Cell Reports
Physical Science

CellPress
OPEN ACCESS

Perspective

The emergent role of explainable artificial intelligence in the materials sciences

Tommy Liu[1,*] and Amanda S. Barnard[1]

SUMMARY

The combination of rational machine learning with creative materials science makes materials informatics a powerful way of discovering, designing, and screening new materials. However, moving from a promising prediction to a practical strategy often requires more than just an instructive structure-property relationship; understanding how a machine learning method uses the structural feature to predict the target properties becomes critical. Explainable artificial intelligence (XAI) is an emerging field in computer science based in statistics that can augment materials informatics workflows. XAI can be used as a forensic analysis to understand the consequences of data, model, and application decisions or as a model refinement method capable of distinguishing important features from nuisance variables. Here, we outline the state of the art in XAI and highlight methods most useful to the physical sciences. This practical guide focuses on characteristics of XAI methods that are relevant to materials informatics and will become increasingly important as more researchers move toward using deeper neural networks and large language models.

INTRODUCTION

At a high-level, machine learning (ML) is a computational tool that seeks to improve performance for a task using data. The many varieties of ML techniques and algorithms are becoming widespread, impacting multiple social, environmental, economic, and scientific domains. As a result of this increased adoption, we are now seeing significant advances in the sciences aided by ML techniques.[1–3]

An example of a typical ML workflow in the physical sciences is shown in Figure 1. Here, we can see the first steps involving the data generation process, which may involve gathering the data experimentally or via physics-based simulations with some process to generate the machine-readable outputs that can then be fed into an analysis model. The input to the ML model (features) can take many forms and can be grouped into descriptors drawn from scientific instrumentation or from computational and statistical analysis. Training ML models then involves splitting the data into training, testing, and validation sets to mitigate the bias and variance of such models.[4] The output is a model capable of assigning a target property (label) with predictable accuracy and performance. Before the advent of ML, scientists needed to know the structure of the model (the mathematical expression) in advance to be able to solve it and make a prediction, or they turned to statistical models and processes, which involve assumptions and knowledge about the underlying process itself.[5] By using ML, scientists are now able to develop the model and solve complex

[1]School of Computing, Australian National University, Canberra, ACT, Australia

*Correspondence: tommy.liu@anu.edu.au

https://doi.org/10.1016/j.xcrp.2023.101630

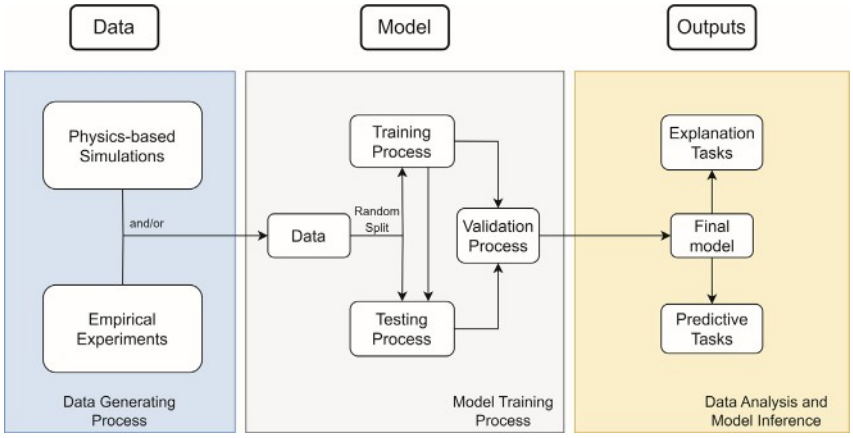

Figure 1. Example scientific workflow combining conventional simulation and machine learning
The input data pipeline could equally be replaced with the synthesis and characterization of experimental samples, followed by appropriate processing to generate observable data. Figure adapted incorporating elements from Brehmer et al.[9] and Huang et al.[10]

challenges at the same time, which represents an acceleration of the scientific method. This offers significant advantages but comes with a new set of challenges that depend on the algorithm. These methods are becoming more common in the materials sciences and have given rise to the sub-field of materials informatics (MI), which seeks to make use of advances in computation power and algorithmic design to drive scientific discovery.[6–8]

A significant hurdle to the further adoption of ML in MI is the core question of how algorithms derive their results and how we can trust that the results are correct. Many advanced algorithms, such as neural networks, are "black boxes" that provide no insights into how the input features relate to the target label. This is because many contemporary ML models are so complex that is becomes impossible to track the relationship between the inputs and outputs. To tackle these challenges, the field of explainable artificial intelligence (XAI) has emerged to provide robust and reproducible techniques to understand how a model operates and insights into how the results can be translated into practice.[11]

In this perspective, we provide an overview of XAI and how ML and XAI can be successfully combined in the context of the materials sciences. Some works in the literature use the term interpretable ML (IML) interchangeably with XAI, which we will also do in this perspective. This work focuses on the cases for "classical" ML techniques used to discover, design, and develop new materials,[1,12,13] as a comprehensive review of deep neural networks (DNNs) in materials science can be found in a review paper by X. Zhong et al.[1]

Background

The overlap between statistics and ML is significant. Statistics begins with data and is concerned with the analysis and insights that can be generated from that data.[5] This is also one of the aims of many studies using ML in modern research, and it is difficult to determine where statistics stops and ML begins. Traditionally the main difference is the division between inference and prediction tasks.[4,5] Inference is the process of creating rigorous mathematical models to more deeply understand or criticize some phenomena, whereas prediction primarily seeks to identify the best course of action to take (or predict) without requiring a significantly deeper understanding of

underlying mechanisms.[14] The techniques from within one field can also be significantly informed by the other. There are many overlapping techniques within these two fields, such as linear regression, which can be used for both inference and prediction,[4,15,16] and the difference manifests in how they are evaluated and analyzed. However, the challenges involved in the sciences are becoming increasingly complex, so many studies are naturally becoming more oriented toward ML. Interpretability lies within the domain of statistics, but using insights from statistics and novel new research, XAI methods can aid in bridging the gap between these domains and further drive innovation in the sciences.[2]

Although a clear consensus regarding definitions has not been reached, we begin with the two core concepts of "interpretability" and "explainability."[17–19] These concepts refer to different aspects of understanding a workflow, although some literature uses these terms interchangeably,[11] and others advocate for focusing on the properties of the techniques rather than attempting to distinguish them.[1] However, there is some value to clear definitions in fields such as MI, where this information is used to make practical decisions in the lab. Interpretability can be defined as "the ability to explain or provide meaning in understandable terms to a human" and is related to the historical differences between statistical and ML approaches. Statistical models are inherently more interpretable since they are both simpler and underpinned by strong assumptions.[20] Explainability more broadly seeks to "provide an interface between humans and the decision maker that is both accurate and comprehensible to humans."[21,22] In practice, regardless of the approach, XAI provides insights into the decisions made by a model, which implies interpretability has a slightly broader scope and includes factors such as model complexity and the transformations or the form that the data themselves take (i.e., data-process interpretability).[23,24] Some models are regarded as highly interpretable, such as decision trees/rules and linear models,[20,21,25] but the main difference between these two definitions depends on the stage of a workflow. Interpretable approaches consider the intrinsic aspects of a learning model, while explainable approaches provide insights into the model outputs and how they were derived (Table 1).

