ACS Partner Journal

ACCOUNTS of materials research

pubs.acs.org/amrcda

Article

Interpretable and Explainable Machine Learning for Materials Science and Chemistry

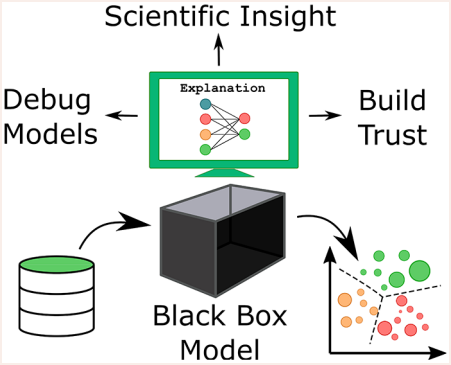
Felipe Oviedo,*,# Juan Lavista Ferres, Tonio Buonassisi, and Keith T. Butler*,#

Cite This: *Acc. Mater. Res.* 2022, 3, 597−607

Read Online

ACCESS | Metrics & More | Article Recommendations

CONSPECTUS: Machine learning has become a common and powerful tool in materials research. As more data become available, with the use of high-performance computing and high-throughput experimentation, machine learning has proven potential to accelerate scientific research and technology development. Though the uptake of data-driven approaches for materials science is at an exciting, early stage, to realize the true potential of machine learning models for successful scientific discovery, they must have qualities beyond purely predictive power. The predictions and inner workings of models should provide a certain degree of explainability by human experts, permitting the identification of potential model issues or limitations, building trust in model predictions, and unveiling unexpected correlations that may lead to scientific insights. In this work, we summarize applications of interpretability and explainability techniques for materials science and chemistry and discuss how these techniques can improve the outcome of scientific studies. We start by defining the fundamental concepts of interpretability and explainability in machine learning and making them less abstract by providing examples in the field. We show how interpretability in scientific machine learning has additional constraints compared to general applications. Building upon formal definitions in machine learning, we formulate the basic trade-offs among the explainability, completeness, and scientific validity of model explanations in scientific problems. In the context of these trade-offs, we discuss how interpretable models can be constructed, what insights they provide, and what drawbacks they have. We present numerous examples of the application of interpretable machine learning in a variety of experimental and simulation studies, encompassing first-principles calculations, physicochemical characterization, materials development, and integration into complex systems. We discuss the varied impacts and uses of interpretabiltiy in these cases according to the nature and constraints of the scientific study of interest. We discuss various challenges for interpretable machine learning in materials science and, more broadly, in scientific settings. In particular, we emphasize the risks of inferring causation or reaching generalization by purely interpreting machine learning models and the need for uncertainty estimates for model explanations. Finally, we showcase a number of exciting developments in other fields that could benefit interpretability in material science problems. Adding interpretability to a machine learning model often requires no more technical know-how than building the model itself. By providing concrete examples of studies (many with associated open source code and data), we hope that this Account will encourage all practitioners of machine learning in materials science to look deeper into their models.



INTRODUCTION

"Where is the knowledge we have lost in information?"
The lamentation on the modern condition in the opening stanza of T. S. Eliot's *The Rock* could just as appropriately, if more prosaically, be used to summarize much of the scepticism of scientists toward machine learning (ML) as applied to traditional scientific subjects. Specifically, scientists are often skeptical about the lack of insight in many publications applying ML. However, is scientific knowledge inevitably lost in machine learning studies? If not how can it be extracted and how does this apply to machine learning in the context of scientific research? In this Account, we set out to provide some answers, illustrated by the work of ourselves and others, to these questions.

There have already been numerous efforts to build a taxonomy of interpretability and explainability methods for machine learning models. Some noteworthy examples that we draw upon for explaining classical ML models and deep neural networks are refs 1−6. In the Key Concepts section, we provide a brief overview of some of the concepts that will be most important for following the rest of this Account, but we recommend these references for readers interested in learning more.

Received: December 15, 2021
Revised: May 11, 2022
Published: June 3, 2022

https://doi.org/10.1021/accountsmr.1c00244
Acc. Mater. Res. 2022, 3, 597−607

In this Account, we draw on the experience of the authors in using machine learning for understanding and guiding experiments and enhancing theory and simulation to give a broad overview of how interpretability and explainability have a role to play across the materials science disciplines. We cover a range of methods, starting from building inherently interpretable models and then introducing techniques for extracting explanations and interpretations from models. Many of the methods we highlight for interpretability are easily implemented using existing software, meaning that we believe that there is no reason why ML applied to materials science should remain a black box. Moreover, we also consider some frontiers in interpretable ML models. We can retrieve and perhaps even expand the knowledge latent in the vast amounts of information currently available in materials science. In addition to this promise, we discuss the potential roadblocks and particular challenges for machine learning interpretability in the field. We suggest that interpreting models may become a routine part of any ML study for materials science alongside other important pillars such as openness and reproducibility.[7]

KEY CONCEPTS

Interpretability of machine learning models is at the forefront of research in computer science. As such, there is an abundance of technical jargon associated with the subject. We try to eliminate unnecessarily technical explanations in this Account, but in the interest of avoiding the quandary of being "divided by a common language",[8] we start by clarifying some of the terms we see as unavoidable for a proper exploration of the subject.

Classical/Deep Machine Learning

When we refer to classical methods, we mean any ML method that is not neural-network-based. When we refer to deep learning (DL) methods, we are referring to any method based on neural network architectures. This definition is important because there is generally a difference in how we interpret classical or deep learning methods. Classical methods are generally trained on structured data, with human-defined and (more-or-less) interpretable features. For example, in materials, a feature could correspond to the mean electronegativity of elements in a compound. Deep learning methods are often trained on less structured data, e.g., images. Deep learning methods learn reduced dimensionality representations of these inputs (known as representation learning) and then use these learned features to perform nonlinear regression or classification. Exploiting representation learning can be a route to building more interpretable and reliable ML models.[9] Because of this difference in how features are developed and used, interpreting classical and deep models often requires different approaches.We note that these differences do have exceptions and deep learning methods can often take structured inputs and classical ML can sometimes work on unstructured data.

Interpretability, Explainability, Completeness, Understandability

There are many definitions and debates about what constitutes a good explanation or an interpretable model. To try to avoid confusion, it is worth defining what we mean when we refer to explainable and interpretable models. Interpretability is often stated, but far less often defined, and as Lipton argues,[1] it is not a monolithic concept. We follow ref 5 that a model is interpretable if its operation can be understood by a human. An example of an interpretable model might be a simple decision-single-tree-based model, the operation of which can be understood by inspecting the structure of the tree, whereas a deep neural network is unlikely to be interpretable by inspection.

Explainable models are models that can have explanations generated, in the case of interpretable models by inspecting the model or by using some other posthoc methods (which we refer to in this work as extrinsic explanations).[4] It could be argued that any ML model can be fully explained by reproducing the algorithmic process that generates its predictions; however, following Kovalerchuk et al.,[6] we call such explanations *quasi-explanations*. Quasi-explanations rely on concepts that are beyond the domain of application of the model. For example, extracting explanations using the activations of hidden layers makes little sense in the context of estimating the formation energy of a given crystal structure. We call the degree to which an explanation conforms to concepts relevant to the domain of application the *understandability* of the explanation. Another important property of an explanation is its completeness, based on the work of ref 10 and borrowing from formal logic; *completeness* is the extent to which all of the underlying model is described by the explanation.

