Journal of Applied Physics EDITORIAL scitation.org/journal/jap

Machine learning for materials design and discovery

Cite as: J. Appl. Phys. **129**, 070401 (2021); doi: 10.1063/5.0043300
Submitted: 7 January 2021 · Accepted: 10 January 2021 ·
Published Online: 18 February 2021

Rama Vasudevan,[1,a)] Ghanshyam Pilania,[2] and Prasanna V. Balachandran[3,4]

AFFILIATIONS

[1]The Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA
[2]Materials Science and Technology Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA
[3]Department of Materials Science and Engineering, University of Virginia, Charlottesville, Virginia 22904, USA
[4]Department of Mechanical and Aerospace Engineering, University of Virginia, Charlottesville, Virginia 22904, USA

Note: This paper is part of the special collection on Machine Learning for Materials Design and Discovery.
[a)]**Author to whom correspondence should be addressed:** rvv@ornl.gov

I. INTRODUCTION

We are excited to present this Special Topic collection on *Machine Learning for Materials Design and Discovery* in the Journal of Applied Physics. With a wide range of exciting and insightful contributions, we anticipate that this timely issue will provide further impetus to the quickly growing field of materials informatics.

While the history of machine learning (ML) and its applications to physical sciences can be traced back more than a half century to the early days of computers [and starting with the introduction of Bayesian- and simple artificial neural network (ANN)-based methods],[1] the last decade has witnessed a paradigm shift in the ways fundamental and applied problems in physics, materials science, and other related fields have traditionally been pursued.[2–5] Unlike the conventional approach that relied heavily on chemical intuition perfected by laborious trial-and-error based optimization cycles for novel materials discovery, the data-enabled route offers a much more efficient and targeted approach to complete the feedback loop underlying the scientific process.[6] More specifically, the informatics-based tools enable one to connect and iterate through different steps of the materials design and optimization process, including hypothesis generation, prediction, synthesis, characterization, and testing. A number of factors within the data-hardware-software-algorithm ecosystem have contributed to this revolutionary transformation. Most definitely, a sustainable and exponential growth in our ability to generate, share, and store large amounts of data, easily accessible computational resources, open source software, synergistic development of computation hardware, and several algorithmic breakthroughs that enabled efficient learning methods to analyze the available *Big Data* have played a dominant role in the advent of the modern informatics era.[2]

This Machine Learning Special Topic collection presents a representative sample of the latest ML related research being pursued within the broader physics and materials communities. Since the editorial is designed for both experts and novices in the field, before going into the details of specific challenges addressed in each individual contribution, in Sec. II, we provide a brief background on various ML and statistical learning methods. In Sec. III, we classify the contributions included in the Special Topic into four broad categories: (i) materials and molecular property predictions, (ii) materials modeling and simulations, (iii) materials design, discovery, and active learning, and (iv) materials characterization and imaging applications. For each of these groups, we then briefly survey the contributing studies while emphasizing the technical challenges addressed by each. Figure 1 shows a *wordcloud* representation of the most frequently observed words in the abstract of the contributed articles. Table I presents an overview of all the articles available within the Special Topic, highlighting the specific learning methods used, nature of the training datasets employed, and a brief description of the problem addressed therein. We believe that it will serve as a useful resource for the reader to quickly identify the most relevant articles of their interest while browsing through this collection.

II. BACKGROUND

In this section, we will provide a background discussing the salient characteristics of a materials informatics approach and the methods used to accomplish the tasks.[7–10] One of the first steps involve building a *dataset* that is representative of the problem that one is intending to address. A dataset will typically have inputs and output(s). Let $\{\mathbf{X}_i, Y_i\}_{i=1}^{n}$ represent the dataset over $\mathbb{R}$, where $\mathbf{X}_i$ is the matrix of material features (or descriptors) and Y_i is the

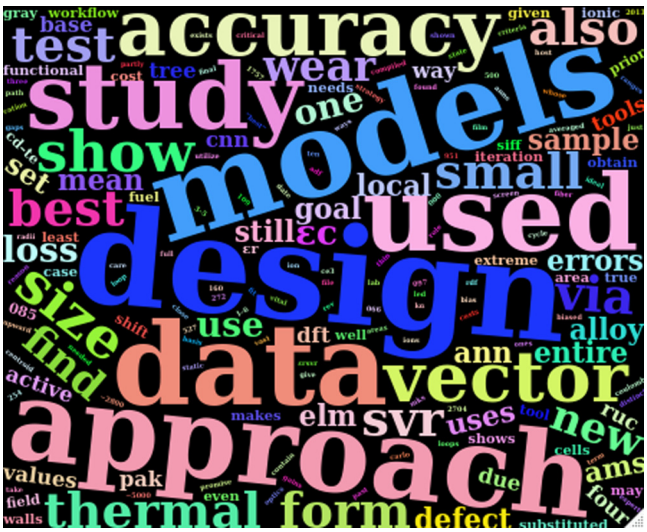

FIG. 1. A wordcloud showing some of the most frequently observed words in the abstract of the contributed articles in the Special Topic on Machine Learning for Materials Design and Discovery.

materials property. The output Y_i can be a vector or a matrix ($\mathbf{Y}_i$) depending on whether the problem formulation is single- or multi-objective. The quantity n is the total number of datapoints in the dataset. The source for $\{\mathbf{X}_i, Y_i\}$ can be experiments, simulations, or combinations of both. One of the key objectives in materials informatics is to transform the raw data of known materials into testable hypotheses.

