are not just more accurate representation but also (1) transferable ML models and (2) generalizable theories for enhanced



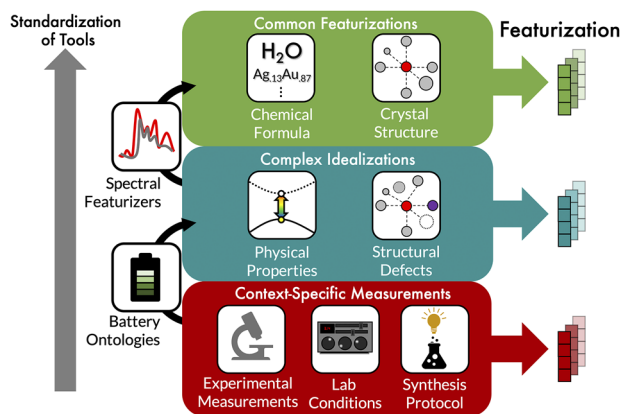
**FIG. 1.** Examples of common forward and inverse problems in materials science, with a focus on structure–property relationship modeling. The forward model maps from knowledge of the material to predictions of consequent properties; the inverse model maps from observations to conceptions of the material congruent with what was measured.

understanding of scientific principles (i.e., knowledge extraction). For (1), presumably all benchmarked models demonstrate acceptable performance on some initial dataset or task, which means that the challenge to demonstrate utility comes from applying them to contexts beyond their initial development. For (2), we note that an example of a useful abstraction that originates from a materials representation is the very notion of periodic atom-containing unit cells that compose crystals: this idealization of the crystal structure is useful for both conceptualizing individual materials and commencing analysis [for instance, matching x-ray diffraction (XRD) patterns with space groups] and therefore enables both understanding and useful predictions.

We structure this perspective around proposing solutions to three identified challenges that researchers may encounter while developing representations of material systems: (A) developing tools to handle the complexity of real-world materials to enable increasing data harvesting and greater interpretability; (B) developing unified representations to combine theory and experimental data sources; and (C) developing representations that can span timescales and length scales. We differentiate between representations suitable for a machine, which we call embeddings (typically vectors of real numbers), and those for a human, which we call idealizations (which follow mathematical and logical structures, such as obeying internally consistent scientific theories). We note that ML can serve as both a guide and a tool to enable the creation of embeddings and idealizations alike.

## II. CHALLENGE A: RICHER DESCRIPTION OF MATERIALS' COMPLEXITY

Understanding the interaction between four key traits of a material: structure, process, property, and performance, is a central focus of modern materials science research. Models that allow us to navigate this complexity effectively might make more experimental systems easily accessible to machine learning methodologies



**FIG. 2.** We vertically order different input data sources according to how well established certain methods of featurization are. Top: Simple idealizations of a materials system, such as the chemical formula or the crystal structure, serve as inputs to established featurization tools or frameworks, such as matminer,<sup>18</sup> featurizing structures is practically an entire subfield.<sup>15</sup> Middle: Complex idealizations of a materials system, such as physical properties of the material that are common across a wider range of systems (such as simple observables about the electronic structure or some knowledge of the defect distribution), are sometimes incorporated into featurization of a material, but standard featurization tools are not in widespread practice yet. Bottom: For systems that have not been well-studied using machine learning, practitioners must make case-by-case decisions on how best to represent their data. This includes synthesis protocol, the set of laboratory conditions that accompanied synthesis, storage, or other miscellaneous measurements. The arrow on the left-hand side represents the development of new tools that can help make featurization of common experimental measurements a standard practice that can be re-used across different projects. For example, growing interest in featurizing spectroscopic data<sup>11,19</sup> may make it more commonplace to featurize spectra for input into machine learning data. For systems such as batteries, where there is a wide variability in the number and kind of measurements that could be made, the emphasis on the community for developing ontologies that can be shared across different systems will help make standards for the field.

and therefore unlock new scientific capabilities (see Fig. 2). In this section, we summarize popular approaches to featurizing data gathered within specific lengths and timescales, such as materials composition and crystal structure, as well as tools to handle convoluted experimental observations that contain information about more than one key traits of materials, such as optical properties or device performance tests. One way to divide the body of recent work on featurizing materials is between those that focus exclusively on the chemical composition,<sup>13,14</sup> those that include some description of the atomistic structure,<sup>15</sup> and those that focus instead on micro- or macro-scale observable properties of the system, including images,<sup>16</sup> spectra,<sup>11</sup> or electrochemical measurements.<sup>17</sup>

Chemical compositions are a simple starting point to represent a material as they are easy to featurize<sup>14,20</sup> and often known in experiments. For input into machine learning models, common approaches include using an element's fractional prevalence within a given composition<sup>21</sup> or as inputs to featurization<sup>14</sup> either internal to a model or via an associated toolkit, such as matminer.<sup>18</sup> We note that when mapping from a chemical composition to an observable property, the composition implicitly encodes structure (more or less depending on the property). This constraint is because of the

fact that all measured properties, of course, rely on some underlying atomic arrangement and that composition–property mappings cannot, in general, be 1:1 without selecting a single structure for each composition (consider the diverse properties presented by pure carbon alone in forms<sup>22</sup> such as graphite, graphene, or diamond).

For the structural representation, we highlight several examples. For input into machine-learning-based models, ample work has been performed on the computational representation of local atomic structures;<sup>23</sup> notably, for use as features in interatomic potential models, we recommend a thorough review from Musil *et al.*<sup>15</sup> In these contexts, the completeness of the descriptor and the computational expense are considerations, which have subsequently given rise to many innovative ideas, such as moment tensor potentials,<sup>24</sup> the atomic cluster expansion,<sup>25</sup> or equivariant descriptors.<sup>26</sup> For structural descriptions, there has also been work centered on graph representations of crystalline materials [e.g., crystal graph convolutional neural networks (CGCNNs)<sup>27–29</sup>] and their applications in predicting site-specific properties.<sup>30</sup>

For atomistic simulations, one typically begins with some prior knowledge of the atomic structure. While macroscopic observables, such as bandgap or surface reactivity, can be very sensitive to individual phases,<sup>31</sup> gaining a detailed mechanistic understanding of the structure–property relationship is challenging because it is experimentally expensive to fully characterize the local atomic structure. This means that representations that correlate with material–property relationships that can sidestep the requirement of full knowledge of the atomic structure are highly desirable. For instance, observables that can give clues to materials structure (e.g., the coordination number of a species in a measured phase)<sup>32,33</sup> can help yield conditions that narrow down the space of possible structures.

A well-chosen representation is itself a tool, as it enables creativity, structured thinking, and useful predictions. An example of this in string serialization of molecules is SMILES<sup>34,35</sup> vs SELFIES,<sup>36</sup> where the latter is purpose-built for traversal of molecule representations in a latent space. Another example is periodic density-functional theory (DFT),<sup>37</sup> where the very idealization of a periodic material as an infinite crystal makes many problems tractable. We note that the computational formulations of representations, such as pymatgen's structure object<sup>38</sup> or ASE's Atoms object,<sup>39</sup> are 21st-century practical advances on an established crystallographic idea in their own right. Making it efficient for researchers to rapidly generate, instantiate, and manipulate these structures on a computer saves thousands of hours of valuable researcher time and enables new feats of cheminformatic and materials informatic work. This capability highlights the serious practical benefits that come from making “human interpretable” idealizations “machine useable.”