Model explanations often involve a trade-off between understandability and completeness. We represent this tension in the radar plot in Figure 1. The NN can be completely explained by a mathematical description of its operations (explanation II), but this is not understandable in the terms of the domain, a quasi-explanation. The quality of an explanation or interpretation can be judged in terms of the domain theory to which it is applied. For example, we recently showed how interpretations for ML models of dielectric constants of materials are consistent with the Penn model of dielectrics.[11] Explanation III does not completely describe the underlying NN but produces an explanation that is understandable in terms of the domain. Some of these explanations may also be actionable, whereas others may not. For example, an analysis that identifies the electronegativity of a compound as the important feature is harder to act upon without changing other properties than an explanation that finds that reaction temperature is important.

Correctness

We introduce an additional factor for consideration, particularly in scientific applications: correctness. According to the famous aphorism, "all models are wrong, but some are useful", correctness is concerned with the degree of scientific wrongness of a given model. In part, this concept is an extension of transferability discussion in explainable artificial intelligence (AI)[1] to scientific ML. In scientific models, there is often a tension between how faithfully a model reproduces measurements and its complexity, as depicted in the radar plot of Figure 1. Example I presents a linear model that has been interpreted with a high degree of completeness, but the model may be limited in terms of adequately capturing physical or chemical phenomena. Thus, explanations of complex neural network models may allow a higher degree of physical correctness, as shown in examples I and II. In all cases, the correct choice depends on the motivation for building the model and how important interpretability and completeness are compared to physical or chemical correctness.

Causation

An important consideration is that, although a "correct" ML model may approximate a physical phenomenon and also give a rough idea of cause and effect, this does not translate necessarily into the discovery of "causation", unless there is specific experimental setup or particular assumptions regarding the

https://doi.org/10.1021/accountsmr.1c00244
Acc. Mater. Res. 2022, 3, 597−607

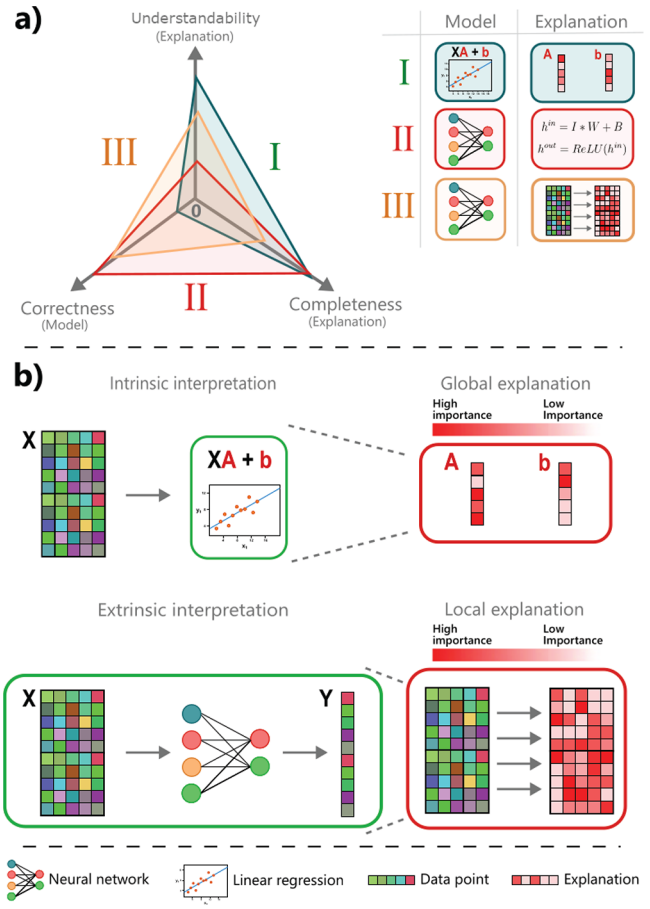


Figure 1. Key concepts. (a) Many explanations often involve a trade-off. In particular, any explanation should balance completeness, i.e., how well the given explanation approximates the operating mechanisms of the actual model and understandability, i.e., how well can a human subject understand the given explanation. In models that approximate real-world phenomena, we argue that a third dimension exists: correctness, i.e., how correct a given model explanation is from the physical or chemical points of view. In the text, we explain the various trade-offs in more detail in examples I, II, and III. (b) Illustration of local/global explanations and intrinsic/extrinsic interpretability. In (I), a linear model is an intrinsically interpretable model. By construction, the vector coefficients **A** and **b** can be interpreted directly, as represented by the color scale, although we add an important warning about the interpretability of linear models in the section on Intrinsically Interpretable Methods. In the same way, under certain assumptions, a linear model can be explained globally as both **A** and **b** are applied to all inputs X and have constant contributions for each input. (II) A neural network requires extrinsic explanations. Because of its nonlinear nature, interpretations of the model require observing the inputs along with the outputs. In the same way, a neural network model is better explained using local explanations: each input interacts in a different way with the model to generate specific outputs.

phenomenon. Confusing interpretability and causality is a common issue in scientific ML. Explaining a model of a physical phenomenon will give an idea of the predictive power of variables but is not the same as giving a causal explanation of the physical phenomenon. We argue that, to provide causal insights, ML explainability techniques have to be supplemented by adequate follow-up experimentation or explicit causal modeling. In this context, the application of scientific ML for understanding physical or chemical problems overlaps with machine learning concepts such an *analytical* and *inductive learning*,[12] i.e., respectively, the identification of relevant features by analyzing particular examples with adequate domain knowledge or by identifying shared commonalities between data points.

Local/Global Explanations

Local explanations tell us why a model reached a certain decision for a given case or data point, global explanations tell us why the model generally behaves as it does. In Figure 1, the coefficients of the linear regression provide a global explanation of model behavior, whereas the salience map provides a local explanation. From classical thermodynamics, the equation

$$\Delta G = \Delta H - T\Delta S$$

provides a global explanation of the relationship among free energy, enthalpy, temperature, and entropy, whereas an individual calorimetry experiment can give a local explanation on how the entropy of a particular system depends on temperature.

Intrinsic/Extrinsic Interpretations

Interpretability can come from examining the model itself or, alternatively, by examining how the model responds to stimuli; the former is an intrinsic interpretation, and the latter is an extrinsic interpretation. Intrinsic methods generally provide global explanations, because the interpretation depends on the construction of the model; intrinsic methods are specific to given types of ML algorithms. The canonical example of an intrinsically interpretable model is linear regression, as illustrated in Figure 1, although we caution that this is not always the case, at the end of the next section on Intrinsically Interpretable Models, we discuss potential issues with this. Extrinsic methods, however, are generally model-agnostic or exploit specific inductive biases in models and often provide local explanations, because they rely on perturbations of the input data and observation of how the model responds. Because deep learning methods rely on huge numbers of learnable parameters, it is unlikely that one could intrinsically examine the model as it is and understand how it works. Therefore, interpretability of deep learning methods comes from extrinsic methods. Because the descriptors of deep learning models are learned inside the model, rather than provided, special methods for uncovering these features are needed.[1,2]

■ INTRINSICALLY INTERPRETABLE MODELS

A range of atomistic ML models has been introduced in recent years. The focus has mainly been on the regression of atom-resolved properties, or global properties as dependent on individual atomic environments.[13] The construction of structural descriptors is often guided by physical ideas, encoding information about environments and symmetries, but this is not an indispensable practice, as complex neural networks have also been used to capture materials structures from raw data inputs. The former naturally lend themselves to interpretable models. The development of physically motivated interatomic potentials from machine learning has been comprehensively covered in other review articles.[14]

Sometimes interpretability can be improved by reducing the number of features, while minimally affecting performance. It is possible to use regularization techniques to limit the number of descriptors to those that are most important for capturing the relationship between the data and the property of interest. Regularization approaches such as SISSO (sure independence screening and sparsifying operator) and LASSO (least absolute shrinkage and selection operator), as well as approaches based

https://doi.org/10.1021/accountsmr.1c00244
Acc. Mater. Res. 2022, 3, 597−607

on perturbing features and retraining models, have all been used to to produce ostensibly more interpretable models. Regularization approaches have been used to reappraise structure prediction heuristics, for perovskites and zinc blende/wurtzite systems.[15] In an example of perturbation methods, backward feature elimination was applied to identify four descriptors most important for predicting superconductor critical temperatures.[16]

In LASSO-type methods, if a group of features are highly correlated, LASSO often arbitrarily chooses one feature at the expense of the others in the group. In perturbation elimination methods, high levels of correlation mean that if an important feature is dropped from the model, it may be compensated by a correlated feature, thus masking the importance. In general, it is good practice to use feature elimination approaches in conjunction with correlation metrics, for example, Pearson or Spearman correlation metrics, although one should also remain vigilant, as low correlation scores do not necessarily mean unrelated features.