After database construction, and prior to learning, *pre-processing* is a necessary intermediary step that serves to prepare the data for model building.[11] One of the common reasons for pre-processing arises from the fact that the materials databases contain attributes of different units and scales. The role of pre-processing is to remove the unintended bias, so that all attributes are treated on equal footing. However, it should be noted that the applicability of data normalization methods is algorithm dependent.

Once the problem statements and the boundary conditions are defined, the next task is *learning*. Broadly speaking, there are two types of learning schemes: supervised learning and unsupervised learning.[12,13] The objective of *supervised learning* is to map $\mathbf{X}_i$ to Y_i, f:$\mathbf{X}_i \mapsto Y_i$. Typically, the mapping, f, is represented in the form of IF-THEN-ELSE rules, hierarchical tree structures, mathematical formulas, or a black-box. The trained model f can then be used to *predict* the property ($\hat{Y}$) for any given $\mathbf{X}$ (that may or may not be in the training dataset). The supervised learning can further be subdivided into two categories. They include: (i) regression and (ii) classification learning. When the property of interest (Y_i) is a numerical quantity, such as yield strength or melting point, then regression-based methods are well suited. On the other hand, when Y_i is a categorical quantity, such as space group of a crystalline material or crystal structure-type, then classification learning methods are better suited for supervised learning. Some of the common ML methods used for supervised learning include Naïve Bayes, k-nearest neighbor (k-NN), decision trees, kernel ridge regression (KRR), random forest regression (RFR), gradient boosting regression (GBR), Gaussian process regression (GPR), support vector machine (SVM), and ANNs.

More recently, there have been a growing body of materials informatics publications focused on deep learning methods.[14,15] Deep learning methods, in contrast to some of the above-mentioned shallow learning methods, can extract features automatically from raw data with little or no pre-processing. On the flip side, more often than not, a deep learning method will require a large training dataset compared to the shallow learning methods. In a shallow ANN method, we have an input layer, one hidden layer, and an output layer. Stringing together many hidden layers will give rise to the "deep neural networks" (DNNs). Modern DNNs can have dozens of layers and can include more complex layers than the simple picture discussed here. For instance, convolutional neural networks (CNNs) scan a window across the input to learn spatial features, relevant for imaging applications (e.g., microscopy), but the essence of their operation is the same. Modern deep learning methods can contain millions of trainable parameters.

One of the emerging themes in the use of supervised learning in materials science for accelerated materials design and discovery is to iteratively guide experiments and computations via Bayesian optimization (also referred to as sequential learning, adaptive learning, or active learning in various contexts). This approach has two key steps: (1) the iterative nature of the learning process that includes a feedback loop and a model update and (2) the introduction of utility or acquisition functions for optimal experimental design. The utility function assesses each unexplored data point in the search space and recommends the most promising data point (that satisfies a well-defined constraint) for the next step of validation and feedback. The input to the utility functions are the outcomes from a supervised learning method. One of the popular supervised learning methods often explored in Bayesian optimization is the GPR method, because it has the capability to quantify prediction uncertainties for each observation. Some of the common utility functions explored in the literature include upper confidence bound, probability of improvement, efficient global optimization, knowledge gradient, and mean objective cost of uncertainty.[16–18]

The key difference between a supervised learning and *unsupervised learning* is that one has no *a priori* information on the predefined target variable (Y_i) in the unsupervised learning scheme. The objective of unsupervised learning is to *assign* a label to each data point in ($\mathbf{X}_i$) without the explicit knowledge of the target variable. This approach is used to find correlations and similarities in datasets and to detect anomalous or outlier data points. The outcome is typically represented in the form of data clustering. Data-dimensionality reduction and clustering constitute the two main workhorses of unsupervised learning schemes. Some of the algorithms used for data-dimensionality reduction include principal component analysis (PCA), multi-dimensional scaling (MDS), Isomap, and t-distributed stochastic neighbor embedding (t-SNE), to name a few. Some of the common clustering algorithms include k-means clustering and hierarchical clustering. Exploratory data analysis can also be associated with unsupervised learning, where the objective is to perform visualization and parametric

TABLE I. Contributions included in the Special Topic at a glance. Category keys are listed at the end of the table. The learning algorithms and the nature of the datasets (experimental and/or simulated) used for training the models are identified for each study. Studies that employed density functional theory (DFT) simulations are explicitly indicated.