## A. Moving forward: Representing disordered systems

A common adage holds that “crystals are like people: it is the defects in them that tend to make them interesting.”<sup>40</sup> There is much interesting work to explore in ML-ready crystal representation beyond representations of average crystal structures. These structures are amenable to methods such as DFT but rely on the idealization of perfect order. The space of defective structures requires serious effort to be able to tractably explore. An idealized single-phase bulk material cannot necessarily contain information that would be germane to experiments, such as the processing history,

if it cannot be captured by defects and disorder in a relatively small unit cell. This detail proves important for observables such as electronic conductivity or catalytic activity,<sup>41</sup> for which very small dopant fractions can play a decisive role in altering the function of a material.<sup>42</sup> Zooming into any real-world material on the atomic scale, it is very likely we would find imperfections in the atomistic ordering. The long-range order of inorganic materials contains countless defects, some by design (e.g., doping in semiconductors),<sup>43</sup> some as a key feature of the material (such as when defects play an entropically stabilizing role in the state),<sup>44</sup> and some by accident, such as thermal strains<sup>45</sup> from unexpected temperature changes.

A recent report by Chen *et al.* demonstrated the machine-learning learned elemental embeddings in materials graph networks to model disorder in materials and the use of multi-fidelity graph neural networks to predict bandgaps.<sup>46</sup> While our focus has been primarily on solid-state systems, we also present a brief case study on how non-solid state disordered solutions such as polymer electrolytes are amenable to novel descriptors, where the dynamics of a polymer system are the object of study. Recent work at Toyota Research Institute<sup>47</sup> has found that representations of trajectories in terms of combined ion clustering and time evolution ion transport properties as a behavior-based descriptor can accelerate molecular dynamics (MD) simulations compared to full MD runs, which also improve the accuracy of predictions compared to other commonly used descriptors, such as SMILES and molecular graphs.<sup>48</sup> A noteworthy feature of this work is that ML efforts that map polymer composition to the result of an MD simulation implicitly capture the full effect of all MD simulation parameters on the outcome. Prediction beyond the set of parameters used to generate the initial dataset—which had common electrolyte composition, temperature, and salt concentration—is simply not possible when the representation is “flattened” into only the identity of the polymer alone. This limitation calls for the development of representations that describe the behavior of the material under study (the polymer matrix) rather than simply the identity of the polymer used in simulation.

## B. Moving forward: Representing processes

The sensitivity of experimental outcomes on processing parameters combined with the expense of data acquisition further challenges the task of evaluating individual materials. Any measurement of a material represents a “snapshot” of its state at a point in time. A holistic record of the time-evolution of a given sample measurement requires knowledge of the full chain of events imposed upon the sample until then: these events could range from individual steps in the synthesis of the sample, a destructive measurement, or even simple storage on the shelf, with each event decorated by descriptive parameters (temperature of sintering and time on shelf). One way to conceptualize this history is via a graph representation, in which a sample is entirely represented by a variable-length series of events, which serves to alter its state in some way. This concept is analogous to the data structure design pattern of event-sourcing, in which a system’s state is described exclusively as an ordered acyclic sequence of changes (git being a common example). Computational tools, such as Aiida<sup>49,50</sup> or StructureNL in pymatgen,<sup>51</sup> are designed to make it easy to navigate the full provenance of parameters, which gave rise to a calculation. For experimental workflows, tools such as DBGen<sup>52</sup> and ESAMP<sup>53</sup> are intended to facilitate data assembly with

this level of exhaustive detail.<sup>54</sup> This form of detailed bookkeeping becoming standard practice could represent an advance in and of itself; more detailed information about the full history of a sample could make it easier to identify causal factors in, e.g., processing that could be decisive during later application. However, the routine use of complete graph-based provenance for experiments and calculations is not yet mainstream.<sup>55</sup> The layer of required overhead when designing a workflow to systematically record every possible state change of the sample may be an inhibiting factor, as well as a lack of common expectation that workflows be documented in exhaustive detail.

## III. CHALLENGE B: UNIFYING REPRESENTATION FOR THEORY AND DIVERSE EXPERIMENTAL DATA SOURCES

Describing the complexity of a scientific challenge or the physical details of a material is complicated by the fact that we are best equipped to think in terms of idealized and abstract representations. We arrive at a central challenge: finding ways to unify representation across theory and diverse experimental data sources (see Fig. 3). Tools that allow researchers to naturally integrate information about experimental data into a representation might allow for more economical use of experimental data and acquisition of more advanced embeddings. Possible directions include combining different modes of input data sources at varying fidelities,<sup>56,57</sup> integration with theoretical representations or theory-generating data,<sup>58</sup> converting theory-generating data into that matches an experiment or vice versa,<sup>59</sup> and principled uncertainty quantification.

Despite differences in the kinds of data from computational and experimental sources, theorists and experimentalists have come up with effective schemas to better communicate with each other. Electron density maps obtained from single-crystal x-ray diffraction and DFT calculations form an example of a unified representation scheme.<sup>60–62</sup> Typical CIFs (crystallographic information files), which have standardized file formats, are universally recognized

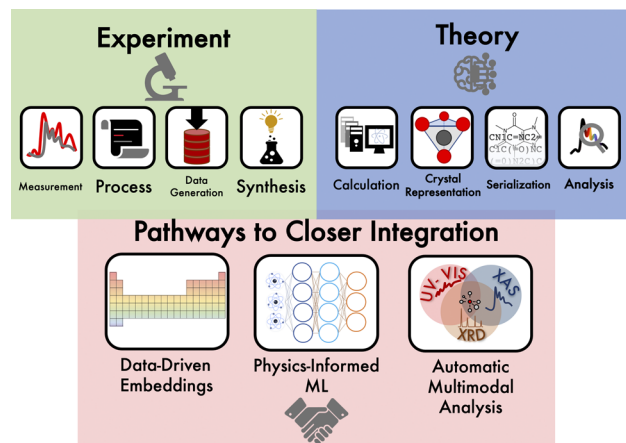


FIG. 3. Pathways toward closer integration of representations for experimental and computational materials.



across the scientific community.<sup>63</sup> CIF files can represent the output of first-principles structure optimization and refinement against single-crystal x-ray diffraction.

Obtaining similar agreements in output formats is much more challenging when it comes to the representation of a material property or device performance. For example, Pourbaix diagrams are widely used as a theoretical guide to deduce thermodynamically stable phases of an aqueous electrochemical system.<sup>64,65</sup> To experimentally detect the degradation of a fuel-cell catalyst,<sup>66</sup> for example, one can measure the concentration of chemical species dissolved in the solvent via techniques such as inductively coupled plasma mass spectrometry,<sup>67</sup> track the deterioration of activity via electrochemical cycling tests,<sup>68</sup> or monitor the microscopic changes via x-ray scattering tomography *in situ*.<sup>69</sup> These three examples of experiments would track events of corrosion and serve as an experimentally measured ground truth to validate a simulated Pourbaix diagram. At the same time, we expect these three measurements to match the theory qualitatively rather than quantitatively: the experimental observations are often a convoluted sum of the property of interests plus the imperfections across scales such as defects, contamination, and the environments in which the tests were conducted, and theory itself has many limitations. In many materials science fields, it remains a challenge to develop “universal languages”: schemas that effectively compare or combine information across multiple sources.