XAI is desirable to ensure that several human-centered considerations surrounding AI are preserved, which is relevant to MI. Several of the target areas are as follows.[49]

(1) Trustworthiness: how can researchers trust that the model produces correct outcomes? To use a model, we need to trust many aspects (including the following points) held in many situations. However, the task of trust is difficult to quantify[32,49] and may need to be defined on an individual researcher level such as advocated by Dazeley et al.[17]
(2) Causality: how can we understand or be sure that the input features truly affect the target outputs? This is a common goal within MI where we seek to understand the causes behind particular phenomena. ML can play a role in providing evidence for causal links but is not typically sufficient for determining causation.[50–52]
(3) Transferability: can the knowledge produced by the model be applied to different situations, and if so, is there evidence that the underlying mechanism is being correctly captured?[53]
(4) Confidence: how can we be sure that the model will perform as expected in different situations or even in the same situation? The concepts of robustness and stability of a model[49] provide errors regarding uncertainty and statistical methods (confidence intervals) or confidence-aware learning are also relevant.[54]

Table 1. Selection of popular XAI methods for tabular data, relevant to materials science applications

Method	Instances or features	Global or local	Intrinsic or post hoc	Description	Example works
Permutation importance	features	global	post hoc	calculates the relative increase in model performance attributed to a particular feature	Breiman, Fisher, Xu et al.[26–28]
Feature importance	features	global	intrinsic	calculates the importance of each feature from internal parameters of a model	Breiman, Groemping et al.[5,29]
Global surrogates	features	global	post hoc	fits a simpler, interpretable model as an approximation to the trained model	Ramprasad, Gorissen, Teichert et al.[7,30,31]
Local surrogates	features	local	post hoc	fits a simpler, interpretable model around a particular region (datums) of interest	Ribeiro, Lorenzi et al.[32,33]
Counterfactual explanations	features, instances	local	post hoc	finds the smallest change to a sample (features) that changes the result a desired amount	Oviedo, Karimi, Wachter, Ribeiro, Wellawatte et al.[12,34–37]
Shapley values	features	local, global	post hoc	finds a set of Shapley values that represent the individual feature contribution to the output of a model according to the game theory concept of how much each player (feature) affects the final output of the model)	Rodríguez-Pérez, Lundberg, Zhang, Huang et al.[38,39,40,41]
Influence statistics	instances	local	intrinsic	finds the "influence" or how much impact that data instances have upon the model parameters or outputs	Cook, Chatterjee, Azari et al.[42–44]
(Data) Shapley values	instances	local	post hoc	finds the Shapley value of the instances instead of the features, determining the most influential instances	Ghorbani, Jia, Barnard, Liu et al.[45–48]

Many of these issues have been previously addressed in the social sciences,[18] including several important aspects such as ethics, bias, and trust.[49,55–60] From the social sciences, we learn that a single explanation method is typically insufficient since different researchers understand concepts differently. This means that the same computational explanation that makes sense to a materials chemist (for example) may not make sense to a materials physicist (and vice versa).

This brings us to the key issue of what constitutes an "explanation." For example, is it sufficient to explain exactly how the model uses data to derive an output, or is some other motivation description required? An ML model is itself a mathematical description of how the outputs are derived, so any simpler explanation is, at best, an approximation of how the model operates and may be an oversimplification of the scientific problem. An infinite number of possible explanations that are more or less correct exist. As a result, the properties of explanations and how they can be interpreted become important. Some works refer to the difference between the true model and the explanation as "completeness."[12,61] Furthermore, there remains a gap between mathematical descriptions or explanations and the human task of understanding; some works advocate for a "conversation" between explanations and the human end user where different viewpoints are presented until the human becomes convinced.[17,18] As ML expands into more areas of materials science, the need for XAI to overcome these knowledge gaps will also increase. The European Union GDPR regulations already state that "a data subject" has the right to "an explanation of the decision reached after [algorithmic] assessment,"[62] and we can foresee a need for the industry to be able to justify investment decisions based on the predictions of ML models of material structures and processes.

Current state of the field of XAI

In computer science, XAI techniques can be broadly divided into intrinsic and post hoc methods.[63] Intrinsic methods make use of simpler models or those underpinned by strong assumptions, or internal architecture, which reduces the number of questions about how outputs were produced.[64–66]

Intrinsic interpretability is concerned with the structure of the model itself. If we consider a linear regression model of the form Ax + y, it is immediately clear how each individual prediction is made; that is, the output consists of each input feature based on the weights present in weight vector A. Another example is that of a decision tree, where each of the individual decisions used to arrive at the final output are determined only by a given threshold on each feature. In general, as models get more complex, the less intrinsically interpretable they are, but this is not always the case since we can impose interpretability constraints on structures of models themselves if necessary.[63] One example is that of localGLMnet,[67] where a particular structure was incorporated into a neural network in order to provide explanations similar to that of the linear regression model.

While there has been a significant focus on DNNs and their XAI interpretations,[1] there is a strong case to be made against universally using DNNs for tabular data. Tree-based models (XGBoost, random forests) have been empirically found to outperform DNNs in many applications,[25] and interpretability concerns regarding complex DNN model architectures have been raised.[65,68] As a result, there are many domains and datasets where state-of-the-art performances, computational limitations, or explainability concerns mean that DNNs may not be the best choice of technique[25,65,68,69]

Post hoc methods seek to probe a given model and ascertain some desired form of information about those outputs. Post hoc explanations take a model as input and generate a useful approximation of how the outputs were produced.[70] These methods are model agnostic and provide an interface between complex un-interpretable black boxes and an understandable explanation to the researcher.[63]

The field of post hoc explanations has seen the greatest research focus in recent years due to their convenient incorporation into any ML workflow and the lack of presumptions about the underlying model. Table 1 lists a selection of popular XAI techniques along with descriptions and some relevant examples. The most straightforward approach is that of model feature importances (how impactful the effect of removing or adding this feature); by directly considering the weights of, say, a linear regression model, it can be determined what are the most important features globally (assuming the data are normalized/standardized). At the same time, very simple models can be considered on a local level as well, particularly in the linear regression case. Permutation importance extends the global notion of importance by considering the removal or permutation of features to determine their overall effect upon the model.[71] Global surrogates seek to approximate the model by fitting a simpler model such as linear regression to a more complex model, and the previous techniques can be applied to interpret this simpler model.[63]

Another much less discussed aspect of XAI is that of the data-process interpretability which use interpretable techniques to transform and pre-process the data, including interpretable dimension reduction or data imputation.[24,72,73] Data-process methods, while not as popular as the two other areas in the current field of XAI, are equally important.[74] Imputation deals with missing or incorrect values in the