Reference	Category[a]	ML method(s)	Dataset	Brief description of the topic(s) covered
Kalidindi[24]	1	Tutorial GPR	Simulated and experimental	Foundational concepts for materials knowledge systems framework
Velli *et al.*[25]	1	Multiple (k-NN, SVM, GBR, etc.)	Simulated and experimental	Effect of laser-based processing parameters on materials' structure
Vanpoucke *et al.*[26]	1	Linear and regularized regression	Simulated and experimental	Ensemble average model performance in learning with small datasets
Zhang *et al.*[27]	1,3	GPR	DFT	Formation enthalpy prediction for binary and ternary intermetallics
Honrao *et al.*[28]	1	SVM	DFT	Configurational representations for formation enthalpy prediction
Huang and Ling[29]	1	Multiple (NN, RFR, etc.)	DFT	Inorganic compounds' formation energy prediction
Magedov *et al.*[30]	1	NN	DFT	Bond order prediction in molecules
Sharma *et al.*[31]	1	RFR	DFT	ML for substitutional defect formation energies in perovskites
Sadat and Wang[32]	1	NN, RFR	Simulated	Bandgap prediction in phononic crystals
Chen *et al.*[33]	1	NN	Simulated	Thermo-mechanical response prediction for unidirectional composites
Parker *et al.*[34]	1	Multiple (clustering, classification, and regression)	Simulated	Structure–property relationships of Pt nanoparticles for catalysis
Zhuo *et al.*[35]	1	Extreme GBR	Experimental	Prediction of 5*d* level centroid shift for Ce doped inorganic phosphors
Costine *et al.*[36]	1	MDS, k-NN, RF	Experimental	Growth performance prediction in transition metal dichalcogenide monolayers
Lightstone *et al.*[37]	1	GPR	Experimental	Polymer refractive index prediction
Sahaluddin *et al.*[38]	1	SVM	Experimental	Density estimation of nitride nanofluids in ethylene glycol
Alade *et al.*[39]	1	SVM, NN	Experimental	ML-based viscosity model of nanofluids
Gurgenc *et al.*[40]	1	NN, SVM	Experimental	Wear loss predictions for spray coated magnesium alloys
Alade *et al.*[41]	1	SVR	Experimental	Prediction of lattice parameters for A_2XY_6 cubic crystals
Jacobs *et al.*[42]	1,2	RFR	Simulated	Detection of delamination failure in composite materials
Santos *et al.*[43]	1,2	NN, extreme GBR and ridge regression	Simulated	Thermal insulating performance prediction for multi-component refractory ceramics
Zagaceta *et al.*[44]	2	NN	DFT	Spectral neural network potentials for Ni–Mo alloys
Mangold *et al.*[45]	2	NN	DFT	Phonons and thermal conductivity predictions in Mn–Ge compounds
Zeledon *et al.*[46]	2	NN and GBR	DFT	ML potential development with information-optimized feature representations
Mazhnik and Oganov[47]	3	NN	DFT	ML-based screening of super hard materials
Dieb *et al.*[48]	3	NN	Simulated	Inverse design of depth-graded multilayer structures for x-ray optics

TABLE I. (*Continued.*)

Reference	Category[a]	ML method(s)	Dataset	Brief description of the topic(s) covered
Zheng *et al.*[49]	3	Gauss–Bayesian model	Simulated and experimental	Metamaterials design for high sound absorption at low frequencies
Tian *et al.*[50]	3	SVM	Experimental	Role of uncertainty estimation in efficient active learning
Ma *et al.*[51]	4	Clustering and NN	Experimental	Linking microstructure to processing conditions in uranium–molybdenum alloys
Ziatdinov *et al.*[52]	4	GPR	Experimental	Gaussian processes to achieve super-resolution in contact Kelvin Probe Force Microscopy
Vasudevan *et al.*[53]	4	Model selection	Experimental	Bayesian inference in band excitation scanning probe microscopy for dynamic imaging
Scheinker and Pokharel[54]	4	NN	Simulated	ML-based reconstruction of three-dimensional crystals from diffraction data
Ciobanu *et al.*[55]	4	Image detection	Experimental	Characterization of TiO_2 layered-nanotube scanning electron microscopy images

[a]Category keys: (1) materials and molecular property predictions, (2) materials modeling and simulations, (3) materials design, discovery, and active learning, and (4) materials characterization and imaging applications.

(or non-parametric) statistical testing to better understand the dataset before applying any supervised learning methods.

In addition to supervised and unsupervised learning methods, there are also rapidly growing applications of semi-supervised learning,[19] transfer learning,[20] multifidelity learning,[21] representation learning,[22] and natural language processing[23] in the materials science domain. Table I provides an overview of various learning algorithms used by the contributions included in the Machine Learning Special Topic, further details of which are discussed in Sec. III.

III. SUMMARY OF AREAS COVERED

This section surveys the areas covered by the articles contributing to our Special Topic. We have organized the papers into four broad categories, as highlighted in Table I. This enables us to not only describe the breath of problems that benefit form ML but to also compare approaches used in different studies that try to address similar challenges. In what follows, we highlight and briefly discuss contributions in each of these four broad classes.

A. Materials and molecular property predictions

Ability to accurately approximate a complex and unknown function given just a subset of relevant data lies at the heart of any ML model building exercise. Unlike a rigid analytical expression, ML-based surrogate models are flexible and dynamic in nature. They can evolve to provide better predictive performance as more data become available. Traditional property prediction approaches for molecules and solids rely on either direct measurements or quantum mechanical-based first principles simulations, which can be extremely demanding in terms of time and resources. On the other hand, once validated and tested rigorously, ML-based surrogates can be extremely efficient in predicting the functional relationships that they approximate. Therefore, development of surrogate models for fast yet reliable prediction of processing–structure–property–performance linkages in molecular and materials systems has been a major theme in materials informatics lately, which is also reflected by the contributions in the Special Topic.