### A. Moving forward: Combining heterogeneous data streams

Moving from an observation to a human-comprehensible representation requires a congruent idealization—a mental picture that agrees with the data. Tools that can automatically combine multiple data streams into a consistent microscopic/microstructural picture, possibly informed by physics, would massively accelerate the process of finding both machine-readable and human-readable representations of data.<sup>70</sup>

Some models are so closely connected to the underlying physics that the idealization comes “for free,” while others require more sophisticated analysis. Combining heterogeneous data sources will require ways to flexibly and automatically combine data from each into a representation. For example, a phenomenon such as EXAFS<sup>71</sup> is well-understood and can be approximated by an equation where individual terms in the equation represent physical quantities in the system [see Eq. (2) of Ref. 72]. This is an example where the model and material representation are implicitly linked, and fitting a good model itself provides insights. X-ray diffraction patterns can be used to establish structural phase conditions that a candidate idealized structure must satisfy. Some forms of characterization can be well-approximated by a closed form expression, such as the EXAFS equation. In these contexts, the act of fitting a model to the data provides readily interpretable features of the material under observation. However, when working with data sources that have nonlinear functional forms such as XANES, the interpretation and mapping causality back to the underlying source are not straightforward, and efforts have been made to craft latent spaces that provide a physical picture for the sake of intermediate representation.<sup>73,74</sup> An experimentalist’s physical or chemical intuition can be used to bridge the gaps among multiple, complementary forms of imaging. This process itself is complicated by epistemic (lack of knowledge)

and aleatoric (random nature of events) uncertainty, as well as the fact that the sample itself can change between measurements or as a result of making a particular measurement.

More flexible, possibly data-driven representations could enable the combination of multi-modal data sources to inform the solution of an ideal material.<sup>75</sup> Toyota Research Institute’s consortium has furthered efforts to make it easier to record state changes within materials, which when multiple data sources are available may make it easier to identify correlations between data sources.<sup>53,76</sup>

### B. Moving forward: Constrained algorithms and flexible theoretical representations

It is challenging to have one form of representation that can mediate among different kinds of measurement, especially when the relationships between the measurements and underlying structures are not easily determined. Representations that could accommodate imprecise knowledge of the underlying structure (that are “fuzzy”) would make it easier to bridge the gap between experiment and theory, such as a physically informed latent space. Data-driven representations of materials that can be rapidly extracted from experimental observables make this possible. Having a concrete idealization that a particular observation will map to (e.g., an XRD pattern revealing a crystal’s space group) necessarily constrains the solution space. This task’s difficulty also depends on the solution space—such as if it has a discrete or a continuous representation or if the fitting process is ill-conditioned. Even when looking within similar systems—molecules—a well-chosen representation of the structure space can enable flexible design, for example, SMILES<sup>34,35</sup> vs SELFIES,<sup>36</sup> where the latter by construction always yields a valid molecule and therefore is a more easily traversable latent space. As more data become available in the materials science community, data-driven spaces could become a viable intermediate space for materials design. For example, Mat2Vec is an example of a space that was derived from literature-based sources,<sup>77</sup> which now sees common use in models such as CrabNet.<sup>78</sup> Furthermore, improved algorithms might flexibly incorporate constraints from experimental observables.<sup>79</sup>

Solution spaces that are designed to be traversable (such as SELFIES) and that also can admit some uncertainty in the underlying structure could have benefits; guesses could be more easily refined in response to new information such as different modes of characterization. In addition, “fuzzy” representations might help to address the issue of noise within experiment and theory. Already, first-principles calculations, such as DFT, owing to their quantum-mechanical and atomistic precision, require idealized unit cells. Forms of representation that describe non-idealized unit cells could aid the interaction between experimentalists and theorists. For instance, compositions are in common use for embedding due to the fact they can represent a material without precise knowledge of the structure.

### C. Moving forward: Improved experimental data generation, collection, and reporting

There are fundamental tensions with the way that ML methods are practiced for training models on large datasets—some work focuses on using available materials databases<sup>80</sup> to train on hundreds or thousands of compounds, but it is expensive to do even

one trial to study one material in great depth experimentally. Crucially, this is contrasted with the high-profile achievements of ML in the commercial software space, where individual trials (e.g., for selling ads) are cost-effective and can be performed at scale. Within fundamental research, one way to rationalize the explosive success of AI in fields such as image recognition/generation<sup>81</sup> and protein folding<sup>82</sup> is the abundance of data available, where the latter is particularly inspiring due to the Protein Data Bank's centrality and importance since 1971. Improving the availability and centrality of data reporting within the materials science community could enable the development of data-driven representations and make it easier to characterize novel materials in light of what has been previously observed by other groups and, thus, to more easily move between modalities of characterization (for example, cross-referencing an XAS measurement made on a particular sample with a database of experimental measurements made in similar systems to gain more structural insight). Over the past few years, an increasing number of open databases of simulated materials structures and properties have been created within the materials community.<sup>83</sup> More recently, several experimental databases<sup>84</sup> and platforms have also become accessible, ranging from functional materials<sup>85</sup> to energy devices.<sup>86</sup>

#### IV. CHALLENGE C: REPRESENTATIONS ACROSS SCALES, FROM MATERIAL TO DEVICE

Scientists are well trained to explain physical phenomena observed using their own eyes by using simpler abstractions as building blocks. For example, when we imagine zooming into a working battery, at a centimeter scale, engineers talk about device architecture and cell design,<sup>87</sup> at a nano- to micrometer scale, materials scientists study degradations using high-resolution microscopy to identify Li dendrite growth,<sup>88</sup> and at an atomic scale, chemists might investigate new crystal structures for a potential cathode material and strive to explain its disordered lattices from electronic structure.<sup>89</sup> Over the hundreds of years of development in modern science, specific languages and models have emerged at each length scale to describe the structures and mechanisms of materials. Challenges arise since models that well represent a material's structure, chemistry, and function at a specific length scale, when zooming out, may only describe what happened locally. Most models, whether a solid sphere model to represent an atom, a SMILES string to represent a molecule,<sup>90</sup> or a crystal graph neural network (e.g., CGCNN) to represent an inorganic material,<sup>28</sup> would have a length scale or timescale limit within which the model would reasonably represent the continuous dimensions in reality.

Materials scientists may be posed to address long-standing difficulties with trying to stitch together representations at different length scales and/or timescales using data-driven methods. Theoretical idealizations on the atomistic level tend to rely on perfect knowledge of the structure and cannot easily integrate real-world timescales and length scales. Device-level models and associated representations come with their own problems depending on the particulars of a given experiment. We may be able to draw inspiration from the multi-scale modeling community, where common representations are used to link individual length scales to the next-larger one, such as coarse-graining atoms or parameterizing individual domains of space; as in Fig. 4, machine learning may make it easier to identify and combine descriptors across length scales. In

recent years, the utilization of ML techniques to extract information from large and diverse datasets, followed by generating abstract representations in the latent space, has garnered substantial attention in the field of materials science. The generated representations can be situated in a shared latent space through the integration of data from various sources, where correlations between individual measurements or calculations are captured by their relative proximity within the space. We shall note that this approach bears resemblance to image-to-text algorithms used in large language models, which likewise rely on a shared latent space to align textual descriptions and visual representations<sup>91</sup>—in one recent case, as many as six modalities.<sup>92</sup>