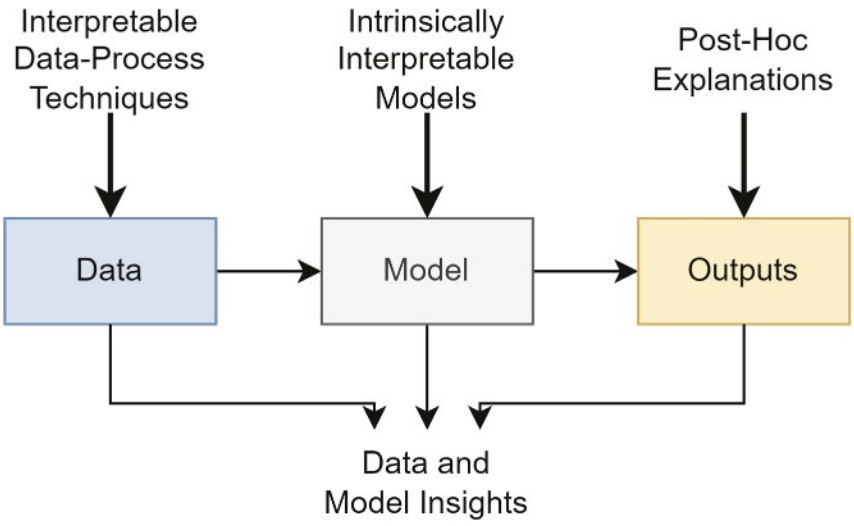

Figure 2. Illustration of data analysis workflow and where the stages of XAI fit
The three groups of XAI methods and which stage of an MI pipeline they may fit into.

data, while dimension reduction reduces the number of features or descriptors in order to better fit models or remove sources of multicollinearity, where features are related to each other.[24] Attempts have been made to interpret commonly used data transformation methods such as imputation[75] and dimensionality reduction.[73,76] This is particularly relevant to materials scientists since many of the data formats commonly used include tabular structured data.

The connection between these three groups of XAI methods and which stage of an MI pipeline they may fit into is illustrated in Figure 2. These XAI techniques can be further sub-divided into two groups, including global techniques, which seek to find patterns across the whole dataset, and local techniques, which seek to explain the process of what happens to an individual sample.[12,63]

Counterfactual explanations are concerned with the question of "what is the minimum change required to change the predicted outcome?"[35] An illustrative question may be "how much of this element I can remove from this reaction before there is no reaction." Counterfactual explanations are "human friendly" in the sense that they are typically the way that humans look at the world or at phenomena. The issue arises in generating counterfactuals from mathematical models, which produce mathematical representations of what needs to be changed. Based on the approach chosen, very different (sets of) counterfactuals may be generated, leaving the end user to search through them until a suitable one is found. Counterfactual explanations can aid in determining what are minimal sets of changes to the data features that will affect the target outputs of a model. This can identify which structural characteristics are not related to properties and can safely be omitted from design strategies or tuned to accommodate other needs such as material manufacturability.

Perhaps the most different approach, yet the one most familiar to materials scientists, is that of influence statistics.[43] Influence statistics is primary concerned with quantities derived from (linear) regression models such as Cooks' distance, leverage,[42] and difference in fits (DFFITS).[77] These quantities are a measure of how much the model changes in the presence of data instances and are derived from closed-form statistical expression. A significant drawback is that the closed

form solution of these quantities must be derived for each new model class and generally rely on the data projection (hat) matrix and do not generalize to other model classes such as random forests, which are the preferred method for regression modeling in contemporary ML tasks. The exception to this is that of influence functions, which approximate the change in the model with respect to a given instance by means of the Hessian matrix and allows this approach to be applied to all models with twice differentiable loss functions such as DNNs; however, random forests remain out of reach. The XAI counterpart to instance importance for tabular data would be that of data Shapley values, which simply apply Shapley values over the instances to decompose the instance contribution to the loss of the model. The significant drawback of data Shapley values, however, lies in the computational cost, as it requires retraining the model many times.[45]

Evaluation

There have been attempts to unify descriptions of various XAI tasks to have consistent evaluation methodologies, as is common in many areas in the physical sciences. Additional requirements that accompany any explanation method may include the functional details (scope, methodology, usage). Operational and usability (how techniques can be used and their properties) criteria should be provided to guide researchers' understanding.[78] Security, privacy, and validation details should be provided to understand the risks and benefits of given methods, particularly given the transnational aspect of a lot of materials science research. Security and privacy in ensuring that data are not leaked or biases enforced are becoming increasingly relevant with ethical concerns around AI and ML usage. Validation needs to be carried out to ensure that results are correct and consistent with the scientific nature of materials sciences. Sokol and Flach[78] argue for a "fact sheet" spanning these five dimensions so that XAI methods may be compared due to a lack of common consensus in the field, which will be a familiar concept to materials scientists accustomed to standards and the use of materials safety data sheets (SDSs). This is relevant because there is little use comparing the actual outputs of individual XAI methods given that they produce explanations with different degrees of correctness.[1]

Shapley value analysis

Shapley values are a concept from co-operative game theory and seek a solution to the problem of giving attribution to a set of actors who produce some final output.[79] In the context of ML, the actors may be the feature values themselves, and the output would be the final model predictions. Shapley values are particularly useful and are increasingly used in many scientific disciplines.[10,80] Shapley values are highly accessible, most notably through the Shapley additive explanations (SHAP) framework,[39] which provides fast and accurate approximations to the true Shapley value. They also satisfy a set of properties that makes interpreting them much simpler compared to LIME,[32] which is a popular local surrogate method[81] that can similarly provide insights into how features are combined. LIME is concerned with fitting an interpretable model such as a linear regression to the local area around a point of interest. It does so by perturbing the point and treating these perturbations as new data. This provides a very similar interpretation compared to the Shapley approach; however, it is only a local approximation and cannot be extrapolated across the whole dataset.

The SHAP framework computes how much of each feature contributes to the final prediction of the model for the local effects of each instance and can be aggregated to describe the global effects of features. An example of such a local data visualization strategy using Shapley values, called the "Force-Plot,"[82] can be seen in Figure 3.

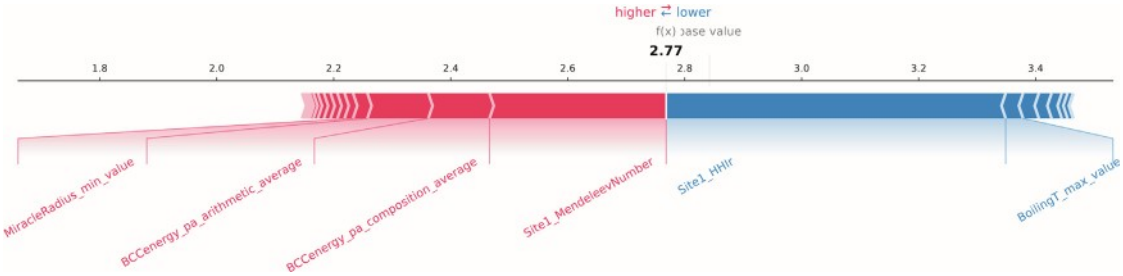

Figure 3. Force plot showing the model predicted value of raw energy (eV) for a single instance within the dilute solution diffusion dataset
Model trained using a random forest[83] and evaluated using the TreeExplainer using SHAP.[39] Reproduced with permission under the MIT license. The meaning of the features (i.e., Site1_mendeleevNumber) is the Mendeleev number of the element present at the first site of the material) are not so relevant but illustrate that each feature contributes a certain amount to the final model output of 2.77 eV.