Closely aligned with this theme, in Ref. 24, Kalidindi presents a pedagogical tutorial describing the foundational concepts underlying a ML-based surrogate model development approach, referred to as the materials knowledge system framework, for capturing the process–structure–property relationships in materials over varying length scales. Since numerical representations of materials at different length scales are a crucial component of such models, a particular emphasis is provided on feature engineering of the material structure at different resolutions. The tutorial also discusses a strategy that allows for a seamless fusion of experimental- and simulation-based materials data within a Bayesian framework to further accelerate the pace of materials innovation. Along the same lines, Velli *et al.*[25] present another example that highlights integration of experimental and simulation data to improve predictive performance of a ML model aimed at mapping the processing parameters in laser-based manufacturing onto the observed material structure.

An important issue frequently encountered in materials property predictions concerns with the training dataset size. Owing to the high cost of accurate data generation, practical materials design problems are typically restricted to relatively small datasets. Under such conditions, one needs to be extremely careful in making statistical inferences and any predictions should be subjected to a

rigorous uncertainty quantification process. Highlighting this issue, Vanpoucke *et al.*[26] suggest that data-limited situations can be particularly benefited from the use of ensemble-averaged ML modes. Using specific examples that employ either experimental or synthetic data, it is argued that ensemble ML models can provide robust predictions within a reasonable accuracy in small dataset learning problems.

Other contributions within this category employ either computational[27–29,31–34] or experimental[35,37–41] datasets to learn a diverse set of properties for a wide range of materials. Zhang *et al.*,[27] Honrao *et al.*,[28] and Huang and Ling[29] consider the problem of learning formation enthalpy of solids using different ML algorithms. Magedov *et al.* use deep learning to address prediction of bond order in organic molecules.[30] Sharma *et al.*[31] employ a density functional theory (DFT) dataset to learn A- and B-site substitutional dopant formation energies in ABO_3 perovskites. Considering a wide range of dopants and different host perovskites, the study identified dopant's ionic size, elemental heat of formation, and oxidation state as the most important factors toward predicting the substitutional dopant formation energetics. Sadat and Wang[32] explore the use of ML to screen materials with a "phononic bandgap." They show that a trained RF-based model was able to predict a finite bandgap crystal structure with a remarkable 89% accuracy compared to a random selection success of only 17%. Chen *et al.*[33] use a simulated thermo-mechanical response data for unidirectional composites to train and validate a deep convolution NN model, which can predict the thermomechanical response of new samples within a relative prediction error of less than 8% of the physics-based finite-volume micromechanics model. In a different study aimed at learning catalytic performance of platinum nanoparticles toward oxygen reduction, hydrogen oxidation, and hydrogen evolution reactions, Parker *et al.*[34] used an open dataset of ordered and disordered platinum nanoparticles simulated using molecule dynamics to develop a classification model, which was followed by class-specific mappings of structure/property relationships for each class. Among many physically meaningful and interesting findings, the study showed that the disordered particles perform better for hydrogen evolution and hydrogen oxidation reactions if the particles are small, while the ordered particles perform better if the {110} surface area is increased.

The contributions utilizing experimental datasets have focused on problems that range from predicting optical properties arising due to the host dielectric screening-dependent placement of the Ce $5d$ states of in Ce-activated inorganic phosphors[35] to growth performance prediction in transition metal dichalcogenide monolayers[36] and from refractive index prediction in polymers[37] to density[38] and viscosity[39] estimation for nanofluids, wear loss prediction in alloy coatings,[40] and lattice parameter predictions in given class of crystal chemistries.[41] These studies encouragingly demonstrate that ML models built on carefully chosen (i.e., domain-knowledge-driven selection) features can be remarkably useful for property predictions even when available training data are limited.

B. Materials modeling and simulations

In addition to enabling cheaper surrogate models for materials properties, ML algorithms have been quite successful in expediting, enhancing, and completing traditional domain-specific modeling and simulation capabilities. The last decade has witnessed a tremendous amount of activity in the field of data-driven atomistic simulations (in particular, in the area of ML forcefields development) to push the frontiers of accessible accuracy, time, and length scales in these simulations. Besides learning interatomic interactions, ML has also been used in conjunction with physics-based simulations to combine information coming in from different sources. For instance, measurements or simulations at different length scales or at the same length scale, but with varying levels of fidelities. An excellent example demonstrating this aspect is presented by Jacobs *et al.*[42] in the Special Topic. They employ two different approaches, namely, the finite element model based mode curvature test and the natural frequency test, to detect delamination failure in a composite material. While the former approach is computationally demanding, it can provide an excellent performance in both identifying and localizing delamination. The latter approach, on the other hand, is simple and inexpensive to conduct but can only identify the presence of delamination and not the location. However, after augmenting the natural frequency test with a ML model allowed for both localization of damage as well as quantification of its severity. In a different contribution, Santos *et al.*[43] combine finite element method with ML to expedite predictions of insulating thermal behavior in multi-component refractory ceramics. The authors demonstrate that physics-based simulations for only a small subset of 2.8% of the total 1.9×10^5 insulating candidates are required to reliably estimate the thermal performance of all insulating system possibilities.