We finish dealing with intrinsically interpretable models by noting that it is also important not to fetishize simpler models in the name of interpretability. Particularly important in this regard is the scenario of model mismatch, where the model form fails to capture the true form of a relationship; i.e., according to our previous definition, it provides low correctness. For example, if a linear model is used to capture a nonlinear relationship, the model will increasingly attribute importance to irrelevant features in an attempt to minimize the difference between the model predictions and the training data and will ultimately produce meaningless explanations. In machine learning literature, a common solution to preserve predictive power and allow high intrinsic interpretability is using generalized linear models with specific linkage functions or generalized additive models (GAMs). For example, GAMs have been used to model and interpret the driving factors of chemical adsorption of subsurface alloys,[17] modeling a nonlinear process with a high degree of interpretability. Linear models are not always as interpretable as they claim to be. For example, if features are heterogeneous and have very different ranges and values, the coefficients of a linear model will probably tell us more about the sizes of various parameters than they will tell us about some underlying physical explanation that is understandable to a domain expert.

■ MODEL EXPLANATION METHODS

Though some ML methods offer intrinsically interpretable results, many more complex models such as deep neural networks (DNNs) are not as easily understood. Even when models are inherently interpretable, extrinsic interpretation methods can provide additional insights impossible by examining the model alone.

What-If Explanations

There are a range of "what-if" analyses that work by examining how the value of the model output changes when one or more of the input values are modified. Partial dependence plots (PDPs) examine how changing a given feature affects the output, ignoring the effects of all other features. For example, we could look at the effect of the mean atomic mass of a material on the dielectric response, marginalizing all other factors using a model such as that presented in ref 19. One drawback is that confounding relationships are missed and can mask effects. For example, imagine increased mean atomic mass increased the dielectric response in a dense material but decreased it in a porous material. These factors would cancel in PDP. Individual conditional expectations (ICE) plots are closely related to PDPs but overcome this limitation and allow group factors to be uncovered.[20]

Feature importance can be calculated as the partial derivative of the output of the model with respect to a given input feature

$$FI_j = \frac{\partial f(X)}{\partial x_j} \tag{1}$$

This equation is solved by taking a Taylor series with the leading term based at a boundary or decision point of the function. By taking derivatives at this point, we see how much a particular feature affects the decision in that region. Typically, this type of feature importance is applied locally on a per sample basis.

In an alternative type of feature importance, the error (e) using the original features (X_{orig}) is calculated

$$e_{orig} = L(y, f(X_{orig})) \tag{2}$$

where the loss function (L) depends on the values of labels (y) and model predictions (X_{orig}). The values are randomly shuffled for each feature (j) and a permuted feature loss (e_{per}^{j}) is calculated. The feature importance is either the ratio or the difference of e_{orig} and e_{per}^{j}. This kind of approach applied globally across the whole data set (it should be the test set) has been been used, for example, to show which features affect the band gap of a material or the dielectric response.[21]

Though feature importance scores can offer insights, they can be misleading. The methods used within many widespread machine learning packages, such as SCIKIT-LEARN have some well-known pathologies. These kinds of feature importance metrics tend to favor continuous over categorical features, and care should be taken in particular when categorical features are used with high dimensionality or continuous features with wide ranges. In materials science, the features that we use are often of vastly different ranges and categorical dimensions, which means that the feature importance obtained by default from these models should be treated with great caution, particularly where counterintuitive results are obtained. In the same way, the actual importance of features and their trends can be masked by observed confounding by other features.

Shapley analysis based on the SHAP method (explained in Figure 2) is becoming very popular in many ML studies. Applying SHAP analysis on support vector regression (SVR) model, it was possible to understand how physical descriptors contribute to the model's ability to predict the dielectric constant of crystals revealing relationships similar to long established empirical models, but with greater predictive power.[11] SHAP values are also now being more commonly applied to give global as well as local explanations, for example, in models that predict atomic charges.[22]

There are limitations to SHAP analysis related to causation and correlation, which should be considered when it is applied. First, like all of the what-if analyses presented, SHAP values can sample unrealistic combinations where parameters are correlated. For example, in a material, an input combination where the HOMO energy is higher than the LUMO energy could be explored despite being physically unrealistic. Second, SHAP values are arrived at by including parameters in sequence, but there is no notion of how one feature may directly cause another, so having a low HOMO may be causally related to having a large band gap, but the SHAP value will be calculated as if neither of these features is causally related to the other. We consider some possible