Furthermore, the types of data are not limited to simple features, and these techniques have found use in visualizing the most important areas of images or most important sets of molecular bonds in compound classification (Figure 4).

One useful aspect of Shapley values is that they are capable of both local and global feature analyses. The Shapley values for each individual prediction can be aggregated to form the global feature importance of a particular feature. These local and global explanations can be presented for post hoc analysis to verify the results of the models developed in a clear and understandable way for almost any workflow. Indeed, we are seeing many such works that incorporate Shapley values as a final step after model development.[40,41,47,84] This contrasts with other feature-based explanations that involve adding additional steps into the ML workflow (such as training a local or global surrogate), which makes interpretations slightly more complex.

Application of XAI in materials science

The number of papers in materials sciences making use of ML has increased exponentially over the past few years,[20] including applications in materials discovery, property predictions, process optimization, and many more.[85] A significant part of the XAI literature is focused on the feature space, determining how much each feature contributed to the model or final prediction results,[71] which is consistent with the aims in materials science and is the basis of structure-property relationships. This suggests an ideal combination of methods and applications, and we have already seen feature selection techniques applied to improve predictive models in materials science[28,86–88] with varying choices of global feature importance used. These global approaches are useful for generally analyzing trends across the entire dataset to identify physicochemical characteristics that are consistently correlated to functional properties, regardless of individual structural nuances. Restricting ML models to these universally important features can reduce model complexity and improve computational efficiency, without compromising accuracy. Global ranking of structural features can be easily achieved using Shapley values by aggregating across instances, or global rankings of individual materials can be obtained by aggregating across features.[47] In both cases, this can be applied to any dataset or model, allowing for simple cross-model comparison, even in the absence of intrinsic interpretability.

Local feature-based approaches can also aid in determining the importance of features for individual data instances that are important, interesting, or different. They can also facilitate more detailed understanding of results, such as which physicochemical features of a material make it particularly important to the model

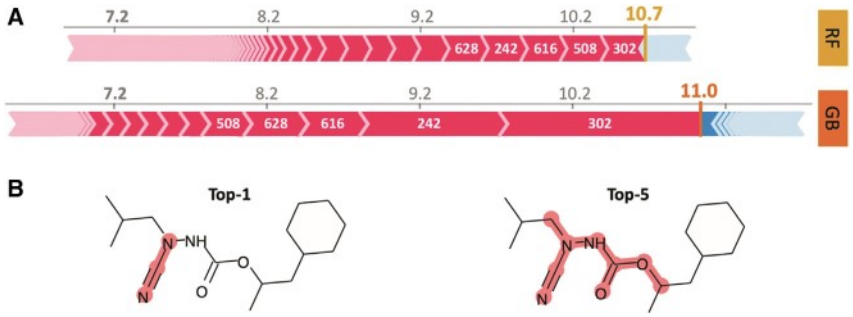

Figure 4. Force plots for the random forest and gradient-boosted machine algorithms to predict the potency of a particular compound

(A) The contribution of the most important tabular features.

(B) The contribution of the top-1 and top-5 most important graphical features.

prediction or which particular materials cause certain features to be so critical. If those materials or features were omitted from the dataset, the prediction would likely change, and so local-feature-based approaches become invaluable in quantifying the impact of these decisions.

Combining these benefits, SHAP is gaining popularity in MI across a variety of subdomains and applications. In pharmaceutical research, Rodríguez-Pérez and Bajorath[89] studied 10 bioactivity classes and structure-activity relationships (SARs) of a range of compounds using random forests (RFs), support vector machines (SVMs), and DNNs. SHAP was used to explain activity predictions and identify features that increase or reduce the probability of predicted activity for mapping onto molecular graphs. Other applications in biomaterials have also benefited from Shapley analysis, including the optimization of small molecules,[90] the binding of macromolecules,[91] and the stability of compounds.[92] SHAP has also been incorporated into pipelines for screening.[10,93]

On a larger length scale, Shapley analysis has been applied to porous materials[94–96] to determine the properties that impact the optimal loading of fuels[97] and catalysts.[98] At the device level, Wang et al.[99] used RFs and a genetic algorithm to study thermally evaporated perovskite solar cells based on both material characteristics and fabrication parameters to improve power conversion efficiency (PCE). SHAP analysis was used to show that the ratio of cations to anions in the perovskite layer and the annealing temperature contribute the most to PCE, leading to the optimum device architecture and fabrication conditions that could exceed state-of-the-art efficiency records. In both cases, the prediction explanation was cited as particularly important to translating the findings into practice.

Future opportunities for XMI

Compared with materials science, or even other areas of ML, the field of XAI is still in its infancy. XAI has still yet to be incorporated into all facets of ML itself since the topic of understanding is still a secondary concern in many predictive tasks. As the challenges in materials science change and become more complex,[1] we can expect to see wider acceptance and adoption of these methods.

A major challenge in many areas of materials sciences is the prohibitive costs of obtaining data (via experiments or electronic structure simulations[100]). Some raw materials are expensive, toxic, flammable, radioactive, or difficult and dangerous to

work with, meaning that experiments are difficult to carry out or verify. This is often compensated by extensive characterization and extraction of a significant number of features since the cost of obtaining additional features is much lower than that of additional material instances. This results in very high dimensionality that can be alleviated using XAI, particularly interpretable data-process techniques such as interpretable dimensionality approaches. The automatic discovery and elimination of uninformative data are particularly desirable, especially when they cannot be related to domain knowledge, making XAI a powerful tool for model refinement. Alternatively, XAI can also drive experimental design to find increasingly relevant features,[101] i.e., by means of the local and global feature contribution scores.

Techniques to reduce the number of instances required for an analysis or to determine what sort of future data to gather are also desirable (as opposed to reducing the number of features). The study of the influence of a data instance has traditionally been seen in statistics.[42,43] Today, it is seeing increasing research focus in XAI domains.[45,46] Many existing works are limited to data valuation tasks,[45,102] which can be useful for improving the quality of the datasets available. However, these techniques provide little insight into why particular data may be important. New XAI techniques must be developed to address the sorts of challenges presented here, as determining the most useful types of data and informing the experimental design process can save time, money, and effort when developing materials datasets.