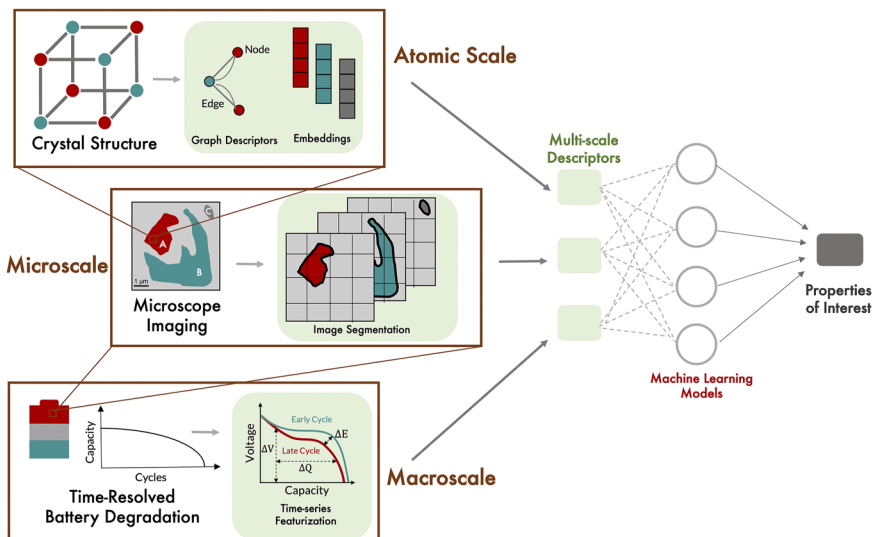
Challenges in representations that can span length scales may originate in the expense of generating datasets, which can be used to capture long timescale and length scale variations in the behavior of a material. Traditional representations are physics-based and thus reflect things we have more easily accessible idealizations of. There are emerging data-driven embeddings<sup>93</sup> for materials and ontologies for devices,<sup>94</sup> which attempt to bridge the gap. Most large datasets correlate theoretical atomic structure and/or compositions with a particular property;<sup>95</sup> the creation of datasets that describe, e.g., battery cycling have helped to enable new fields of informatics work.<sup>96</sup> Closer experiment–theory connection, and unified representations, may help to expand the space of available data and targets.

#### A. Moving forward: Economical models and benchmarking

Most materials design relies on structure–property relationships that live in the same length scale.<sup>97</sup> For example, it is easier to connect electronic structure in a unit cell to properties such as the bandgap than to the compound's ductility. It would be desirable to find ways to decorate idealized bulk structures such that there could be some way to connect defects to the bulk,<sup>98</sup> possibly drawing inspiration from the field of multi-scale modeling. Since defects and short-range phenomena govern so many important performance criteria, benchmarking and accounting for these accurately are key to improving our discovery process.<sup>99,100</sup>

There is still a need to be able to create accurate and especially transferable models from small amounts of data, as accurate materials data (especially experimental data) can be expensive to generate.<sup>101</sup> As the field matures, we expect to see increased use of constraints in features and models to reduce data hunger and therefore increase the scope of applicable problems ML can be applied to, as well as an increasing awareness of the diversity of data-economical models beyond artificial neural networks. Additionally, statisticians have known that simpler models tend to extrapolate better, and incorporating physical and chemical knowledge into model structure may help to simplify the form of models, improving generalizability and efficiency.

One challenge is that it remains unclear how multi- or cross-scale models should be benchmarked against each other and against the prior art. By contrast, there is established work that benchmarks the effectiveness of different models in active discovery against the goal of acceleration. For example, Rohr *et al.*<sup>102</sup> introduced three different metrics: active learning metrics that quantify the discovery of any “good” material, enhancement factors that quantify the improvement of the method introduced (compared to the benchmark) at a given budget, and acceleration factors that quantify the



**FIG. 4.** The goal of combining features from multiple length scales into an end-to-end machine learning framework.

savings in the budget of the method introduced to achieve the same results as the benchmark. However, when it comes to combining representations obtained at multiple timescales or length scales into a single machine learning model to predict materials' behavior, we have a set of questions to answer prior to conducting a new experiment/study. They are as follows: (1) What is the benchmark that we are comparing against? (2) What value have we added using our method compared to the benchmark, specifically, how do we decide which metrics define "success"? (3) How are the materials properties in the lab or simulation connected to the actual device performance in our ML model? New benchmarks enabled by datasets—experimental datasets in particular—would be worthy targets for the community going forward.

## B. Moving forward: Embeddings, proxies, and mesoscale descriptors

Multiple reports recently have discussed the lack of mesoscale models bridging the gap between our understanding at an atomic level to a device level.<sup>103</sup> Indeed, many common systems in materials science lack good physical models to fully explain the complex phenomena, such as interfacial dynamics and microstructural heterogeneity in batteries.<sup>104</sup> Data-driven methods have recently emerged as a way to overcome this challenge of learning in a domain without decent physical models.<sup>105,106</sup> One example to achieve propagation of scientific laws across length scales is through embeddings. Learning the embedding of smaller constituents of a large structure followed by a combination of the embeddings provides a viable way to represent complex materials. For example, combining the learned embeddings of organic linkers<sup>107</sup> and inorganic nodes<sup>23</sup> allows us to describe hybrid organic–inorganic framework materials, such as metal–organic frameworks.<sup>108</sup>

Inside laboratories, low-fidelity proxies are often used when high-fidelity measurements are expensive.<sup>109</sup> One example is the use of color change as a means of representation instead of precise bandgap measurements to track perovskite degradation under

elevated temperatures and humidity.<sup>110</sup> Another example where tailored representations can help bridge the scale gaps is through mesoscale descriptors. Yang and Buehler have recently reported methods correlating the atomic structure with mesoscale crystal structures using large graph neural networks.<sup>98</sup> Through features extracted from microscopic imaging,<sup>16</sup> such as the shape, size, and orientation of grains in a polycrystalline alloy, one can build data-driven models to correlate compositions with the microstructural features under uniform processing conditions and to correlate microstructural features with bulk materials' properties being measured. Here, descriptors that encode microscopic information serve as an intermediate step in assisting the understanding of composition (atomic scale) and property (macroscale) relationships. The field of descriptor engineering is rapidly evolving, benefiting from advancements in high-performance computing.<sup>111</sup> One area of enormous opportunities lies in combining physics and data-driven representations for explainable property predictions, which may have the effect of allowing researchers to discover new empirical laws.<sup>112</sup>

## V. CONCLUSION

In conclusion, toward the goal of improved inverse and forward models, we articulate three central challenges for representation development for ML in materials science: representations that support a richer description of materials' complexity, unifying representations for theory and diverse experimental data sources, and representations that can span multiple timescales and length scales. We emphasize that a significant benefit would be easier integration of ML into regular experimental practice, as we should not lose sight of the fact that machine learning, while an interesting object of study in its own right, still has tremendous untapped potential in enabling better materials science and engineering. In this Perspective, we identify promising directions that have emerged for each of these challenges and hope that this can serve as an inspiration for future researchers engaging with these topics.



## VI. METHODS

The authors facilitated a workshop with around 40 researchers in a joint academic–industrial virtual consortium over two days to share ideas from the cutting edge of the field and solicit viewpoints about the future of material representations. We broke participants into virtual breakout rooms using Zoom and asked them to compile thoughts on their discussion topics into a series of recommendations and challenge statements for the field, which informed the drafting of this Perspective.

## ACKNOWLEDGMENTS

We gratefully acknowledge the contributions of Zhe “Andy” Wang, Huan Tran, Brian Storey, Ahmet Kusoglu, Abraham Anapolsky, Colin Ophus, Santosh Suram, Weike Ye, Shakirul Islam, Madhur Bloor, Tzu-chen Liu, Adolfo Salgado-Casanova, Ashton Aleman, Gavin Winter, Michaela Burke Stevens, Patrick Asinger, Melissa Kreider, Sheng Gong, Yash Samantaray, Hilda Mera, Kumudra Aung, Giacomo Galuppini, Juner Zhu, Shakul Pathak, Avtar Singh, Michael Li, Andrew Lee, Huada Lian, Dohun Kang, and Zachary Ulissi to discussions leading to this publication.