https://doi.org/10.1021/accountsmr.1c00244
Acc. Mater. Res. 2022, 3, 597−607

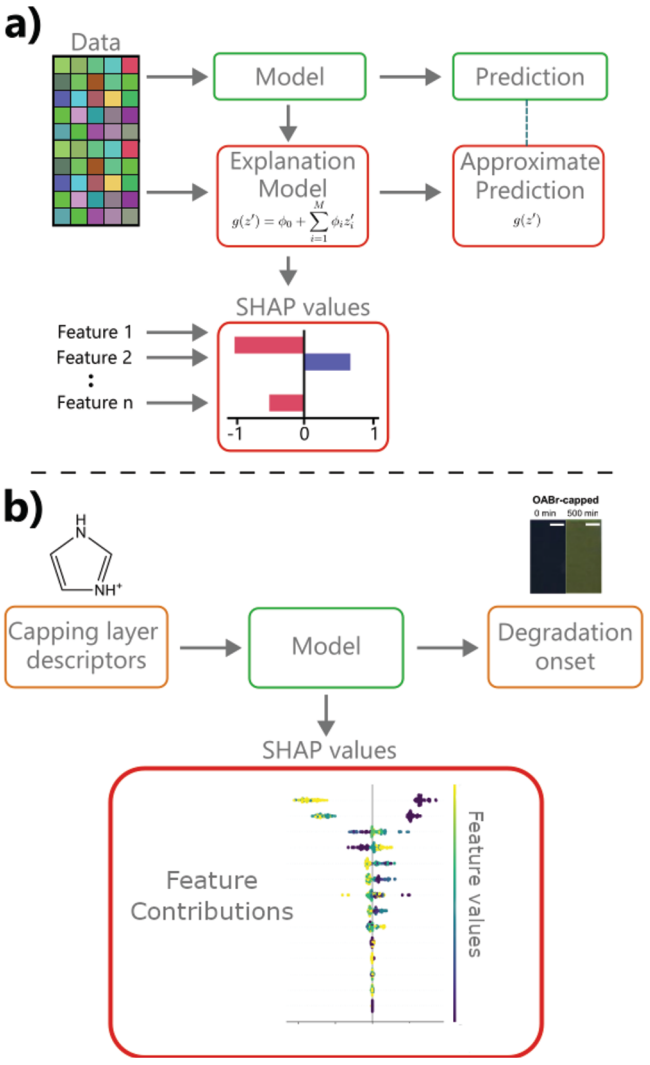


Figure 2. Interpretability with SHAP values. (a) SHAP values are a generalization of various black-box explainability methods. SHAP values work by approximating the output of a model with a local linear explanation model. The coefficients of this explanation model quantify the local effect of each feature on the output. The coefficients can be aggregated to get global feature contributions. (b) Case study of SHAP analysis in material science.[18] A model is built to relate the physiochemical descriptors of a capping layer of lead halide perovskite solar cells. Then, a machine learning model is trained to predict the onset time of degradation of the solar cells under ambient conditions. SHAP analysis allows us to identify the dominant descriptors in the model, shown in the figures as a distribution of local SHAP values. *Top polar surface area* and *H-bond donor* have the most significant impact the output's prediction and are demonstrated to have dominant importance by additional experimentation. The SHAP analysis is reproduced under the CC-BY 4.0 License from ref 18.

solutions to these limitations in the Physical Knowledge Beyond Model Explanation section.

Related to this kind of analysis are the class of counterfactual analysis methods, which examine how small changes in input features can result in different outcomes. There are not many examples of counterfactual approaches in materials science yet, although they have recently been applied in the field of molecular chemistry.[23] Counterfactual and contrastive explanations provide insights of model operation by determining examples or cases that explain the difference between a desired outcome and actual outcome.[24] As these explanations are often very actionable by humans and can lead to valid hypothesis or experiments, we believe they should be more frequently utilized in material science and chemistry.

Deep Explanations

As we described in the Key Concepts section, deep learning methods learn nonlinear representations rather than relying on handcrafted inputs. Because of this, there are a number of DL-specific methods for interpretability. Methods such as feature importance from gradients, discussed above, breakdown for DL models, because small changes in input can result in discontinuous changes, the so-called shattered gradients problem.[26,27] Interpretation methods for DL models typically take one of two approaches, *processing* methods or *representation* methods.[3] Processing methods examine how the model processes a given input in manner similar to a what-if analyses, whereas representation methods attempt to interpret learned representations or intrinsically learn representations that have some degree of interpretability.

The most popular approach to understanding how the DL model processes data are salience methods. Since the early efforts in this area, a range of methods have been developed that examine a balance between the areas of the network that respond most strongly to a given input (the activations) and the areas that are most sensitive, i.e., where changes in activations would change the output most (the gradients). An overview of various methods for salience mapping is available elsewhere.[28] The class activation map (CAM)/grad-CAM approach builds a map of the input regions that are responsible for a classification by calculating how the different convolutional filters contribute to that classification, the operation of CAM is presented schematically in Figure 3.

Recently, transformer architectures have demonstrated outstanding performance on a variety of vision and language tasks. Transformers utilize attention mechanisms, which learn a weighted masking of different sections of the input data during an encoding procedure (see Figure 4). The transformer begins by learning an "embedding" for each element of the input. A simple example could be a vector for an atom of the length of all other elements in the periodic table. The embedding then represents how a given atom is related to all other atoms. This embedding then passes through the attention layers, which learn how much attention elements of the vectors pay to each other. These representations, known as *attention masks*, can be interpreted in similar way to salience maps and determine sections of the input data that a model exploits for making predictions. The authors of a transformer model trained on chemical reaction data were able to perform atom-mapping and learn chemical grammars,[29] i.e., identify atoms during a chemical reaction, by interpreting its learned attention map.

With salience and attention-based approaches, there is a danger of overinterpretation, particularly in cases where physical explanations are searched for. There can be cases where salience maps produce the same explanation for *all* classes. This can happen, for example, if the network is particularly responsive to edges in the input, as opposed to more meaningful features. There is a tendency in the literature to produce salience maps only for the top-ranked class, but to ensure that the network really is picking a class because of certain region it is important to consider what parts of the image activate for other classes that are not the top class. Another challenge of deep interpretation techniques is that they may lead to unexplainable results. In

https://doi.org/10.1021/accountsmr.1c00244
Acc. Mater. Res. 2022, 3, 597−607

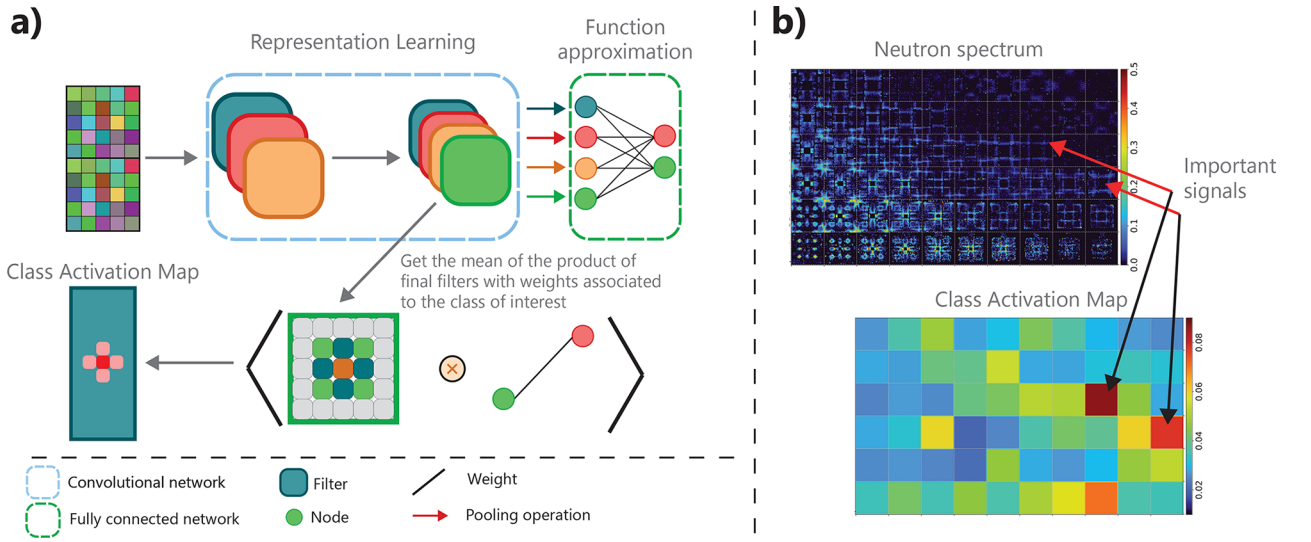


Figure 3. Interpreting the results from a deep convolutional neural network. (a) Schematic of class activation maps (CAM). The trained model is presented with a new instance that is passed through the network. The filters in the final convolutional layer are pooled to a single node, the weights connecting this node to a given class are multiplied by the filter, and the average of all resulting weighted filters is taken and projected back onto the original image, to show the important regions for the classification to that given class. (b) CAM in action. A CNN was trained to classify magnetic Hamiltonians based on inelastic neutron scattering spectra (upper) and highlight the regions of energy transfer in Q-space that are important for making distinctions using a CAM (lower). The regions identified by the CNN/CAM match with the regions that a trained physicist identifies, but in a fraction of the time. Spectra and maps reproduced under the CC-BY 4.0 License from ref 25.

