IEEE Access
Multidisciplinary Rapid Review Open Access Journal

Received January 21, 2020, accepted February 11, 2020, date of publication February 24, 2020, date of current version March 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2976199

Explainable Machine Learning for Scientific Insights and Discoveries

RIBANA ROSCHER[1,2], **(Member, IEEE), BASTIAN BOHN**[3],
MARCO F. DUARTE[4], **(Senior Member, IEEE), AND JOCHEN GARCKE**[3,5]

[1]Institute of Geodesy and Geoinformation, University of Bonn, 53115 Bonn, Germany
[2]Institute of Computer Science, University of Osnabrueck, 49074 Osnabrück, Germany
[3]Institute for Numerical Simulation, University of Bonn, 53115 Bonn, Germany
[4]Department of Electrical and Computer Engineering, University of Massachusetts Amherst, Amherst, MA 01003, USA
[5]Fraunhofer Center for Machine Learning and Fraunhofer SCAI, 53757 Sankt Augustin, Germany

Corresponding author: Jochen Garcke (jochen.garcke@scai.fraunhofer.de)

This work was supported in part by the Fraunhofer Cluster of Excellence Cognitive Internet Technologies, and in part by the Sonderforschungsbereich 1060-The Mathematics of Emergent Effects-funded by the Deutsche Forschungsgemeinschaft.

ABSTRACT Machine learning methods have been remarkably successful for a wide range of application areas in the extraction of essential information from data. An exciting and relatively recent development is the uptake of machine learning in the natural sciences, where the major goal is to obtain novel scientific insights and discoveries from observational or simulated data. A prerequisite for obtaining a scientific outcome is domain knowledge, which is needed to gain explainability, but also to enhance scientific consistency. In this article, we review explainable machine learning in view of applications in the natural sciences and discuss three core elements that we identified as relevant in this context: *transparency*, *interpretability*, and *explainability*. With respect to these core elements, we provide a survey of recent scientific works that incorporate machine learning and the way that explainable machine learning is used in combination with domain knowledge from the application areas.

INDEX TERMS Explainable machine learning, informed machine learning, interpretability, scientific consistency, transparency.

I. INTRODUCTION

Machine learning methods, especially with the rise of neural networks (NNs), are nowadays used widely in commercial applications. This success has led to a considerable uptake of machine learning (ML) in many scientific areas. Usually these models are trained with regard to high accuracy, but there is a recent and ongoing high demand for understanding the way a specific model operates and the underlying reasons for the decisions made by the model. One motivation behind this is that scientists increasingly adopt ML for optimizing and producing scientific outcomes. Here, explainability is a prerequisite to ensure the scientific value of the outcome. In this context, research directions such as explainable artificial intelligence (AI) [1], informed ML [2], or intelligible intelligence [3] have emerged. Though related, the concepts, goals, and motivations vary, and core technical terms are defined in different ways.

In the natural sciences, the main goals for utilizing ML are scientific understanding, inferring causal relationships from observational data, or even achieving new scientific insights. With ML approaches, one can nowadays (semi-)automatically process and analyze large amounts of scientific data from experiments, observations, or other sources. The specific aim and scientific outcome representation will depend on the researchers' intentions, purposes, objectives, contextual standards of accuracy, and intended audiences. Regarding conditions for an adequate scientific representation, we defer to the philosophy of science [4].

This article provides a survey of recent ML approaches that are meant to derive scientific outcomes, where we specifically focus on the natural sciences. Given the scientific outcomes, novel insights can be derived to deepen understanding, or scientific discoveries can be revealed that were not known before. *Gaining scientific insights and discoveries* from an ML algorithm means gathering information from its output and/or its parameters regarding the scientific process or experiments underlying the data.

The associate editor coordinating the review of this manuscript and approving it for publication was Massimo Cafaro.