As MI continues to develop, the advances in XAI techniques will certainly become increasingly critical. We speculate that such a sub-branch incorporating these elements may be called XMI and will further accelerate scientific developments in the same way that ML did for materials science. An immediate possibility is the inclusion of post hoc analysis techniques, which can be incorporated into almost any workflow and can provide further guarantees, particularly in the presence of domain experts that models are performing correctly. By including these approaches, we are already seeing cases where materials science discoveries are aided by XAI.[28,86–88,97] Furthermore, there remains a gap between scientific understanding and numerical explanations from current XAI techniques. Current XAI techniques provide an explanation of the model outputs, and it is up to the users to translate this into meaningful science. With the advent of large language models (LLMs), this understanding may be incorporated into XAI workflows to critique the models until suitable explanations may be found.

While existing XAI techniques may not be sufficient to achieve all the materials science goals of today, they may necessarily provide additional evidence to verify the results of ML methodologies. Despite the pushback from some domains such as in health informatics,[103,104] the benefits that XAI techniques can provide to the materials sciences are numerous. Approaches such as Shapley values are particularly relevant and can be used as additional evidence of the underlying phenomena,[10,105,106] including highly prized structure-property relationships. The adoption of ML has met some resistance from such domains due to scientific, regulatory, or social concerns[103,107] but has significant potential in MI to translate predictions into meaningful insights that can inform investment decisions and future research directions. We anticipate that the demand for XAI in MI will increase in future years as more advanced ML methods are applied to this domain, such as large pre-trained transformer models.[108]

ACKNOWLEDGMENTS

This research was supported by an Australian Government Research Training Program (RTP) Scholarship.

AUTHOR CONTRIBUTIONS

Conceptualization, methodology, and writing - original draft, T.L.; writing - review & editing, supervision, and project administration, A.S.L.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Zhong, X., Zhao, X., Zhang, L., Liu, N., Shi, S., and Wang, Y. (2022). Explainable machine learning in materials science. Biochem. Biophys. Res. Commun. *606*, 1–9. https://doi.org/10.1038/s41524-022-00884-7.
2. Pilania, G. (2021). Machine learning in materials science: From explainable predictions to autonomous design. Comput. Mater. Sci. *193*, 110360. https://doi.org/10.1016/j.commatsci.2021.110360.
3. Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., Vogt-Maranto, L., and Zdeborová, L. (2019). Machine learning and the physical sciences. Rev. Mod. Phys. *91*, 045002.
4. Bishop, C.M. (2006). Pattern Recognition and Machine Learning (Information Science and Statistics) (Springer-Verlag).
5. Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Stat. Sci. *16*, 199–231. https://doi.org/10.1214/ss/1009213726.
6. Barnard, A.S., Motevalli, B., Parker, A.J., Fischer, J.M., Feigl, C.A., and Opletal, G. (2019). Nanoinformatics, and the big challenges for the science of small things. Nanoscale *11*, 19190–19201. https://doi.org/10.1039/c9nr05912a.
7. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A., and Kim, C. (2017). Machine learning in materials informatics: recent applications and prospects. npj Comput. Mater. *3*, 54. https://doi.org/10.1038/s41524-017-0056-5.
8. Agrawal, A., and Choudhary, A. (2016). Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. Apl. Mater. *4*, 053208. https://doi.org/10.1063/1.4946894.
9. Brehmer, J., Cranmer, K., Louppe, G., and Pavez, J. (2018). Constraining effective field theories with machine learning. Physical review letters *121*, 111801. https://doi.org/10.1103/PhysRevLett.121.111801.
10. Huang, W., Suominen, H., Liu, T., Rice, G., Salomon, C., and Barnard, A.S. (2023). Explainable discovery of disease biomarkers: The case of ovarian cancer to illustrate the best practice in machine learning and Shapley analysis. J. Biomed. Inf. *141*, 104365. https://doi.org/10.1016/j.jbi.2023.104365.
11. Molnar, C., Casalicchio, G., and Bischl, B. (2020). Interpretable Machine Learning - A Brief History, State-of-the-Art and Challenges. inproceedings, I. Koprinska, M. Kamp, A. Appice, C. Loglisci, L. Antonie, A. Zimmermann, R. Guidotti, z. zgbek, R.P. Ribeiro, and R. Gavald, et al., eds. (Springer), pp. 417–431. https://doi.org/10.1007/978-3-030-65965-3_28.
12. Oviedo, F., Ferres, J.L., Buonassisi, T., and Butler, K.T. (2022). Interpretable and Explainable Machine Learning for Materials Science and Chemistry. Acc. Mater. Res. *3*, 597–607. https://doi.org/10.1021/accountsmr.1c00244.
13. Dybowski, R. (2020). Interpretable machine learning as a tool for scientific discovery in chemistry. New J. Chem. *44*, 20914–20920. https://doi.org/10.1039/D0NJ02592E.
14. Bzdok, D., Altman, N., and Krzywinski, M. (2018). Statistics Versus Machine Learning. Nat. Methods *15*, 233–234. https://doi.org/10.1038/nmeth.4642.
15. Gregori, D., Petrinco, M., Bo, S., Desideri, A., Merletti, F., and Pagano, E. (2011). Regression models for analyzing costs and their determinants in health care: an introductory review. Int. J. Qual. Health Care *23*, 331–341. https://doi.org/10.1093/intqhc/mzr010.
16. Maulud, D., and Abdulazeez, A.M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. Journal of Applied Science and Technology Trends *1*, 140–147. https://doi.org/10.38094/jastt1457.
17. Dazeley, R., Vamplew, P., Foale, C., Young, C., Aryal, S., and Cruz, F. (2021). Levels of explainable artificial intelligence for human-aligned conversational explanations. Artif. Intell. *299*, 103525. https://doi.org/10.1016/j.artint.2021.103525.
18. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007.
19. Roscher, R., Bohn, B., Duarte, M.F., and Garcke, J. (2020). Explainable Machine Learning for Scientific Insights and Discoveries. IEEE Access *8*, 42200–42216. https://doi.org/10.1109/ACCESS.2020.2976199.
20. Schleder, G.R., Padilha, A.C.M., Reily Rocha, A., Dalpian, G.M., and Fazzio, A. (2020). Ab Initio Simulations and Materials Chemistry in the Age of Big Data. J. Chem. Inf. Model. *60*, 452–459. https://doi.org/10.1021/acs.jcim.9b00781.
21. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. ACM Comput. Surv. *51*, 1–42. https://doi.org/10.1145/3236009.
22. Doshi-Velez, F., and Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning. Preprint at ArXiv. https://doi.org/10.48550/arXiv.1702.08608.
23. Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., and Baesens, B. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. Decis. Support Syst. *51*, 141–154. https://doi.org/10.1016/j.dss.2010.12.003.
24. Liu, T. (2021). Dimension Reduction and Data Augmentation Methods for the Physical Sciences. https://doi.org/10.25911/VBTY-Z808.
25. Shwartz-Ziv, R., and Armon, A. (2022). Tabular data: Deep learning is not all you need. Inf. Fusion *81*, 84–90. https://doi.org/10.1016/j.inffus.2021.11.011.
26. Breiman, L. (2001). Random Forests. Mach. Learn. *45*, 5–32. https://doi.org/10.1023/A:1010933404324.
27. Fisher, A., Rudin, C., and Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. J. Mach. Learn. Res. *20*, 177–177:181. https://doi.org/10.48550/arXiv.1801.01489.
28. Xu, Y., Jiang, L., and Qi, X. (2021). Machine learning in thermoelectric materials identification: Feature selection and analysis. Comput. Mater. Sci. *197*, 110625. https://doi.org/10.1016/j.commatsci.2021.110625.
29. Groemping, U. (2006). Relative Importance for Linear Regression in R: The Package relaimpo. J. Stat. Software *17*, 1–27. https://doi.org/10.18637/jss.v017.i01.
30. Gorissen, D., Couckuyt, I., Demeester, P., Dhaene, T., and Crombecq, K. (2010). A Surrogate Modeling and Adaptive Sampling Toolbox for Computer Based Design. J. Mach. Learn. Res. *11*, 2051–2055. https://doi.org/10.5555/1756006.1859919.
31. Teichert, G.H., and Garikipati, K. (2019). Machine learning materials physics: Surrogate optimization and multi-fidelity algorithms predict precipitate morphology in an alternative to phase field dynamics. Comput. Methods Appl. Mech. Eng. *344*, 666–693. https://doi.org/10.1016/j.cma.2018.10.025.
32. Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). Why Should I Trust You? In Explaining the Predictions of Any Classifier. inproceedings, B. Krishnapuram, M. Shah, A.J. Smola, C.C. Aggarwal, D. Shen, and R.