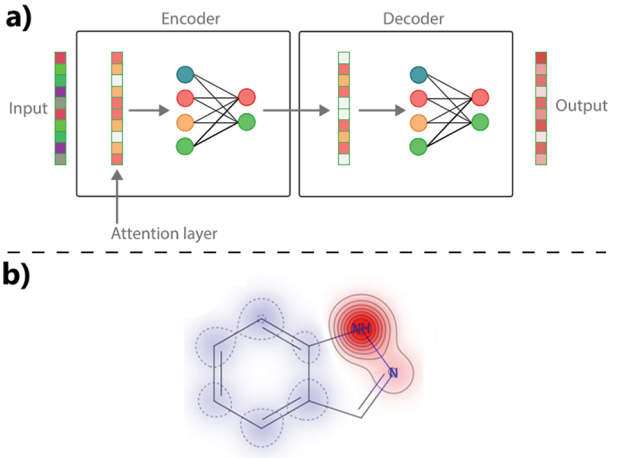


Figure 4. (a) Attention mechanism in a transformer neural network. The network learns to translate between sequences of arbitrary length, through encoder and decoder networks. The encoder network takes an embedding of vector of each element of the input and passes it through an attention layer, which learns how much to correlate different members of the input sequence. This then goes through a standard multilayer perceptron and can be repeated an arbitrary number of times. (b) Attention-derived map of the molecular fragments found to be important for predicting log *P* values. Part b of the figure is reproduced under theCC-BY 4.0 License from ref 30.

these cases, a physical explanation of feature attribution may be difficult to infer as the model may be exploiting correlations in the data distribution that do not have a physical explanation or cause, a problem commonly known as shortcut learning.[31]

EXPERIMENTAL PREDICTIONS AND EXPLANATIONS

Interpretability is a desirable characteristic of ML models in experimental contexts as it facilitates tasks such as characterization, optimization, sensitivity analysis, and hypothesis testing. In materials science, physical models are often developed to approximate input−output relations in experimental processes and interpret, identify, or optimize dominant physical parameters. Examples of such models include molecular dynamics simulations of carbon nanotube synthesis or drift-diffusion models of semiconductor devices. This category of models is derived from known mathematical relations with chosen inductive biases and is often designed with intrinsic experimental interpretability in mind. However, explanations of these models are subject to the tension among correctness, interpretability, and completeness, which we introduced in Figure 1.

We argue that a principled adaptation of ML models to the experimental context may preserve a useful degree of interpretability. For example, a combination of intrinsic physical parametrization and a surrogate ML model has been applied to the complex problem of mapping fabrication variables to final figures-of-merit in layered semiconductors[32] or first-principles calculations have been used to constrain surrogate-based compositional optimization.[33] In these hybrid models, ML enhances the model capacity of a physical parametrization to better approximate experimental data,[32] account for uncertainty, or deal with noise.[33] It also allows the smooth integration of first-principles calculations into experimentation.[33]

In scenarios where hypotheses are tested experimentally, predictive power may be second to induction or explainability. Traditionally, ML models with intrinsic interpretability have proven useful in this setting even if the absence of causal models, for example, in the case of inference and explanation of degradation processes.[34] In this case, the explanations produced by SHAP analysis provided meaningful actionable insights that allowed materials scientists to design new better materials for perovskite solar cells. ML models with a high degree of intrinsic interpretability have been used to identify dominant material descriptors in high dimensional material screening spaces[35] or to actively guide experimental interventions with physical or causal

https://doi.org/10.1021/accountsmr.1c00244
Acc. Mater. Res. 2022, 3, 597−607

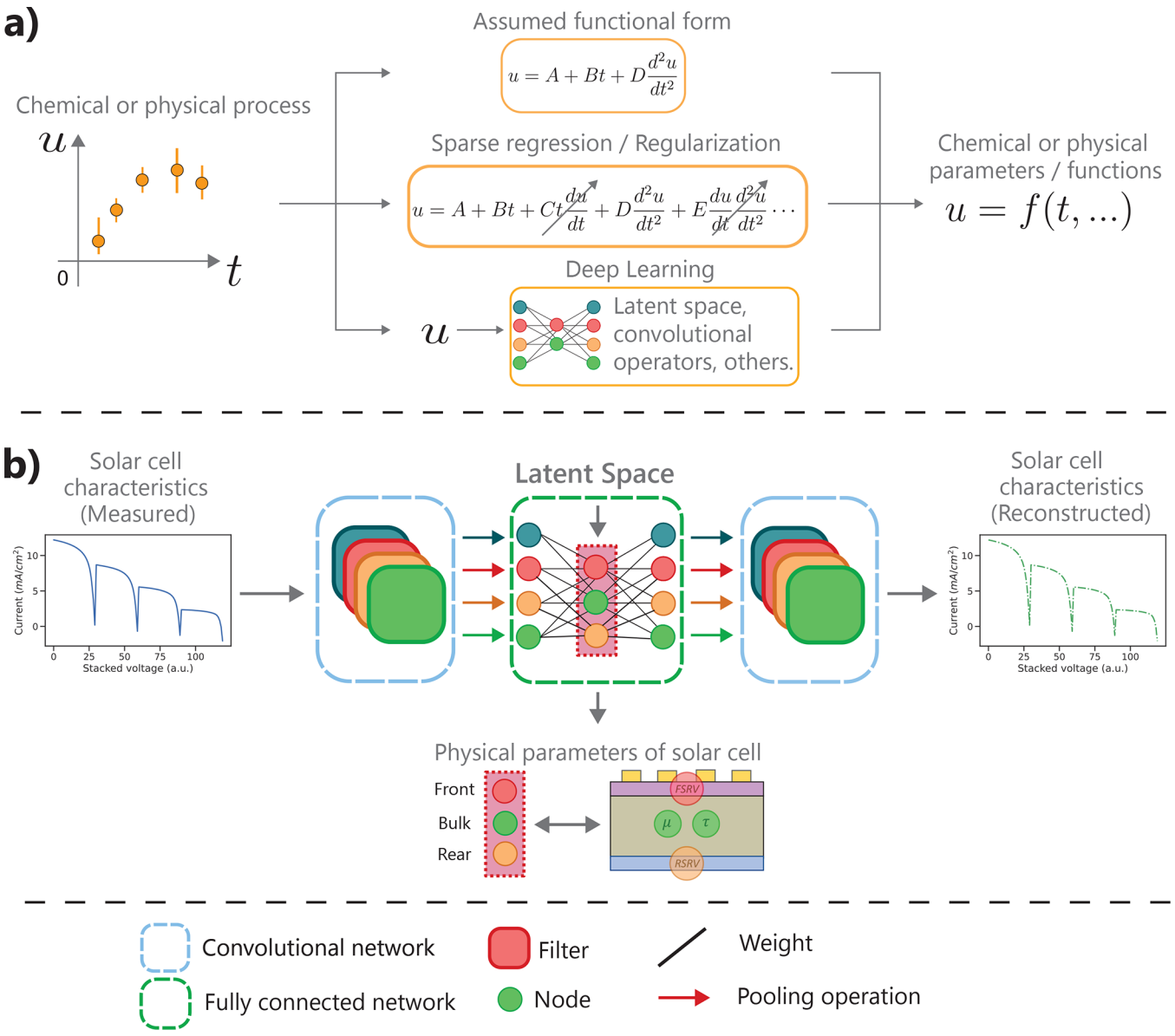


Figure 5. (a) General taxonomy for the methods of direct knowledge extraction from physical or chemical data. (b) Example of the application of deep autoencoders to learn useful and interpretable representations of solar cell data. Current voltage characteristics of solar cells can be encoded in physical parameters. By analyzing these physical parameters, we gain operational knowledge of solar cells and are able to infer paths for device optimization.[32]

constraints.[33] In the same way, the capacity of ML models to approximate complex conditional distributions may facilitate understanding of complicated physical and chemical systems for which high-performing physical simulations are limited, as demonstrated by recent advances in likelihood-free inference.