Several contributions in the Special Topic have focused on the ML-based potential development challenge. Zagaceta *et al.*[44] present a numerical implementation of the atom-centered representations introduced earlier by Bartók *et al.*[56] and subsequently apply this to develop ML-based interatomic potentials for binary Ni–Mo alloys for large scale simulations. Mangold *et al.*[45] develop and validate a transferable neural network ML potential for germanium manganese compounds—a class of materials exhibiting a variety of stable and metastable phases with different stoichiometries and interesting electronic, magnetic, and thermal properties—that can successfully reproduce structural and thermal behavior of the systems with different local chemical environments. Further continuing along this theme, Zeledon *et al.*[46] present a feature-engineered approach that emphasizes on optimizing physically relevant information storage within the feature representations and decoupling of the representation (or feature vector) dimensionality from the size of the materials structure to be modeled. The learning performance of the proposed structural information filtered features (SIFFs) potential is demonstrated on several datasets consisting of molecules, clusters, and periodic solids.

C. Materials design, discovery, and active learning

Further building on efficient property predictions, one or more ML-based surrogate models can be used to explore, design, and screen promising materials candidates, for a further in-depth analysis, given an application. Most commonly adopted approaches for ML-based materials screening have resorted to combinatorial enumeration, inverse design, and active learning-based strategies.

As an example of enumeration-based screening strategy, Zhang *et al.*[27] use a ML model developed to predict formation enthalpy of intermetallic compounds. The model trained on a set of binary intermetallics was applied to an enumerated set of all possible ternary intermetallics falling within the domain of applicability of the model to screen potentially formable novel compounds. Mazhnik and Oganov[47] trained a graph-based neural network model on elastic data from available from the Materials Project database[57] to develop efficient and general models for predicting hardness and fracture toughness in compounds. This model was then applied to screen all crystal structures in the database to identify a number of potentially interesting new superhard materials, while confirm that diamond and its polytypes are indeed the hardest materials in the database.

Unlike combinatorial enumeration where property predictions are made on the entire set of compounds, an inverse design strategy aims to identify compounds given desired properties. This further requires a coupling of the ML model with an optimization routine (such as evolutionary algorithms, minima-hopping, or swarm optimization-based routines) to help explore the structure–property landscape. The Special Topic highlights studies from Dieb *et al.*[48] and Zheng *et al.*[49] that employ an inverse design approach to address interesting functional materials design challenges. Dieb *et al.*[48] use Monte Carlo tree search with policy gradient reinforcement learning method in combination with a simulated reflectivity dataset to design depth-graded multilayer structures (also known as supermirrors) for x-ray optics applications. Zheng *et al.*[49] take up the challenge of designing acoustic metamaterials that facilitate high sound absorption at low frequencies. The design employs a typical acoustic metamaterial absorber with multiple structural parameters that are optimized via adopting adaptive ML framework guided by a physics-based Gauss–Bayesian model. The final results of the high absorption performance were further verified by explicit numerical simulations and experimental measurements.

Active learning adopts an adaptive design procedure with a feedback loop where predictions using a current ML model are used to guide the data collection in the next iteration to further improve the model in terms of its domain of applicability and predictive accuracy. An active learning framework heavily relies on uncertainty quantification in model's mean predictions and a judiciously selected acquisition function that prioritizes the decision-making process on unseen data using model uncertainties. Further exploring this aspect, Tian *et al.*[50] discuss the nuances of statistical uncertainty estimation using approaches based on bootstrapping and jackknife-based estimators in relation to its vital role in accelerating materials development via active learning.

D. Materials characterization and imaging applications

Owing to recent advances in atomically resolved materials characterization techniques such as scanning transmission electron microscopy, scanning tunneling microscopy, and atomic force microscopy, electron backscatter diffraction imaging, today state-of-the-art materials imaging, and characterization techniques can provide high quality data on functional materials in large quantities for both static and dynamic conditions. On the other hand, advances in ML-based image recognition methods have opened up new avenues for not only efficient on-the-fly analysis of materials characterization data to help address workflow bottlenecks in materials development but also to extract knowledge from materials characterization data with physics-informed models. Furthermore, ML models can also be used to efficiently interpolate and augment the acquired data to construct two- and three-dimensional maps of the material being probed. A number of contributions in the special topic have been devoted to the broad theme of materials characterization and imaging applications.

Using an example of uranium–molybdenum nuclear fuel alloy, Ma *et al.*[51] consider the problem of ML a mapping between the microstructure images and processing conditions. Employing a new microstructure representation, their approach was shown to score an excellent classification performance, with an F1 score of 95.1%, in distinguishing between micrographs corresponding to ten different thermo-mechanical material processing conditions. Ziatdinov *et al.*[52] employed a Gaussian process regression-based extrapolation scheme, across both the spatial and parameter space, to enhance resolution for the contact Kelvin probe force microscopy technique. The authors also provide their methodology implementation as an interactive Google Colab notebook that goes through all the details of the analyses presented as a part of this study. Vasudevan *et al.*[53] tackle the challenge of ML assisted reliable data interpretation in complex scanning probe microscopy measurements by taking up a specific example of Bayesian inference approach for the analysis of the image formation mechanisms in a ferroelectric thin film. It is shown that the Bayesian framework allows for the incorporation of extant materials knowledge and allows for model selection.