Rastogi, eds. {ACM}), pp. 1135–1144. https://doi.org/10.1145/2939672.2939778.

33. Lorenzi, J.M., Stecher, T., Reuter, K., and Matera, S. (2017). Local-metrics error-based Shepard interpolation as surrogate for highly non-linear material models in high dimensions. J. Chem. Phys. *147*, 164106. https://doi.org/10.1063/1.4997286.

34. Karimi, A., Barthe, G., Balle, B., and Valera, I. (2020). In Model-Agnostic Counterfactual Explanations for Consequential Decisions. inproceedings, S. Chiappa and R. Calandra, eds. {PMLR}), pp. 895–905. https://doi.org/10.48550/arXiv.1905.11190.

35. Wachter, S., Mittelstadt, B.D., and Russell, C. (2017). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. CoRR *Abs/1711*, p. 00399. https://doi.org/10.48550/arXiv.1711.00399.

36. Ribeiro, M.T., Singh, S., and Guestrin, C. (2018). In Anchors: High-Precision Model-Agnostic Explanations. inproceedings, S.A. McIlraith and K.Q. Weinberger, eds. ({AAAI} Press), pp. 1527–1535. https://doi.org/10.1609/aaai.v32i1.11491.

37. Wellawatte, G.P., Seshadri, A., and White, A.D. (2022). Model agnostic generation of counterfactual explanations for molecules. Chem. Sci. *13*, 3697–3705. https://doi.org/10.1039/d1sc05259d.

38. Rodríguez-Pérez, R., and Bajorath, J. (2020). Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. J. Comput. Aided Mol. Des. *34*, 1013–1026.

39. Lundberg, S.M., and Lee, S. (2017). In A Unified Approach to Interpreting Model Predictions. inproceedings, I. Guyon, U. von Luxburg, S. Bengio, H.M. Wallach, R. Fergus, S.V.N. Vishwanathan, and R. Garnett, eds., pp. 4765–4774. https://doi.org/10.1048550/arXiv.1705.07874.

40. Zhang, M., Li, J., Kang, L., Zhang, N., Huang, C., He, Y., Hu, M., Zhou, X., and Zhang, J. (2020). Machine learning-guided design and development of multifunctional flexible Ag/poly (amic acid) composites using the differential evolution algorithm. Nanoscale *12*, 3988–3996. https://doi.org/10.1039/c9nr09146g.

41. Huang, H., Wang, X., Shi, J., Huang, H., Zhao, Y., Xu, H., Zhang, P., Long, Z., Bai, B., Fa, T., et al. (2021). Material informatics for uranium-bearing equiatomic disordered solid solution alloys. Mater. Today Commun. *29*, 102960. https://doi.org/10.1016/j.mtcomm.2021.102960.

42. Cook, R.D. (1979). Influential Observations in Linear Regression. J. Am. Stat. Assoc. *74*, 169–174. https://doi.org/10.2307/2286747.

43. Chatterjee, S., and Hadi, A.S. (1986). Influential Observations, High Leverage Points, and Outliers in Linear Regression. Stat. Sci. *1*, 379–393. https://doi.org/10.1214/ss/1177013622.

44. Azari, A., Nabizadeh, R., Nasseri, S., Mahvi, A.H., and Mesdaghinia, A.R. (2020). Comprehensive systematic review and meta-analysis of dyes adsorption by carbon-based adsorbent materials: Classification and analysis of last decade studies. Chemosphere *250*, 126238. https://doi.org/10.1016/j.chemosphere.2020.126238.

45. Ghorbani, A., and Zou, J.Y. (2019). In Data Shapley: Equitable Valuation of Data for Machine Learning. inproceedings, K. Chaudhuri and R. Salakhutdinov, eds. {PMLR}), pp. 2242–2251. https://doi.org/10.48550/arXiv.1904.02868.

46. Jia, R., Dao, D., Wang, B., Hubis, F.A., Hynes, N., Grel, N.M., Li, B., Zhang, C., Song, D., and Spanos, C.J. (2019). In Towards Efficient Data Valuation Based on the Shapley Value. inproceedings, K. Chaudhuri and M. Sugiyama, eds. {PMLR}), pp. 1167–1176. https://doi.org/10.48550/arXiv.1902.10275.

47. Barnard, A.S. (2022). Explainable prediction of N-V-related defects in nanodiamond using neural networks and Shapley values. Cell Reports Physical Science *3*, 100696. https://doi.org/10.1016/j.xcrp.2021.100696.

48. Liu, T., and Barnard, A. (2023). Shapley Based Residual Decomposition for Instance Analysis. https://doi.org/10.48550/arXiv.2305.18818.

49. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

50. Pearl, J. (2009). Causality (Cambridge university press).

51. Ting, J.Y.C., Li, S., and Barnard, A.S. (2022). Causal Paths Allowing Simultaneous Control of Multiple Nanoparticle Properties Using Multi-Target Bayesian Inference. Adv. Theory Simul. *5*, 2200330. https://doi.org/10.1002/adts.202200330.

52. Ting, J.Y.C., Parker, A.J., and Barnard, A.S. (2023). Data-Driven Design of Classes of Ruthenium Nanoparticles Using Multitarget Bayesian Inference. Chem. Mater. *35*, 728–738. https://doi.org/10.1021/acs.chemmater.2c03435.