Extrinsic methods such as SHAP and salience are proving powerful in coupling ML models to experimental procedures. SHAP analysis has been used to understand ML models that predict the efficacy of organic capping layers for increasing stability of halide perovskites solar cells, highlighting the importance of low numbers of hydrogen bond donors and small topological polar surface ares.[18] Salience methods have been used to identify the regions in 3D neutron spectroscopy signals that are most important for deciding the magnetic structure in a double perovskite, these regions are found to match with the regions identified by a trained physicist, but are found in a fraction of the time.[25] Salience methods were also used to identify the regions responsible for misclassifications in an X-ray diffraction analysis deep neural networks, allowing human intervention where the model is likely to perform poorly.[36]

These examples demonstrate how building interpretability is the key to successful application of ML for enhanced experimentation. Interpretable models not only increase the level of trust in the ML approach but also help to strengthen the relationship between the algorithms and the humans in the experimental loop.

PHYSICAL KNOWLEDGE BEYOND MODEL EXPLANATIONS

Most applications of interpretability in the material science field have been driven by predictability goals, having interpretability as a second-order goal. We believe that some machine learning problems might be better framed with the explicit goal of extracting actual knowledge or causal interpretations.[34,37] In

https://doi.org/10.1021/accountsmr.1c00244
Acc. Mater. Res. 2022, 3, 597−607

exploratory scientific research, this often constitutes a better trade-off of the dimensions we explore in Figure 1.

One approach in the broader physics community, summarized in Figure 5a, consists in designing or learning models that directly extract knowledge from noisy experimental data. An initial approximation to the problem is to assume an functional form and fit various coefficients to experimental data. A more robust approach consists in using sparse regression[34] or genetic algorithms[38] to explore many potential functional forms and find those that better explain the data. An evolution of these techniques is based on deep learning methods, commonly on deep autoencoders (AEs), which, for example, have been shown successful in extracting order parameters for phase transitions[39] or disentangling physical phenomena in microscopy data.[40] AEs work by learning to reconstruct an input, while passing through a reduced dimensional space, termed the latent space, thus learning compressed representations of the data (see Figure 5). This latent space can be constrained by various methods, for example, by adding penalty losses to the latent space that penalize for physical variables,[32] explicitly defining hierarchical (or even causal) graphical structures in the latent space,[41] etc. Simple operations in this interpretable latent space, such as clustering or regression, make it possible to find physical insights or perform physically relevant predictions. Figure 5b presents an example of using AEs to extract physically relevant knowledge from noisy experimental data and a physical model and use these learning to design an optimal solar cell fabrication process.[32]

As another example, the so-called β-VAE,[42] introduces additional constraints to enforce orthogonality and sparsity on the latent space, so that the dimensions are uncorrelated and the VAE will only use the minimum number of dimensions required for reconstruction of the data. This kind of β-VAE type approach was recently shown to extract parameters that are interpretable as the driving parameters of ordinary differential equations from data of dynamic processes.[43]

Another challenge in the field is related to the lack of confidence intervals or error distributions for model explanations. In contrast to classical statistical approaches on linear models, interpretations of modern machine models do not produce any notion of uncertainty. This fact greatly limits the confidence of any insights extracted from the model, as there is no inherent notion of uncertainty to them. Various works have explored uncertainty and bias in model explanations and have proposed ways to account for them.[44] We expect that future interpretability approaches in material science will integrate this inherent notions of uncertainty into the insights extracted from ML models.

Finally, a continuing and fundamental challenge in ML interpretability is that explanations do not have strong causal guarantees or resilience against cofounding effects. Thus, the real-world insights gained from interpretability tend to be limited by the judgment of the scientist or secondary confirmation by experiments or simulation. The field of causal inference is witnessing a renaissance in fields of AI where explainable chains of action are legally necessary, such as autonomous vehicles. Though most ML methods work on identifying correlations in data, they say nothing about cause and effect. The leading proponent of causal inference Judea Pearl has pronounced ML methods to be "profoundly dumb" for this reason.[45] This constitutes a very active area of research in mainstream machine learning, and we are optimiztic about future progress in the field. By combining ML with the tools of causal inference, it may be possible to learn new cause and effect relationships from materials data. A recent pioneering example suggesting the potential for this kind of approach was reported on electron microscopy data by Ziatdinov and colleagues,[46] who report combining ML with causal inference to uncover mechanisms driving ferroelectric distortions, based on experimental micrographs. ML was used to obtain descriptors of the data, combined with other physical descriptors, such as composition, and information-geometric causal inference was used to infer causal relationships between the different descriptors. We believe that one great advantage of causality approaches in hard experimental science is that there is significant control of confounding factors by means of traditional experimental design.

These concepts of causal inference are also being explored in the context of generating extrinsic explanations of ML models. We have described how SHAP analysis provides a powerful, principled approach to asking what-if questions of models and producing insightful interpretations (see section What-If Interpretations). However, these values can be susceptible to sampling unrealistic parameter combinationsl. For example, the SHAP method may try to calculate the dielectric constant for a metal with a band gap of 2 eV, if metal/nonmetal and band gap were separate input parameters. VAEs have been explored as a method for ensuring that sampled scenarios for arriving at final SHAP values fall within reasonable distributions.[47] SHAP values also have no concept of causality, so the density of a material may just as well result in composition as vice versa. By relaxing the symmetry constrains for SHAP values, it becomes possible to develop interpretations that respect known causal chains.[48] These developments have thus far only been applied in computer science; however, their applicability of developing more robust and meaningful explanations for materials science ML is clear.

Other fields such as econometrics have a long tradition of developing statistical models to get insights about certain phenomena.[49] We imagine these approaches extracting meaningful, actionable information from complex materials data, such as processing conditions, complex compositional landscapes and large scale simulations. We also urge critical appraisal of the application of statistical methods in the light of known physical constraints, as has been highlighted in econometrics.[50]

CONCLUSIONS

The technological advances of machine learning have been felt in all areas of science and technology in the past decade. As the initial excitement at these disruptive successes starts to fade, the long process of realizing the true potential of ML for understanding the world begins. As with any largely empirically developed technological advance, the process of understanding just why it works so well will only increase the breadth and depth of the application of ML. In this paper, we have outlined some of our experience and observations of the nascent field of interpretable ML, in the context of materials science. The methods that we have outlined here cover interpretability for the range of ML methods that are becoming increasingly popular in materials science. Many of the methods we have presented are as easily implemented as the ML models they interpret. As such, we hope that in future every ML paper in materials science will include some efforts to understand the derived models and to extract more knowledge from the information. We have also tried to provide a balanced critique of the potential shortcomings of these methods, interpretable ML is not a silver bullet for model understanding and constitutes an initial approx-

https://doi.org/10.1021/accountsmr.1c00244
Acc. Mater. Res. 2022, 3, 597−607

imation to causal hypothesis generation. We are convinced that for ML to achieve full impact in materials science, understanding the results of our models will be imperative. In the final analysis, to paraphrase David E. Womble (who may have been paraphrasing Max Planck), interpretable ML will not replace human experts, but human experts who embrace interpretable ML will replace those who do not.