A yet another exciting contribution is presented by Scheinker and Pokharel[54] that deals with the problem of three-dimensional reconstruction of crystals from coherent diffraction imaging. Coherent diffraction imaging is a nondestructive x-ray imaging technique that records only the intensity of the complex diffraction pattern originating from the sample volume to provide sub-nm atomic displacement estimates in the specimen. However, since the phase information is lost in a coherent diffraction imaging experiment, construction of the underlying 3D electron density given the diffraction data is a challenge and conventional iterative numerical methods are expensive. Scheinker and Pokharel[54] put forward a new approach that utilizes spherical harmonics based representation of crystals along with convolutional neural networks to address this reconstruction challenge. Finally, Ciobanu *et al.*[55] present a computer-vision-based morphological analysis technique to assist in and expedite high-resolution scanning electron microscopy image analysis.

IV. CONCLUSIONS

In the last decade, use of ML and related techniques for materials design and development has matured in a main stream field from a narrow and highly specialized area of interdisciplinary research. After discussing the current status of the field and a brief background on various commonly used ML techniques, this editorial presented a brief overview of the papers included in our Special Topic on Machine Learning for Materials Design and Discovery. The collection presents a representative snapshot of

the latest ML related research that is currently being pursued within the community. We hope that this editorial will serve as a useful guide to the exciting and informative content presented in the Special Topic. and the collection of papers will be enjoyed by the scientific community.

AUTHORS' CONTRIBUTIONS

All the guest editors contributed equally to this editorial.

ACKNOWLEDGMENTS

G.P. acknowledges funding from the Laboratory Directed Research and Development (LDRD) program of the Los Alamos National Laboratory (LANL) via multiple projects (Nos. 20190001DR and 20190043DR). LANL is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (DOE) (Contract No. 89233218CNA000001). A portion of this work was performed at Oak Ridge National Laboratory's Center for Nanophase Materials Sciences, which is a U.S. DOE Office of Science User Facility (R.V.). We thank our colleagues and collaborators for making this Special Topic possible. We also sincerely thank the staff and editors of the Journal of Applied Physics for their assistance in putting this Special Topic together.