53. Yamada, H., Liu, C., Wu, S., Koyama, Y., Ju, S., Shiomi, J., Morikawa, J., and Yoshida, R. (2019). Predicting Materials Properties with Little Data Using Shotgun Transfer Learning. ACS Cent. Sci. *5*, 1717–1730. https://doi.org/10.1021/acscentsci.9b00804.

54. Moon, J., Kim, J., Shin, Y., and Hwang, S. (2020). Confidence-Aware Learning for Deep Neural Networks (inproceedings), pp. 7034–7044. ({PMLR}). https://doi.org/10.48550/arXiv.2007.01458.

55. Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., and Floridi, L. (2018). Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. Sci. Eng. Ethics *24*, 505–528. https://doi.org/10.1007/s11948-017-9901-7.

56. Keskinbora, K.H. (2019). Medical ethics considerations on artificial intelligence. J. Clin. Neurosci. *64*, 277–282. https://doi.org/10.1016/j.jocn.2019.03.001.

57. Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds Mach. *30*, 99–120. https://doi.org/10.1007/s11023-020-09517-8.

58. Stahl, B.C., and Wright, D. (2018). Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation. IEEE Secur. Priv. *16*, 26–33. https://doi.org/10.1109/MSP.2018.2701164.

59. Chen, L., Cruz, A., Ramsey, S., Dickson, C.J., Duca, J.S., Hornak, V., Koes, D.R., and Kurtzman, T. (2019). Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. PLoS One *14*, e0220113. https://doi.org/10.1371/journal.pone.0220113.

60. Das, A., and Rad, P. (2020). Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. CoRR *Abs/2006*, p. 11371. https://doi.org/10.48550/arXiv.2006.11371.

61. Lipton, Z.C. (2018). The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery. Queue *16*, 31–57. https://doi.org/10.1145/3236386.3241340.

62. Union, E. (2016). EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Orkesterjournalen *L119*, 1–88.

63. Molnar, C. (2022). Interpretable Machine Learning, 2 Edition.

64. Vellido, A., MartnGuerrero, J.D., and Lisboa, P.J.G. (2012). Making machine learning models interpretable. 20th European Symposium on Artificial Neural Networks Held in Bruges, 163–172.

65. Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nat. Mach. Intell. *1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x.

66. Angelov, P., and Soares, E. (2020). Towards explainable deep neural networks (xDNN). Neural Network. *130*, 185–194. https://doi.org/10.1016/j.neunet.2020.07.010.

67. Richman, R., and Wüthrich, M.V. (2022). LocalGLMnet: interpretable deep learning for tabular data. Scand. Actuar. J. *2023*, 71–95. https://doi.org/10.1080/03461238.2022.2081816.

68. Zhang, X., Wang, N., Shen, H., Ji, S., Luo, X., and Wang, T. (2020). In Interpretable Deep Learning under Fire. inproceedings, S. Capkun and F. Roesner, eds. {USENIX}

Association), pp. 1659–1676. https://doi.org/10.48550/arXiv.1812.00891.

69. Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? inproceedings. https://doi.org/10.48550/arXiv.2207.08815.

70. Moradi, M., and Samwald, M. (2021). Post-hoc explanation of black-box classifiers using confident itemsets. Expert Syst. Appl. *165*, 113941. https://doi.org/10.1016/j.eswa.2020.113941.

71. Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. Bioinformatics *26*, 1340–1347. https://doi.org/10.1093/bioinformatics/btq134.

72. Cohen, S.B., Ruppin, E., and Dror, G. (2005). In Feature Selection Based on the Shapley Value. inproceedings, L.P. Kaelbling and A. Saffiotti, eds. (Professional Book Center), pp. 665–670.

73. Chipman, H.A., and Gu, H. (2005). Interpretable dimension reduction. J. Appl. Stat. *32*, 969–987. https://doi.org/10.1080/02664760500168648.

74. Zytek, A., Arnaldo, I., Liu, D., Berti-Equille, L., and Veeramachaneni, K. (2022). The Need for Interpretable Features: Motivation and Taxonomy. SIGKDD Explor. Newsl. *24*, 1–13. https://doi.org/10.1145/3544903.3544905.

75. Ahmad, M.A., Eckert, C., and Teredesai, A. (2019). In The Challenge of Imputation in Explainable Artificial Intelligence Models, H. E., H. Y., X. H., F. L., C. C., J. H.O., S. higeartaigh inproceedings and R. Mallah, eds. CEUR-WS.org. https://doi.org/10.48550/arXiv.1907.12669.

76. Marcílio-Jr, W.E., and Eler, D.M. (2021). Explaining dimensionality reduction results using Shapley values. Expert Syst. Appl. *178*, 115020. https://doi.org/10.1016/j.eswa.2021.115020.

77. (1980). Detecting Influential Observations and Outliers. In Regression DiagnosticsRegression Diagnostics, pp. 6–84. https://doi.org/10.1002/0471725153.ch2.

78. Sokol, K., and Flach, P.A. (2020). In Explainability fact sheets: a framework for systematic assessment of explainable approaches. inproceedings, M. Hildebrandt, C. Castillo, L.E. Celis, S. Ruggieri, L. Taylor, and G. ZanfirFortuna, eds. {ACM}), pp. 56–67. https://doi.org/10.1145/3351095.3372870.

79. Shapley, L.S. (1953). 17. A Value for n-Person Games. In Contributions to the Theory of Games (AM-28), Volume II, K. Harold William and T. Albert William, eds. (Princeton University Press), pp. 307–318. https://doi.org/10.1515/9781400881970-018.

80. Ke, G.Y., Hu, X.-F., and Xue, X.-L. (2022). Using the Shapley Value to mitigate the emergency rescue risk for hazardous materials. Group Decis. Negot. *31*, 137–152. https://doi.org/10.1007/s10726-021-09760-z.

81. Frmling, K., Westberg, M., Jullum, M., Madhikermi, M., and Malhi, A. (2021). In Comparison of Contextual Importance and Utility with LIME and Shapley Values. inproceedings, D. Calvaresi, A. Najjar, M. Winikoff, and K. Frmling, eds. (Springer), pp. 39–54. https://doi.org/10.1007/978-3-030-82017-6_3.

82. Lundberg, S.M., Nair, B., Vavilala, M.S., Horibe, M., Eisses, M.J., Adams, T., Liston, D.E., Low, D.K.W., Newman, S.F., Kim, J., and Lee, S.I. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat. Biomed. Eng. *2*, 749–760. https://doi.org/10.1038/s41551-018-0304-0.

83. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

84. Lamu, A.N., and Olsen, J.A. (2016). The relative importance of health, income and social relations for subjective well-being: An integrative analysis. Soc. Sci. Med. *152*, 176–185. https://doi.org/10.1016/j.socscimed.2016.01.046.

85. Cai, J., Chu, X., Xu, K., Li, H., and Wei, J. (2020). Machine learning-driven new material discovery. Nanoscale Adv. *2*, 3115–3130. https://doi.org/10.1039/d0na00388c.