## AUTHOR DECLARATIONS

### Conflict of Interest

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

## Author Contributions

**Steven B. Torrisi:** Conceptualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Martin Z. Bazant:** Conceptualization (supporting). **Alexander E. Cohen:** Conceptualization (supporting). **Min Gee Cho:** Conceptualization (supporting). **Jens S. Hummelshøj:** Conceptualization (supporting). **Linda Hung:** Conceptualization (supporting); Writing – review & editing (supporting). **Gaurav Kamat:** Conceptualization (supporting). **Arash Khajeh:** Conceptualization (supporting); Writing – review & editing (supporting). **Adeesh Kolluru:** Conceptualization (supporting). **Xiangyun Lei:** Conceptualization (supporting). **Handong Ling:** Conceptualization (supporting). **Joseph H. Montoya:** Conceptualization (supporting); Writing – review & editing (supporting). **Tim Mueller:** Conceptualization (supporting); Writing – review & editing (supporting). **Aini Palizhati:** Conceptualization (supporting). **Benjamin A. Paren:** Conceptualization (supporting). **Brandon Phan:** Conceptualization (supporting). **Jacob Pietryga:** Conceptualization (supporting). **Elodie Sandraz:** Conceptualization (supporting). **Daniel Schweigert:** Conceptualization (supporting). **Yang Shao-Horn:** Conceptualization (supporting). **Amalie Trewartha:** Conceptualization (supporting). **Ruijie Zhu:** Conceptualization (supporting). **Debbie Zhuang:** Conceptualization (supporting). **Shijing Sun:** Conceptualization (equal); Writing – original draft (equal); Writing – review & editing (equal).

## DATA AVAILABILITY

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

## REFERENCES

- <sup>1</sup>K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, “Machine learning for molecular and materials science,” *Nature* **559**, 547–555 (2018).
- <sup>2</sup>K. Hippalgaonkar, Q. Li, X. Wang, J. W. Fisher III, J. Kirkpatrick, and T. Buonassisi, “Knowledge-integrated machine learning for materials: Lessons from gameplaying and robotics,” *Nat. Rev. Mater.* **8**, 241 (2023).
- <sup>3</sup>A. D. Sendek, B. Ransom, E. D. Cubuk, L. A. Pellouchoud, J. Nanda, and E. J. Reed, “Machine learning modeling for accelerated battery materials design in the small data regime,” *Adv. Energy Mater.* **12**, 2200553 (2022).
- <sup>4</sup>S. Kong, D. Guevarra, C. P. Gomes, and J. M. Gregoire, “Materials representation and transfer learning for multi-property prediction,” *Appl. Phys. Rev.* **8**, 021409 (2021).
- <sup>5</sup>K. Ryan, J. Lengyel, and M. Shatruk, “Crystal structure prediction via deep learning,” *J. Am. Chem. Soc.* **140**, 10158–10168 (2018).
- <sup>6</sup>P. Schlexer Lamoureux, K. T. Winther, J. A. Garrido Torres, V. Streibel, M. Zhao, M. Bajdich, F. Abild-Pedersen, and T. Bligaard, “Machine learning for computational heterogeneous catalysis,” *ChemCatChem* **11**, 3581–3601 (2019).
- <sup>7</sup>V. L. Deringer, M. A. Caro, and G. Csányi, “Machine learning interatomic potentials as emerging tools for materials science,” *Adv. Mater.* **31**, 1902765 (2019).
- <sup>8</sup>J. Vandermause, S. B. Torrisi, S. Batzner, Y. Xie, L. Sun, A. M. Kolpak, and B. Kozinsky, “On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events,” *npj Comput. Mater.* **6**, 20 (2020).
- <sup>9</sup>K. A. Severson, P. M. Attia, N. Jin, N. Perkins, B. Jiang, Z. Yang, M. H. Chen, M. Aykol, P. K. Herring, D. Fraggedakis *et al.*, “Data-driven prediction of battery cycle life before capacity degradation,” *Nat. Energy* **4**, 383–391 (2019).
- <sup>10</sup>Z. Ren, S. I. P. Tian, J. Noh, F. Oviedo, G. Xing, J. Li, Q. Liang, R. Zhu, A. G. Aberle, S. Sun *et al.*, “An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties,” *Matter* **5**, 314–335 (2022).
- <sup>11</sup>S. B. Torrisi, M. R. Carbone, B. A. Rohr, J. H. Montoya, Y. Ha, J. Yano, S. K. Suram, and L. Hung, “Random forest machine learning models for interpretable x-ray absorption near-edge structure spectrum-property relationships,” *npj Comput. Mater.* **6**, 109 (2020).
- <sup>12</sup>A. Tihoonen, S. J. Cox-Vazquez, Q. Liang, M. Ragab, Z. Ren, N. T. P. Hartono, Z. Liu, S. Sun, C. Zhou, N. C. Incandela *et al.*, “Predicting antimicrobial activity of conjugated oligoelectrolyte molecules via machine learning,” *J. Am. Chem. Soc.* **143**, 18917–18931 (2021).
- <sup>13</sup>R. A. Patel, C. H. Borca, and M. A. Webb, “Featurization strategies for polymer sequence or composition design by machine learning,” *Mol. Syst. Des. Eng.* **7**, 661–676 (2022).
- <sup>14</sup>L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, “A general-purpose machine learning framework for predicting properties of inorganic materials,” *npj Comput. Mater.* **2**, 16028 (2016).
- <sup>15</sup>F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, “Physics-inspired structural representations for molecules and materials,” *Chem. Rev.* **121**, 9759–9815 (2021).
- <sup>16</sup>A. Bruefach, C. Ophus, and M. C. Scott, “Analysis of interpretable data representations for 4D-stem using unsupervised learning,” *Microsc. Microanal.* **28**, 1998–2008 (2022).
- <sup>17</sup>N. H. Paulson, J. Kubal, L. Ward, S. Saxena, W. Lu, and S. J. Babinec, “Feature engineering for machine learning enabled early prediction of battery lifetime,” *J. Power Sources* **527**, 231127 (2022).
- <sup>18</sup>L. Ward, A. Dunn, A. Faghaninia, N. E. R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla *et al.*, “Matminer: An open source toolkit for materials data mining,” *Comput. Mater. Sci.* **152**, 60–69 (2018).