REFERENCES

[1]A. M. Turing, "Computing machinery and intelligence," Mind **59**(236), 433 (1950).
[2]D. Morgan and R. Jacobs, "Opportunities and challenges for machine learning in materials science," Annu. Rev. Mater. Res. **50**, 71–103 (2020).
[3]R. Batra, L. Song, and R. Ramprasad, "Emerging materials intelligence ecosystems propelled by machine learning," Nat. Rev. Mater. (published online 2020).
[4]J. Schmidt, M. R. Marques, S. Botti, and M. A. Marques, "Recent advances and applications of machine learning in solid-state materials science," npj Comput. Mater. **5**(1), 1–36 (2019).
[5]K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," Nature **559**(7715), 547–555 (2018).
[6]R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, "Machine learning in materials informatics: Recent applications and prospects," npj Comput. Mater. **3**(1), 1–13 (2017).
[7]K. Rajan, "Materials informatics: An introduction," in *Informatics for Materials Science and Engineering* (Elsevier, 2013), pp. 1–16.
[8]O. Isayev, A. Tropsha, and S. Curtarolo, *Materials Informatics: Methods, Tools, and Applications* (John Wiley & Sons, 2019).
[9]G. Pilania, P. V. Balachandran, J. E. Gubernatis, and T. Lookman, "Data-based methods for materials design and discovery: Basic ideas and general methods," Synth. Lect. Mater. Opt. **1**(1), 1–188 (2020).
[10]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," Chem. Mater. **32**(12), 4954–4965 (2020).
[11]S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining* (Springer, 2015).
[12]T. M. Mitchell, *Machine Learning* (McGraw-Hill Higher Education, New York, 1997).
[13]J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, Springer Series in Statistics Vol. 1 (Springer, New York, 2001).
[14]A. Agrawal and A. Choudhary, "Deep materials informatics: Applications of deep learning in materials science," MRS Commun. **9**(3), 779–792 (2019).
[15]B. Nikolay, N. Sabine, S. V. Kalinin, O. S. Ovchinnikova, R. K. Vasudevan, and S. Jesse, "Deep neural networks for understanding noisy data applied to physical property extraction in scanning probe microscopy," npj Comput. Mater. **5**(1), 25 (2019).
[16]T. Lookman, P. V. Balachandran, D. Xue, and R. Yuan, "Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design," npj Comput. Mater. **5**(1), 1–17 (2019).
[17]P. V. Balachandran, "Adaptive machine learning for efficient materials design," MRS Bull. **45**(7), 579–586 (2020).
[18]J. E. Saal, A. O. Oliynyk, and B. Meredig, "Machine learning in materials discovery: Confirmed predictions and their underlying approaches," Annu. Rev. Mater. Res. **50**, 49–69 (2020).
[19]H. Huo, Z. Rong, O. Kononova, W. Sun, T. Botari, T. He, V. Tshitoyan, and G. Ceder, "Semi-supervised machine-learning classification of materials synthesis procedures," npj Comput. Mater. **5**(1), 1–7 (2019).
[20]E. D. Cubuk, A. D. Sendek, and E. J. Reed, "Screening billions of candidates for solid lithium-ion conductors: A transfer learning approach for small data," J. Chem. Phys. **150**(21), 214701 (2019).
[21]G. Pilania, J. E. Gubernatis, and T. Lookman, "Multi-fidelity machine learning models for accurate bandgap predictions of solids," Comput. Mater. Sci. **129**, 156–163 (2017).
[22]C. J. Court, B. Yildirim, A. Jain, and J. M. Cole, "3D inorganic crystal structure generation and property prediction via representation learning," J. Chem. Inf. Model. **60**, 4518–4535 (2020).
[23]E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, and E. Olivetti, "Materials synthesis insights from scientific literature via text extraction and machine learning," Chem. Mater. **29**(21), 9436–9444 (2017).
[24]S. R. Kalidindi, "Feature engineering of material structure for AI-based materials knowledge systems," J. Appl. Phys. **128**(4), 041103 (2020), ISSN: 0021-8979, 1089-7550.
[25]M. C. Velli, G. D. Tsibidis, A. Mimidis, E. Skoulas, Y. Pantazis, and E. Stratakis, "Predictive modeling approaches in laser-based material processing," arXiv:2006.07686 (2020).
[26]D. E. P. Vanpoucke, O. S. J. van Knippenberg, K. Hermans, K. V. Bernaerts, and S. Mehrkanoon, "Small data materials design with machine learning: When the average model knows best," J. Appl. Phys.**128**(5), 054901 (2020), ISSN: 0021-8979 1089-7550.
[27]Z. Zhang, M. Li, K. Flores, and R. Mishra, "Machine learning formation enthalpies of intermetallics," J. Appl. Phys. **128**(10), 105103 (2020), ISSN: 0021-8979, 1089-7550.
[28]S. J. Honrao, S. R. Xie, and R. G. Hennig, "Augmenting machine learning of energy landscapes with local structural information," J. Appl. Phys. **128**(8), 085101 (2020), ISSN: 0021-8979, 1089-7550.
[29]L. Huang and C. Ling, "Practicing deep learning in materials science: An evaluation for predicting the formation energies," J. Appl. Phys. **128**(12), 124901 (2020), ISSN: 0021-8979, 1089-7550.
[30]S. Magedov, C. Koh, W. Malone, N. Lubbers, and B. Nebgen, "Bond order predictions using deep neural networks," J. Appl. Phys. **128**(1), 064701 (2020), ISSN: 0021-8979, 1089-7550.
[31]V. Sharma, P. Kumar, P. Dev, and G. Pilania, "Machine learning substitutional defect formation energies in ABO_3 perovskites," J. Appl. Phys. **128**(3), 034902 (2020), ISSN: 0021-8979, 1089-7550.
[32]S. M. Sadat and R. Y. Wang, "A machine learning based approach for phononic crystal property discovery," J. Appl. Phys. **128**(2), 025106 (2020), ISSN: 0021-8979, 1089-7550.
[33]Q. Chen, W. Tu, and M. Ma, "Deep learning in heterogeneous materials: Targeting the thermo-mechanical response of unidirectional composites," J. Appl. Phys. **127**(17), 175101 (2020), ISSN: 0021-8979, 1089-7550.
[34]A. J. Parker, G. Opletal, and A. S. Barnard, "Classification of platinum nanoparticle catalysts using machine learning," J. Appl. Phys. **128**(1), 014301 (2020), ISSN: 0021-8979, 1089-7550.