AUTHOR INFORMATION

Corresponding Authors

Felipe Oviedo − *Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; Microsoft AI for Good Research Lab, Redmond, Washington 98052, United States*; Email: foviedo@alum.mit.edu

Keith T. Butler − *SciML, Scientific Computing Department, Rutherford Appleton Laboratory, Didcot OX110D, U.K.; Department of Chemistry, University of Reading, Reading RG6 6AD, U.K.*; orcid.org/0000-0001-5432-5597; Email: keith.butler@stfc.ac.uk

Authors

Juan Lavista Ferres − *Microsoft AI for Good Research Lab, Redmond, Washington 98052, United States*

Tonio Buonassisi − *Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0001-8345-4937

Complete contact information is available at: https://pubs.acs.org/10.1021/accountsmr.1c00244

Author Contributions

#F.O. and K.T.B. have equal contributions.

Author Contributions

K.T.B. and F.O. conceived the work and wrote the manuscript with key intellectuals contributions from T.B. and J.L.F.

Notes

The authors declare no competing financial interest.

Biographies

Felipe Oviedo is an applied researcher at Microsoft AI for Good, focusing on scientific machine learning for sustainability and healthcare applications. Prior to joining Microsoft, Felipe completed a PhD at the intersection of material science and computer science at MIT under the guidance of Prof. Tonio Buonassisi (MIT Mechanical Engineering) and Dr. John Fisher (MIT CSAIL). His dissertation was focused on accelerated development of photovoltaics by physics-informed machine learning. Felipe developed and deployed machine learning algorithms to accelerate the experimental screening and optimization of renewable energy materials and technologies. Before MIT, Felipe briefly worked in the energy industry and CERN.

Juan Lavista Ferres is the Chief Scientist and Lab Director of the Microsoft AI for Good Research Lab, where he works in AI, Machine Learning, and statistical modeling, working across Microsoft AI for Good efforts. Before joining Microsoft, Juan was the CTO and cofounder of alerts.com. He spent 6 years in Washington working at the InterAmerican Development Bank applying data science to understand the impact of programs for reducing poverty and inequality in Latin-America and the Caribbean. Juan has two computer science degrees from the Catholic University in Uruguay and a graduate degree in Data Mining and Machine Learning from Johns Hopkins University.

Tonio Buonassisi is a Professor of Mechanical Engineering at the Massachusetts Institute of Technology (MIT). He is pioneering the application of artificial intelligence to develop new materials for societally beneficial applications. His research in solar photovoltaics and technoeconomic analysis assisted technology developments in dozens of companies, earning him a US Presidential Early Career Award for Scientists and Engineers (PECASE), a National Science Foundation CAREER Award, and a Google Faculty Award. Tonio received his PhD from Berkeley before working as a researcher in the Fraunhofer Institute in Freiburg and a crystal grower for Evergreen Solar.

Keith T. Butler is a senior scientist at the Rutherford Appleton Laboratory, in the Scientific Machine Learning (SciML) team, where he leads projects that apply machine learning to the discovery and characterization of materials. Keith obtained his bachelor's from Trinity College Dublin and his PhD from University College London. Before joining RAL, Keith spent time as a postdoctoral researcher in the groups of Aron Walsh (University of Bath/Imperial College London) and John Harding (University of Sheffield) and was a visiting researcher in The University of Toronto and Tokyo Institute of Technology. Keith's research focuses on using machine learning to accelerate the characterization of materials and also to predict new, previously undiscovered materials for renewable energy applications, such as photovoltaics and photocatalysts. Keith is an active developer of several open source materials design packages (SMACT, SuperResTomo, MacroDenisty) and a strong advocate of open science.

ACKNOWLEDGMENTS

We thank Professor Volker Deringer and Dr. Noor Titan Putri Hartono for useful discussion. We thank Pedro Costa for his contributions to figure design. This worked was supported by the National Research Foundation (NRF), the Singapore Massachusetts Institute of Technology (MIT) Alliance for Research, and Technology's Low Energy Electronic Systems research program, Microsoft AI for Good.

REFERENCES

(1) Lipton, Z. C. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* **2018**, *16*, 31−57.

(2) Molnar, C. *Interpretable Machine Learning*; 2019; https://christophm.github.io/interpretable-ml-book/.

(3) Gilpin, L. H.; Bau, D.; Yuan, B. Z.; Bajwa, A.; Specter, M.; Kagal, L. Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*; IEEE, 2018; pp 80−89.

(4) Biran, O.; Cotton, C. Explanation and justification in machine learning. A survey; Reference: IJCAI-17. *Workshop on Explainable AI (XAI)* **2017**, *8*, 8−13.

(5) Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* **2019**, *267*, 1−38.

(6) Kovalerchuk, B.; Ahmad, M. A.; Teredesai, A. Survey of explainable machine learning with visual and granular methods beyond quasi-explanations. *Interpretable artificial intelligence: A perspective of granular computing*; Springer, 2021; pp 217−267.

(7) Artrith, N.; Butler, K. T.; Coudert, F.-X.; Han, S.; Isayev, O.; Jain, A.; Walsh, A. Best practices in machine learning for chemistry. *Nat. Chem.* **2021**, *13*, 505−508.

(8) The original quote about England and Americaand America being "divided by a common language" is variously attributed to Oscar Wilde, George Bernard Shaw, and Winston Churchill.

(9) Allotey, J.; Butler, K. T.; Thiyagalingam, J. Entropy-based active learning of graph neural network surrogate models for materials properties. *J. Chem. Phys.* **2021**, *155*, 174116.

(10) Kulesza, T.; Stumpf, S.; Burnett, M.; Yang, S.; Kwan, I.; Wong, W.-K. Too much, too little, or just right? Ways explanations impact end