86. Khmaissia, F., Frigui, H., Sunkara, M., Jasinski, J., Garcia, A.M., Pace, T., and Menon, M. (2018). Accelerating band gap prediction for solar materials using feature selection and regression techniques. Comput. Mater. Sci. *147*, 304–315. https://doi.org/10.1016/j.commatsci.2018.02.012.

87. Balachandran, P.V., Xue, D., Theiler, J., Hogden, J., Gubernatis, J.E., and Lookman, T. (2018). Importance of Feature Selection in Machine Learning and Adaptive Design for Materials. In Materials Discovery and Design: By Means of Data Science and Optimal Learning (Springer International Publishing), pp. 59–79. https://doi.org/10.1007/978-3-319-99465-9_3.

88. De Breuck, P.-P., Hautier, G., and Rignanese, G.-M. (2021). Materials property prediction for limited datasets enabled by feature selection and joint learning with MODNet. npj Comput. Mater. *7*, 83. https://doi.org/10.1038/s41524-021-00552-2.

89. Rodríguez-Pérez, R., and Bajorath, J. (2019). Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values. J. Med. Chem. *63*, 8761–8777. https://doi.org/10.1021/acs.jmedchem.9b01101.

90. Grimberg, H., Tiwari, V.S., Tam, B., Gur-Arie, L., Gingold, D., Polachek, L., and Akabayov, B. (2022). Machine learning approaches to optimize small-molecule inhibitors for RNA targeting. J. Cheminf. *14*, 4. https://doi.org/10.1186/s13321-022-00583-x.

91. Yazdani, K., Jordan, D., Yang, M., Fullenkamp, C.R., Calabrese, D.R., Boer, R., Hilimire, T., Allen, T.E.H., Khan, R.T., and Schneekloth, J.S., Jr. (2023). Machine Learning Informs RNA-Binding Chemical Space. Angew. Chem., Int. Ed. Engl. *62*, e202211358. https://doi.org/10.1002/anie.202211358.

92. Wojtuch, A., Jankowski, R., and Podlewska, S. (2021). How can SHAP values help to shape metabolic stability of chemical compounds? J. Cheminf. *13*, 74-20. https://doi.org/10.1186/s13321-021-00542-y.

93. Burroughs, L., Amer, M.H., Vassey, M., Koch, B., Figueredo, G.P., Mukonoweshuro, B., Mikulskis, P., Vasilevich, A., Vermeulen, S., Dryden, I.L., et al. (2021). Discovery of synergistic material-topography combinations to achieve immunomodulatory osteoinductive biomaterials using a novel in vitro screening method: The ChemoTopoChip. Biomaterials *271*, 120740. https://doi.org/10.1016/j.biomaterials.2021.120740.

94. Korolev, V.V., Mitrofanov, A., Marchenko, E.I., Eremin, N.N., Tkachenko, V., and Kalmykov, S.N. (2020). Transferable and Extensible Machine Learning-Derived Atomic Charges for Modeling Hybrid Nanoporous Materials. Chem. Mater. *32*, 7822–7831. https://doi.org/10.1021/acs.chemmater.0c02468.

95. Jablonka, K.M., Ongari, D., Moosavi, S.M., and Smit, B. (2020). Big-Data Science in Porous Materials: Materials Genomics and Machine Learning. Chem. Rev. *120*, 8066–8129. https://doi.org/10.1021/acs.chemrev.0c00004.

96. Guo, S., Huang, X., Situ, Y., Huang, Q., Guan, K., Huang, J., Wang, W., Bai, X., Liu, Z., Wu, Y., and Qiao, Z. (2023). Interpretable Machine-Learning and Big Data Mining to Predict Gas Diffusivity in Metal-Organic Frameworks. Adv Sci (Weinh). n/a *10*, e2301461. https://doi.org/10.1002/advs.202301461.

97. Maulana Kusdhany, M.I., and Lyth, S.M. (2021). New insights into hydrogen uptake on porous carbon materials via explainable machine learning. Carbon *179*, 190–201. https://doi.org/10.1016/j.carbon.2021.04.036.

98. Chai, M., Moradi, S., Erfani, E., Asadnia, M., Chen, V., and Razmjou, A. (2021). Application of Machine Learning Algorithms to Estimate Enzyme Loading, Immobilization Yield, Activity Retention, and Reusability of Enzyme–Metal–Organic Framework Biocatalysts. Chem. Mater. *33*, 8666–8676. https://doi.org/10.1021/acs.chemmater.1c02476.

99. Wang, J., Qi, Y., Zheng, H., Wang, R., Bai, S., Liu, Y., Liu, Q., Xiao, J., Zou, D., and Hou, S. (2023). Advancing vapor-deposited perovskite solar cells via machine learning. J. Mater. Chem. A *11*, 13201–13208. https://doi.org/10.1039/D3TA00027C.

100. Wu, H., Mayeshiba, T., and Morgan, D. (2016). High-throughput ab-initio dilute solute diffusion database. Sci. Data *3*, 160054. https://doi.org/10.1038/sdata.2016.54.

101. Yin, H., Sun, Z., Wang, Z., Tang, D., Pang, C.H., Yu, X., Barnard, A.S., Zhao, H., and Yin, Z. (2021). The data-intensive scientific revolution occurring where two-dimensional materials meet machine learning. Cell Reports Physical

Science *2*, 100482. https://doi.org/10.1016/j.xcrp.2021.100482.

102. Koh, P.W., and Liang, P. (2017). In Understanding Black-box Predictions via Influence Functions. inproceedings, D. Precup and Y.W. Teh, eds. {PMLR}), pp. 1885–1894. https://doi.org/10.48550/arXiv.1703.04730.

103. Ghassemi, M., Oakden-Rayner, L., and Beam, A.L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. Lancet. Digit. Health *3*, e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9.

104. Panch, T., Mattie, H., and Celi, L.A. (2019). The "inconvenient truth" about AI in healthcare. NPJ Digit. Med. *2*, 77. https://doi.org/10.1038/s41746-019-0155-4.

105. Ma, S., and Tourani, R. (2020). In Predictive and Causal Implications of using Shapley Value for Model Interpretation. inproceedings, T.D. Le, L. Liu, K. Zhang, E. Kiciman, P. Cui, and A. Hyvrinen, eds. {PMLR}), pp. 23–38. https://doi.org/10.48550/arXiv.2008.05052.

106. Leung, C.K., Madill, E.W.R., Souza, J., and Zhang, C.Y. (2022). Towards Trustworthy Artificial Intelligence in Healthcare. Inproceedings (IEEE), pp. 626–632. https://doi.org/10.1109/ICHI54592.2022.00127.

107. Alufaisan, Y., Marusich, L.R., Bakdash, J.Z., Zhou, Y., and Kantarcioglu, M. (2021). Does Explainable Artificial Intelligence Improve Human Decision-Making? Inproceedings ({AAAI} Press), pp. 6618–6626. https://doi.org/10.48550/arXiv.2006.11194.

108. Korolev, V., and Protsenko, P. (2023). Toward Accurate Interpretable Predictions of Materials Properties within Transformer Language Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2303.12188.