[35]Y. Zhuo, S. Hariyani, S. You, P. Dorenbos, and J. Brgoch, "Machine learning 5D-level centroid shift of Ce^{3+} inorganic phosphors," J. Appl. Phys. **128**(1), 013104 (2020), ISSN: 0021-8979, 1089-7550.
[36]A. Costine, P. Reinke, and P. V. Balachandran, "Data-driven assessment of chemical vapor deposition grown MoS_2 monolayer thin films," J. Appl. Phys. **128**(1), 235303 (2020), ISSN: 0021-8979, 1089-7550.
[37]J. P. Lightstone, L. Chen, C. Kim, R. Batra, and R. Ramprasad, "Refractive index prediction models for polymers using machine learning," J. Appl. Phys. **127**(21), 215105 (2020), ISSN: 0021-8979, 1089-7550.
[38]M. Sahaluddin, I. O. Alade, M. O. Oyedeji, and U. S. Aliyu, "A machine learning-based model to estimate the density of nanofluids of nitrides in ethylene glycol," J. Appl. Phys. **127**(20), 205105 (2020), ISSN: 0021-8979, 1089-7550.
[39]I. O. Alade, M. A. A. Rahman, A. Hassan, and T. A. Saleh, "Modeling the viscosity of nanofluids using artificial neural network and Bayesian support vector regression," J. Appl. Phys. **128**(8), 085306 (2020a), ISSN: 0021-8979, 1089-7550.
[40]T. Gurgenc, O. Altay, M. Ulas, and C. Ozel, "Extreme learning machine and support vector regression wear loss predictions for magnesium alloys coated using various spray coating methods," J. Appl. Phys. **127**(18), 185103 (2020), ISSN: 0021-8979, 1089-7550.
[41]I. O. Alade, I. A. Olumegbon, and A. Bagudu, "Lattice constant prediction of A_2 XY_6 cubic crystals (A = K, Cs, Rb, TI; X = tetravalent cation; Y = F, Cl, Br, I) using computational intelligence approach," J. Appl. Phys. **127**(1), 015303 (2020b), ISSN: 0021-8979, 1089-7550.
[42]E. W. Jacobs, C. Yang, K. G. Demir, and G. X. Gu, "Vibrational detection of delamination in composites using a combined finite element analysis and machine learning approach," J. Appl. Phys. **128**(12), 125104 (2020), ISSN: 0021-8979, 1089-7550.
[43]D. P. Santos, P. I. B. G. B. Pelissari, R. F. de Mello, and V. C. Pandolfelli, "Estimating the thermal insulating performance of multi-component refractory ceramic systems based on a machine learning surrogate model framework," J. Appl. Phys. **127**(21), 215104 (2020), ISSN: 0021-8979, 1089-7550.
[44]D. Zagaceta, H. Yanxon, and Q. Zhu, "Spectral neural network potentials for binary alloys," J. Appl. Phys. **128**(4), 045113 (2020), ISSN: 0021-8979, 1089-7550.
[45]C. Mangold, S. Chen, G. Barbalinardo, J. Behler, P. Pochet, K. Termentzidis, Y. Han, L. Chaput, D. Lacroix, and D. Donadio, "Transferability of neural network potentials for varying stoichiometry: Phonons and thermal conductivity of Mn_xGe_y compounds," J. Appl. Phys. **127**(24), 244901 (2020), ISSN: 0021-8979, 1089-7550.
[46]J. A. H. Zeledon, A. H. Romero, P. Ren, X. Wen, Y. Li, and J. P. Lewis, "The structural information filtered features (SIFF) potential: Maximizing information stored in machine-learning descriptors for materials prediction," J. Appl. Phys. **127**(21), 215108 (2020), ISSN: 0021-8979, 1089-7550.
[47]E. Mazhnik and A. R. Oganov, "Application of machine learning methods for predicting new superhard materials," J. Appl. Phys. **128**(7), 075102 (2020), ISSN: 0021-8979, 1089-7550.
[48]S. Dieb, Z. Song, W.-J. Yin, and M. Ishii, "Optimization of depth-graded multilayer structure for x-ray optics using machine learning," J. Appl. Phys. **128**(7), 074901 (2020), ISSN: 0021-8979, 1089-7550.
[49]B. Zheng, J. Yang, B. Liang, and J.-C. Cheng, "Inverse design of acoustic metamaterials based on machine learning using a Gauss–Bayesian model," J. Appl. Phys. **128**(13), 134902 (2020), ISSN: 0021-8979, 1089-7550.
[50]Y. Tian, R. Yuan, D. Xue, Y. Zhou, X. Ding, J. Sun, and T. Lookman, "Role of uncertainty estimation in accelerating materials development via active learning," J. Appl. Phys. **128**(1), 014103 (2020), ISSN: 0021-8979, 1089-7550.
[51]W. Ma, E. J. Kautz, A. Baskaran, A. Chowdhury, V. Joshi, B. Yener, and D. J. Lewis, "Image-driven discriminative and generative machine learning algorithms for establishing microstructure–processing relationships," J. Appl. Phys. **128**(13), 134901 (2020), ISSN: 0021-8979, 1089-7550.
[52]M. Ziatdinov, D. Kim, S. Neumayer, L. Collins, M. Ahmadi, R. K. Vasudevan, S. Jesse, M. Hyun Ann, J. H. Kim, and S. V. Kalinin, "Super-resolution and signal separation in contact Kelvin probe force microscopy of electrochemically active ferroelectric materials," J. Appl. Phys. **128**(5), 055101 (2020), ISSN: 0021-8979, 1089-7550.
[53]R. K. Vasudevan, K. P. Kelley, E. Eliseev, S. Jesse, H. Funakubo, A. Morozovska, and S. V. Kalinin, "Bayesian inference in band excitation scanning probe microscopy for optimal dynamic model selection in imaging," J. Appl. Phys. **128**(5), 054105 (2020), ISSN: 0021-8979, 1089-7550.
[54]A. Scheinker and R. Pokharel, "Adaptive 3D convolutional neural network-based reconstruction method for 3D coherent diffraction imaging," arXiv:2008.10094 (2020).
[55]A. Ciobanu, M. Luca, C. T. Konrad-Soare, G. Stoian, and D. Luca, "Computer-aided detection and morphological characterization of nanotube layers using scanning electron microscopy images," J. Appl. Phys. **127**(10), 105102 (2020), ISSN: 0021-8979, 1089-7550.
[56]A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," Phys. Rev. B **87**(18), 184115 (2013).
[57]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**(1), 011002 (2013).