- <sup>19</sup>J. Timoshenko, A. Anspoks, A. Cintins, A. Kuzmin, J. Purans, and A. I. Frenkel, "Neural network approach for characterizing structural transformations by x-ray absorption fine structure spectroscopy," *Phys. Rev. Lett.* **120**, 225502 (2018).
- <sup>20</sup>C. B. Wahl, M. Aykol, J. H. Swisher, J. H. Montoya, S. K. Suram, and C. A. Mirkin, "Machine learning-accelerated design and synthesis of polyelemental heterostructures," *Sci. Adv.* **7**, eabj5505 (2021).
- <sup>21</sup>Y. Zhang, T. C. Peck, G. K. Reddy, D. Banerjee, H. Jia, C. A. Roberts, and C. Ling, "Descriptor-free design of multicomponent catalysts," *ACS Catal.* **12**, 10562–10571 (2022).
- <sup>22</sup>S. Kaciulis, A. Mezzi, P. Calvani, and D. M. Trucchi, "Electron spectroscopy of the main allotropes of carbon," *Surf. Interface Anal.* **46**, 966–969 (2014).
- <sup>23</sup>J. Damewood, J. Karaguesian, J. R. Lunger, A. R. Tan, M. Xie, J. Peng, and R. Gómez-Bombarelli, "Representations of materials for machine learning," *Ann. Rev. Mater. Res.* (published online, 2023).
- <sup>24</sup>A. V. Shapeev, "Moment tensor potentials: A class of systematically improvable interatomic potentials," *Multiscale Model. Simul.* **14**, 1153–1173 (2016).
- <sup>25</sup>R. Drautz, "Atomic cluster expansion of scalar, vectorial, and tensorial properties including magnetism and charge transfer," *Phys. Rev. B* **102**, 024104 (2020).
- <sup>26</sup>U. Schmidt and S. Roth, "Learning rotation-aware features: From invariant priors to equivariant descriptors," in *2012 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2012), pp. 2050–2057.
- <sup>27</sup>T. Xie and J. C. Grossman, "Hierarchical visualization of materials space with graph convolutional neural networks," *J. Chem. Phys.* **149**, 174111 (2018).
- <sup>28</sup>T. Xie and J. C. Grossman, "Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties," *Phys. Rev. Lett.* **120**, 145301 (2018).
- <sup>29</sup>R. E. A. Goodall and A. A. Lee, "Predicting materials properties without crystal structure: Deep representation learning from stoichiometry," *Nat. Commun.* **11**, 6280 (2020).
- <sup>30</sup>R. Maulik and P. Balaprakash, "Site-specific graph neural network for predicting protonation energy of oxygenate molecules," *arXiv:2001.03136* (2019).
- <sup>31</sup>A. Stukowski, "Structure identification methods for atomistic simulations of crystalline materials," *Modell. Simul. Mater. Sci. Eng.* **20**, 045021 (2012).
- <sup>32</sup>T. L. Pham, H. Kino, K. Terakura, T. Miyake, K. Tsuda, I. Takigawa, and H. C. Dam, "Machine learning reveals orbital interaction in materials," *Sci. Technol. Adv. Mater.* **18**, 756 (2017).
- <sup>33</sup>M. R. Carbone, S. Yoo, M. Topsakal, and D. Lu, "Classification of local chemical environments from x-ray absorption spectra using supervised machine learning," *Phys. Rev. Mater.* **3**, 033604 (2019).
- <sup>34</sup>D. Weininger, "Smiles, a chemical language and information system. I. Introduction to methodology and encoding rules," *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
- <sup>35</sup>R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, "Automatic chemical design using a data-driven continuous representation of molecules," *ACS Cent. Sci.* **4**, 268–276 (2018).
- <sup>36</sup>M. Krenn, F. Häse, A. Nigam, P. Friederich, and A. Aspuru-Guzik, "Self-referencing embedded strings (selfies): A 100% robust molecular string representation," *Mach. Learn.: Sci. Technol.* **1**, 045024 (2020).
- <sup>37</sup>R. O. Jones, "Density functional theory: Its origins, rise to prominence, and future," *Rev. Mod. Phys.* **87**, 897 (2015).
- <sup>38</sup>S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, "Python materials genomics (pymatgen): A robust, open-source Python library for materials analysis," *Comput. Mater. Sci.* **68**, 314–319 (2013).
- <sup>39</sup>A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus *et al.*, "The atomic simulation environment—A Python library for working with atoms," *J. Phys.: Condens. Matter* **29**, 273002 (2017).
- <sup>40</sup>C. J. Humphreys, "Stem imaging of crystals and defects," in *Introduction to Analytical Electron Microscopy* (Springer, New York, 1979), pp. 305–332.
- <sup>41</sup>S. R. Ede and Z. Luo, "Tuning the intrinsic catalytic activities of oxygen-evolution catalysts by doping: A comprehensive review," *J. Mater. Chem. A* **9**, 20131–20163 (2021).
- <sup>42</sup>Y. Zhou, K. Neyerlin, T. S. Olson, S. Pylypenko, J. Bult, H. N. Dinh, T. Gennett, Z. Shao, and R. O'Hayre, "Enhancement of Pt and Pt-alloy fuel cell catalyst activity and durability via nitrogen-modified carbon supports," *Energy Environ. Sci.* **3**, 1437–1446 (2010).
- <sup>43</sup>S. C. Erwin, L. Zu, M. I. Hafte, A. L. Efron, T. A. Kennedy, and D. J. Norris, "Doping semiconductor nanocrystals," *Nature* **436**, 91–94 (2005).
- <sup>44</sup>Y.-H. Kye, C.-J. Yu, U.-G. Jong, K.-C. Ri, J.-S. Kim, S.-H. Choe, S.-N. Hong, S. Li, J. N. Wilson, and A. Walsh, "Vacancy-driven stabilization of the cubic perovskite polymorph of CsPbI<sub>3</sub>," *J. Phys. Chem. C* **123**, 9735–9744 (2019).
- <sup>45</sup>G. K. White, "Solids: Thermal expansion and contraction," *Contemp. Phys.* **34**, 193–204 (1993).
- <sup>46</sup>C. Chen, Y. Zuo, W. Ye, X. Li, and S. P. Ong, "Learning properties of ordered and disordered materials from multi-fidelity data," *Nat. Comput. Sci.* **1**, 46–53 (2021).
- <sup>47</sup>A. Khajeh, D. Schweigert, S. Torrisi, L. Hung, B. Storey, and H.-K. Kwon, "Early prediction of ion transport properties in solid polymer electrolytes using machine learning and system behavior-based descriptors of molecular dynamics simulations," *chemRxiv:10.26434* (2022).
- <sup>48</sup>T. Xie, A. France-Lanord, Y. Wang, J. Lopez, M. A. Stolberg, M. Hill, G. M. Lev-erick, R. Gomez-Bombarelli, J. A. Johnson, Y. Shao-Horn, and J. C. Grossman, "Accelerating amorphous polymer electrolyte screening by learning to reduce errors in molecular dynamics simulated properties," *Nat. Commun.* **13**, 3415 (2022).
- <sup>49</sup>G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari, and B. Kozinsky, "AiiDA: Automated interactive infrastructure and database for computational science," *Comput. Mater. Sci.* **111**, 218–230 (2016).
- <sup>50</sup>S. P. Huber, S. Zoupanos, M. Uhrin, L. Talirz, L. Kahle, R. Häuselmann, D. Gresch, T. Müller, A. V. Yakutovich, C. W. Andersen *et al.*, "AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance," *Sci. Data* **7**, 300 (2020).
- <sup>51</sup>S. P. Ong, S. Cholia, A. Jain, M. Brafman, D. Gunter, G. Ceder, and K. A. Persson, "The materials application programming interface (API): A simple, flexible and efficient API for materials data based on representational state transfer (REST) principles," *Comput. Mater. Sci.* **97**, 209–215 (2015).
- <sup>52</sup>M. Statt, K. Brown, S. Suram, L. Hung, D. Schweigert, J. Gregoire, and B. Rohr, "DBgen: A Python library for defining scalable, maintainable, accessible, reconfigurable, transparent (SMART) data pipelines," *chemRxiv:10.33774* (2021).
- <sup>53</sup>M. Statt, B. A. Rohr, K. S. Brown, D. Guevarra, J. S. Hummelshøj, L. Hung, A. Anapolsky, J. Gregoire, and S. Suram, "ESAMP: Event-sourced architecture for materials provenance management and application to accelerated materials discovery," *chemRxiv:14583258* (2021).
- <sup>54</sup>In the natural language processing community, fixed-length embeddings of variable-length sequences represented a major breakthrough; while experimental data does not often exist in sufficiently large quantities in academic contexts to allow for a fully flexible sequence-to-vector to be learned, perhaps industrial-scale laboratories could utilize this concept.
- <sup>55</sup>M. Saad, Y. Zhang, J. Tian, and J. Jia, "A graph database for life cycle inventory using Neo4j," *J. Cleaner Prod.* **393**, 136344 (2023).
- <sup>56</sup>A. Palizhati, S. B. Torrisi, M. Aykol, S. K. Suram, J. S. Hummelshøj, and J. H. Montoya, "Agents for sequential learning using multiple-fidelity data," *Sci. Rep.* **12**, 4694 (2022).
- <sup>57</sup>A. E. Siemenn, Z. Ren, Q. Li, and T. Buonassisi, "Fast Bayesian optimization of needle-in-a-haystack problems using zooming memory-based initialization (ZoMBI)," *Npj Comput. Mater.* **9**, 79 (2023).
- <sup>58</sup>A. Jain, J. Montoya, S. Dwaraknath, N. E. R. Zimmermann, J. Dagdelen, M. Horton, P. Huck, D. Winston, S. Cholia, S. P. Ong, and K. Persson, "The materials project: Accelerating materials design through theory-driven data and tools," in *Handbook of Materials Modeling: Methods: Theory and Modeling* (Springer, Cham, 2020), pp. 1751–1784.
- <sup>59</sup>A. S. Anker, K. T. Butler, M. D. Le, T. G. Perring, and J. Thiyagalingam, "Using generative adversarial networks to match experimental and simulated inelastic neutron scattering data," *Digital Discovery* (published online, 2023).
- <sup>60</sup>G. M. Sheldrick, "SHELXT—Integrated space-group and crystal-structure determination," *Acta Crystallogr., Sect. A: Found. Adv.* **71**, 3–8 (2015).
- <sup>61</sup>A. Kokalj, "XCrySDen—A new program for displaying crystalline structures and electron densities," *J. Mol. Graphics Modell.* **17**, 176–179 (1999).