users' mental models. *2013 IEEE Symposium on visual languages and human centric computing*; IEEE, 2013; pp 3−10.
(11) Morita, K.; Davies, D. W.; Butler, K. T.; Walsh, A. Modeling the dielectric constants of crystals using machine learning. *J. Chem. Phys.* **2020**, *153*, 024503.
(12) Mitchell, T. M. *Machine Learning*; McGraw-Hill, 1997.
(13) Antunes, L. M.; Grau-Crespo, R.; Butler, K. T. Distributed representations of atoms and materials for machine learning. *npj Computational Materials* **2022**, *8*, 44.
(14) Deringer, V. L.; Caro, M. A.; Csányi, G. Machine learning interatomic potentials as emerging tools for materials science. *Adv. Mater.* **2019**, *31*, 1902765.
(15) Ouyang, R.; Curtarolo, S.; Ahmetcik, E.; Scheffler, M.; Ghiringhelli, L. M. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Phys. Rev. Mater.* **2018**, *2*, 083802.
(16) Stanev, V.; Oses, C.; Kusne, A. G.; Rodriguez, E.; Paglione, J.; Curtarolo, S.; Takeuchi, I. Machine learning modeling of superconducting critical temperature. *npj Computational Materials* **2018**, *4*, 29.
(17) Esterhuizen, J. A.; Goldsmith, B. R.; Linic, S. Theory-Guided Machine Learning Finds Geometric Structure-Property Relationships for Chemisorption on Subsurface Alloys. *Chem.* **2020**, *6*, 3100−3117.
(18) Hartono, N. T. P.; Thapa, J.; Tiihonen, A.; Oviedo, F.; Batali, C.; Yoo, J. J.; Liu, Z.; Li, R.; Marrón, D. F.; Bawendi, M. G.; et al. How machine learning can help select capping layers to suppress perovskite degradation. *Nat. Commun.* **2020**, *11*, 4172.
(19) Takahashi, A.; Kumagai, Y.; Miyamoto, J.; Mochizuki, Y.; Oba, F. Machine learning models for predicting the dielectric constants of oxides based on high-throughput first-principles calculations. *Phys. Rev. Mater.* **2020**, *4*, 103801.
(20) Goldstein, A.; Kapelner, A.; Bleich, J.; Pitkin, E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics* **2015**, *24*, 44−65.
(21) Davies, D. W.; Butler, K. T.; Walsh, A. Data-driven discovery of photoactive quaternary oxides using first-principles machine learning. *Chem. Mater.* **2019**, *31*, 7221−7230.
(22) Korolev, V. V.; Mitrofanov, A.; Marchenko, E. I.; Eremin, N. N.; Tkachenko, V.; Kalmykov, S. N. Transferable and Extensible Machine Learning-Derived Atomic Charges for Modeling Hybrid Nanoporous Materials. *Chem. Mater.* **2020**, *32*, 7822−7831.
(23) Wellawatte, G. P.; Seshadri, A.; White, A. D. Model agnostic generation of counterfactual explanations for molecules. *Chemrxiv* **2021**, DOI: 10.26434/chemrxiv-2021-4qkg8-v2.
(24) Verma, S.; Dickerson, J.; Hines, K. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596* **2020**, DOI: 10.48550/arXiv.2010.10596.
(25) Butler, K. T.; Le, M. D.; Thiyagalingam, J.; Perring, T. G. Interpretable, calibrated neural networks for analysis and understanding of inelastic neutron scattering data. *J. Phys.: Condens. Matter* **2021**, *33*, 194006.
(26) Balduzzi, D.; Frean, M.; Leary, L.; Lewis, J.; Ma, K. W.-D.; McWilliams, B.The shattered gradients problem: If resnets are the answer, then what is the question? *International Conference on Machine Learning*; MLR Press, 2017; pp 342−350.
(27) White, A. D. *Deep Learning for Molecules and Materials*; thewhitelab.org, 2021; https://dmol.pub/intro.html.
(28) Ancona, M.; Ceolini, E.; Öztireli, C.; Gross, M. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. *arXiv.1711.06104* **2018**, DOI: 10.48550/arXiv.1711.06104.
(29) Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances* **2021**, *7*, No. eabe4166.
(30) Tang, B.; Kramer, S. T.; Fang, M.; Qiu, Y.; Wu, Z.; Xu, D. A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility. *Journal of cheminformatics* **2020**, *12*, 15.
(31) Robinson, C.; Trivedi, A.; Blazes, M.; Ortiz, A.; Desbiens, J.; Gupta, S.; Dodhia, R.; Bhatraju, P. K.; Liles, W. C.; Lee, A. Deep learning models for COVID-19 chest x-ray classification: Preventing shortcut learning using feature disentanglement. *medRxiv* **2021**, DOI: 10.1101/2021.02.11.20196766.
(32) Ren, Z.; Oviedo, F.; Thway, M.; Tian, S. I.; Wang, Y.; Xue, H.; Perea, J. D.; Layurova, M.; Heumueller, T.; Birgersson, E.; et al. Embedding physics domain knowledge into a Bayesian network enables layer-by-layer process innovation for photovoltaics. *npj Computational Materials* **2020**, *6*, 9.
(33) Sun, S.; Tiihonen, A.; Oviedo, F.; Liu, Z.; Thapa, J.; Zhao, Y.; Hartono, N. T. P.; Goyal, A.; Heumueller, T.; Batali, C.; et al. A data fusion approach to optimize compositional stability of halide perovskites. *Matter* **2021**, *4*, 1305−1322.
(34) Naik, R. R.; Tiihonen, A.; Thapa, J.; Batali, C.; Liu, Z.; Sun, S.; Buonassisi, T. Discovering Equations that Govern Experimental Materials Stability under Environmental Stress using Scientific Machine Learning. *arXiv preprint arXiv:2106.10951* **2021**, DOI: 10.48550/arXiv.2106.10951.
(35) Vasudevan, R. K.; Choudhary, K.; Mehta, A.; Smith, R.; Kusne, G.; Tavazza, F.; Vlcek, L.; Ziatdinov, M.; Kalinin, S. V.; Hattrick-Simpers, J. Materials science in the artificial intelligence age: high-throughput library generation, machine learning, and a pathway from correlations to the underpinning physics. *MRS Commun.* **2019**, *9*, 821−838.
(36) Oviedo, F.; Ren, Z.; Sun, S.; Settens, C.; Liu, Z.; Hartono, N. T. P.; Ramasamy, S.; DeCost, B. L.; Tian, S. I.; Romano, G.; et al. Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. *npj Computational Materials* **2019**, *5*, 60.
(37) Kalinin, S. V.; Ghosh, A.; Vasudevan, R.; Ziatdinov, M. Describing condensed matter from atomically resolved imaging data: from structure to generative and causal models. *arXiv preprint arXiv:2109.07350* **2021**, DOI: 10.48550/arXiv.2109.07350.
(38) Atkinson, S.; Subber, W.; Wang, L.; Khan, G.; Hawi, P.; Ghanem, R. Data-driven discovery of free-form governing differential equations. *arXiv preprint arXiv:1910.05117* **2019**, DOI: 10.48550/arXiv.1910.05117.
(39) Wetzel, S. J. Unsupervised learning of phase transitions: From principal component analysis to variational autoencoders. *Phys. Rev. E* **2017**, *96*, 022140.
(40) Kalinin, S. V.; Steffes, J.; Liu, Y.; Huey, B.; Ziatdinov, M. Disentangling ferroelectric domain wall geometries and pathways in dynamic piezoresponse force microscopy via unsupervised machine learning. *Nanotechnology* **2021**, *33*, 055707.
(41) Hsu, W.-N.; Zhang, Y.; Glass, J. Unsupervised learning of disentangled and interpretable representations from sequential data. *Advances in Neural Information Processing Systems*, **2017**, *30*.
(42) Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. *Proceedings of ICLR*; OpenReview.net, 2016.
(43) Lu, P. Y.; Kim, S.; Soljačić, M. Extracting interpretable physical parameters from spatiotemporal systems using unsupervised learning. *Physical Review X* **2020**, *10*, 031056.
(44) Schwab, P.; Karlen, W. Cxplain: Causal explanations for model interpretation under uncertainty. *arXiv preprint arXiv:1910.12336* **2019**, DOI: 10.48550/arXiv.1910.12336.
(45) Pearl, J. Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016* **2018**, DOI: 10.48550/arXiv.1801.04016.
(46) Ziatdinov, M.; Nelson, C.; Zhang, X.; Vasudevan, R.; Eliseev, E.; Morozovska, A. N.; Takeuchi, I.; Kalinin, S. V. Causal analysis of competing atomistic mechanisms in ferroelectric materials from high-resolution Scanning Transmission Electron Microscopy data. *npj Comput. Mater.* **2020**, *6*, 127.
(47) Frye, C.; de Mijolla, D.; Cowton, L.; Stanley, M.; Feige, I. Shapley-based explainability on the data manifold. *arXiv preprint arXiv:2006.01272* **2020**, DOI: 10.48550/arXiv.2006.01272.

(48) Frye, C.; Feige, I.; Rowat, C. Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. *arXiv preprint arXiv:1910.06358* **2019**, DOI: 10.48550/arXiv.1910.06358.
(49) Athey, S. Machine learning and causal inference for policy evaluation. *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*; ACM, 2015; pp 5−6.
(50) Leamer, E. E. Let's take the con out of econometrics. *American Economic Review* **1983**, *73* (1), 31−43.

https://doi.org/10.1021/accountsmr.1c00244
Acc. Mater. Res. 2022, 3, 597−607