- <sup>62</sup>A. Kirfel, T. Lippmann, P. Blaha, K. Schwarz, D. F. Cox, K. M. Rosso, and G. V. Gibbs, "Electron density distribution and bond critical point properties for forsterite, Mg<sub>2</sub>SiO<sub>4</sub>, determined with synchrotron single crystal X-ray diffraction data," *Phys. Chem. Miner.* **32**, 301–313 (2005).
- <sup>63</sup>S. R. Hall, F. H. Allen, and I. D. Brown, "The crystallographic information file (CIF): A new standard archive file for crystallography," *Acta Crystallogr., Sect. A: Found. Crystallogr.* **47**, 655–685 (1991).
- <sup>64</sup>K. A. Persson, B. Waldwick, P. Lazic, and G. Ceder, "Prediction of solid-aqueous equilibria: Scheme to combine first-principles calculations of solids with experimental aqueous states," *Phys. Rev. B* **85**, 235438 (2012).
- <sup>65</sup>A. M. Patel, J. K. Nørskov, K. A. Persson, and J. H. Montoya, "Efficient Pourbaix diagrams of many-element compounds," *Phys. Chem. Chem. Phys.* **21**, 25323–25327 (2019).
- <sup>66</sup>M. E. Kreider and M. Burke Stevens, "Material changes in electrocatalysis: An *in situ*/operando focus on the dynamics of cobalt-based oxygen reduction and evolution catalysts," *ChemElectroChem* **10**, e202200958 (2023).
- <sup>67</sup>M. E. Kreider, G. A. Kamat, J. A. Zamora Zeledón, L. Wei, D. Sokaras, A. Gallo, M. B. Stevens, and T. F. Jaramillo, "Understanding the stability of manganese chromium antimonate electrocatalysts through multimodal *in situ* and operando measurements," *J. Am. Chem. Soc.* **144**, 22549–22561 (2022).
- <sup>68</sup>F.-Y. Chen, Z.-Y. Wu, Z. Adler, and H. Wang, "Stability challenges of electrocatalytic oxygen evolution reaction: From mechanistic understanding to reactor design," *Joule* **5**, 1704–1731 (2021).
- <sup>69</sup>I. Martens, A. Vamvakeros, N. Martinez, R. Chattot, J. Pusa, M. V. Blanco, E. A. Fisher, T. Asset, S. Escibano, F. Micoud *et al.*, "Imaging heterogeneous electrocatalyst stability and decoupling degradation mechanisms in operating hydrogen fuel cells," *ACS Energy Lett.* **6**, 2742–2749 (2021).
- <sup>70</sup>D. Unruh, V. S. C. Kolluru, A. Baskaran, Y. Chen, and M. K. Chan, "Theory+ AI/ML for microscopy and spectroscopy: Challenges and opportunities," *MRS Bull.* **47**, 1024 (2023).
- <sup>71</sup>M. Newville, "Fundamentals of XAFS," *Rev. Mineral. Geochem.* **78**, 33–74 (2014).
- <sup>72</sup>J. J. Rehr and R. C. Albers, "Theoretical approaches to x-ray absorption fine structure," *Rev. Mod. Phys.* **72**, 621 (2000).
- <sup>73</sup>P. K. Routh, Y. Liu, N. Marcella, B. Kozinsky, and A. I. Frenkel, "Latent representation learning for structural characterization of catalysts," *J. Phys. Chem. Lett.* **12**, 2086–2094 (2021).
- <sup>74</sup>Z. Liang, M. R. Carbone, W. Chen, F. Meng, E. Stavitski, D. Lu, M. S. Hybertsen, and X. Qu, "Decoding structure-spectrum relationships with physically organized latent spaces," *Phys. Rev. Mater.* **7**, 053802 (2023).
- <sup>75</sup>C. Fare, P. Fenner, M. Benatan, A. Varsi, and E. O. Pyzer-Knapp, "A multi-fidelity machine learning approach to high throughput materials screening," *npj Comput. Mater.* **8**, 257 (2022).
- <sup>76</sup>E. Soedarmadji, H. S. Stein, S. K. Suram, D. Guevarra, and J. M. Gregoire, "Tracking materials science data lineage to manage millions of materials experiments and analyses," *npj Comput. Mater.* **5**, 79 (2019).
- <sup>77</sup>V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, and A. Jain, "Unsupervised word embeddings capture latent knowledge from materials science literature," *Nature* **571**, 95–98 (2019).
- <sup>78</sup>A. Y.-T. Wang, S. K. Kauwe, R. J. Muddock, and T. D. Sparks, "Compositionally restricted attention-based network for materials property predictions," *npj Comput. Mater.* **7**, 77 (2021).
- <sup>79</sup>D. Chen, Y. Bai, W. Zhao, S. Ament, J. Gregoire, and C. Gomes, "Deep reasoning networks for unsupervised pattern de-mixing with constraint reasoning," in *Proceedings of the 37th International Conference on Machine Learning, Proceedings of Machine Learning Research*, edited by H. Daumé III and A. Singh (PMLR, 2020), Vol. 119, pp. 1500–1509.
- <sup>80</sup>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 Mater.* **1**, 011002 (2013).
- <sup>81</sup>S. Chakraborty and K. Mali, "An overview of biomedical image analysis from the deep learning perspective," in *Research Anthology on Improving Medical Imaging Techniques for Analysis and Intervention* (Medical Information Science Reference, 2023), pp. 43–59.
- <sup>82</sup>J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Židek, A. Potapenko *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature* **596**, 583–589 (2021).
- <sup>83</sup>W. Ye, X. Lei, M. Aykol, and J. H. Montoya, "Novel inorganic crystal structures predicted using autonomous simulation agents," *Sci. Data* **9**, 302 (2022).
- <sup>84</sup>A. Zakutayev, N. Wunder, M. Schwarting, J. D. Perkins, R. White, K. Munch, W. Tumas, and C. Phillips, "An open experimental database for exploring inorganic materials," *Sci. Data* **5**, 180053 (2018).
- <sup>85</sup>K. R. Talley, R. White, N. Wunder, M. Eash, M. Schwarting, D. Evenson, J. D. Perkins, W. Tumas, K. Munch, C. Phillips, and A. Zakutayev, "Research data infrastructure for high-throughput experimental materials science," *Patterns* **2**, 100373 (2021).
- <sup>86</sup>T. J. Jacobsson, A. Hultqvist, A. García-Fernández, A. Anand, A. Al-Ashouri, A. Hagfeldt, A. Crovetto, A. Abate, A. G. Ricciardulli, A. Vijayan *et al.*, "An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles," *Nat. Energy* **7**, 107–115 (2022).
- <sup>87</sup>C.-H. Hung, P. Huynh, K. Teo, and C. L. Cobb, "Are three-dimensional batteries beneficial? Analyzing historical data to elucidate performance advantages," *ACS Energy Lett.* **8**, 296–305 (2022).
- <sup>88</sup>T. Foroozan, S. Sharifi-Asl, and R. Shahbazian-Yassar, "Mechanistic understanding of Li dendrites growth by *in situ*/operando imaging techniques," *J. Power Sources* **461**, 228135 (2020).
- <sup>89</sup>A. Urban, A. Abdellahi, S. Dacek, N. Artrith, and G. Ceder, "Electronic-structure origin of cation disorder in transition-metal oxides," *Phys. Rev. Lett.* **119**, 176402 (2017).
- <sup>90</sup>M. Quirós, S. Gražulis, S. Girdzijauskaitė, A. Merkys, and A. Vaitkus, "Using smiles strings for the description of chemical connectivity in the crystallography open database," *J. Cheminf.* **10**, 23 (2018).
- <sup>91</sup>A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," [arXiv:2204.06125](https://arxiv.org/abs/2204.06125) (2022).
- <sup>92</sup>R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "ImageBind: One embedding space to bind them all," [arXiv:2305.05665](https://arxiv.org/abs/2305.05665) (2023).
- <sup>93</sup>C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, "Graph networks as a universal machine learning framework for molecules and crystals," *Chem. Mater.* **31**, 3564–3572 (2019).
- <sup>94</sup>S. Clark, F. L. Bleken, S. Stier, E. Flores, C. W. Andersen, M. Marcinek, A. Szczesna-Chrzan, M. Gaberscek, M. R. Palacin, M. Uhrin, and J. Friis, "Toward a unified description of battery data," *Adv. Energy Mater.* **12**, 2102702 (2022).
- <sup>95</sup>A. Dunn, Q. Wang, A. Ganose, D. Dopp, and A. Jain, "Benchmarking materials property prediction methods: The Matbench test set and Automatminer reference algorithm," *npj Comput. Mater.* **6**, 138 (2020).
- <sup>96</sup>P. M. Attia, A. Grover, N. Jin, K. A. Severson, T. M. Markov, Y.-H. Liao, M. H. Chen, B. Cheong, N. Perkins, Z. Yang *et al.*, "Closed-loop optimization of fast-charging protocols for batteries with machine learning," *Nature* **578**, 397–402 (2020).
- <sup>97</sup>Q. Yao, X. Yuan, T. Chen, D. T. Leong, and J. Xie, "Engineering functional metal materials at the atomic level," *Adv. Mater.* **30**, 1802751 (2018).
- <sup>98</sup>Z. Yang and M. J. Buehler, "Linking atomic structural defects to mesoscale properties in crystalline solids using graph neural networks," *npj Comput. Mater.* **8**, 198 (2022).
- <sup>99</sup>A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," *Phys. Rev. B* **87**, 184115 (2013).
- <sup>100</sup>L. Himanen, M. O. Jäger, E. V. Morooka, F. F. Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster, "DScribe: Library of descriptors for machine learning in materials science," *Comput. Phys. Commun.* **247**, 106949 (2020).
- <sup>101</sup>M. Politi, F. Baum, K. Vaddi, J. Vasquez, E. Antonio, B. P. Bishop, N. Peek, V. C. Holmberg, and L. D. Pozzo, "High-throughput workflow for the synthesis of CdSe nanocrystals using a sonochemical materials acceleration platform," [chemRxiv:10.26434](https://chemrxiv.org/abs/202302.06434) (2023).

- <sup>102</sup>B. Rohr, H. S. Stein, D. Guevarra, Y. Wang, J. A. Haber, M. Aykol, S. K. Suram, and J. M. Gregoire, "Benchmarking the acceleration of materials discovery by sequential learning," *Chem. Sci.* **11**, 2696–2706 (2020).
- <sup>103</sup>A. D. Rollett, G. S. Rohrer, and R. M. Suter, "Understanding materials microstructure and behavior at the mesoscale," *MRS Bull.* **40**, 951–960 (2015).
- <sup>104</sup>O. Borodin, X. Ren, J. Vatamanu, A. von Wald Cresce, J. Knap, and K. Xu, "Modeling insight into battery electrolyte electrochemical stability and interfacial structure," *Acc. Chem. Res.* **50**, 2886–2894 (2017).
- <sup>105</sup>H. Xu, J. Zhu, D. P. Finegan, H. Zhao, X. Lu, W. Li, N. Hoffman, A. Bertei, P. Shearing, and M. Z. Bazant, "Guiding the design of heterogeneous electrode microstructures for Li-ion batteries: Microscopic imaging, predictive modeling, and machine learning," *Adv. Energy Mater.* **11**, 2003908 (2021).
- <sup>106</sup>A. Bhowmik, I. E. Castelli, J. M. Garcia-Lastra, P. B. Jørgensen, O. Winther, and T. Vegge, "A perspective on inverse design of battery interphases using multi-scale modelling, experiments and generative deep learning," *Energy Storage Mater.* **21**, 446–456 (2019).
- <sup>107</sup>S. Chong, S. Lee, B. Kim, and J. Kim, "Applications of machine learning in metal-organic frameworks," *Coord. Chem. Rev.* **423**, 213487 (2020).
- <sup>108</sup>C. Altintas, O. F. Altundal, S. Keskin, and R. Yildirim, "Machine learning meets with metal organic frameworks for gas storage and separation," *J. Chem. Inf. Model.* **61**, 2131–2146 (2021).
- <sup>109</sup>N. Taherimaksousi, M. Fievez, B. P. MacLeod, E. P. Booker, E. Fayard, M. Matheron, M. Manceau, S. Cros, S. Berson, and C. P. Berlinguette, "A machine vision tool for facilitating the optimization of large-area perovskite photovoltaics," *npj Comput. Mater.* **7**, 190 (2021).
- <sup>110</sup>R. Keeseey, A. Tiihonen, A. E. Siemenn, T. W. Colburn, S. Sun, N. T. Putri Hartono, J. Serdy, M. Zeile, K. He, C. A. Gurtner, A. C. Flick, C. Batali, A. Encinas, R. R. Naik, Z. Liu, F. Oviedo, I. M. Peters, J. Thapa, S. I. Parker Tian, and R. H. Dauskardt, "An open-source environmental chamber for materials-stability testing using an optical proxy," *Digital Discovery*, **2**, 422 (2023).
- <sup>111</sup>A. Roy, M. F. N. Taufique, H. Khakurel, R. Devanathan, D. D. Johnson, and G. Balasubramanian, "Machine-learning-guided descriptor selection for predicting corrosion resistance in multi-principal element alloys," *npj Mater. Degrad.* **6**, 9 (2022).
- <sup>112</sup>K. Low, M. L. Coote, and E. I. Izgorodina, "Explainable solvation free energy prediction combining graph neural networks with chemical intuition," *J. Chem. Inf. Model.* **62**, 5457–5470 (2022).