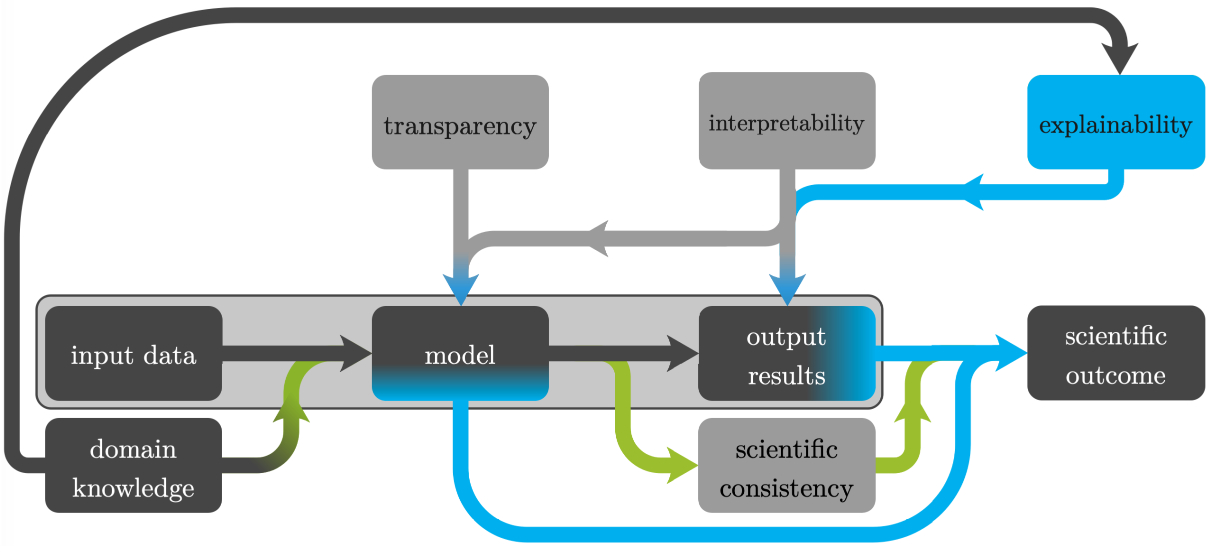

FIGURE 1. Major ML-based chains from which scientific outcomes can be derived: The commonly used, basic ML chain (light gray box) learns a black box model from given input data and provides an output. Given the black box model and input-output relations, a scientific outcome can be derived by explaining the output results utilizing domain knowledge. Alternatively, a transparent and interpretable model can be explained using domain knowledge leading to scientific outcomes. Additionally, the incorporation of domain knowledge can promote scientifically consistent solutions (green arrows).

One should note that a data-driven effort of scientific discovery is nothing new, but mimics the revolutionary work of Johannes Kepler and Sir Isaac Newton, which was based on a combination of data-driven and analytical work. As stated by [5],

> *Data science is not replacing mathematical physics and engineering, but is instead augmenting it for the twenty-first century, resulting in more of a renaissance than a revolution.*

What is new, however, is the abundance of high-quality data in the combination with scalable computational and data processing infrastructure.

The main contribution of this survey is the discussion of commonly used ML-based chains leading to scientific outcomes that have been used in the natural sciences (see Fig. 1). The three elements *transparency*, *interpretability*, and *explainability* play a central role. These concepts will be defined and discussed in detail in this survey. Another essential component is *domain knowledge*, which is necessary to achieve explainability, but can also be used to foster *scientific consistency* of the model and the result. We provide diverse examples from the natural sciences of approaches that can be related to these topics. Moreover, we define several groups of ML chains based on the presence of the components from Fig. 1. Our goal is to foster a better understanding and a clearer overview of ML algorithms applied to data from the natural sciences.

The paper is structured as follows. Section II discusses transparency, interpretability, and explainability in the context of this article. While these terms are more methodology-driven and refer to properties of the model and the algorithm, we also describe the role of additional information and domain knowledge, as well as scientific consistency. Section III highlights several applications in the natural sciences that use these concepts to gain new scientific insights, while organizing the ML workflows into characteristic groups based on the different uptakes of interpretability and explainability. Section IV closes the paper with a discussion.

II. TERMINOLOGY

Several descriptive terms are used in the literature about explainable ML with diverse meanings, e.g., [6]–[11]. Nonetheless, distinct ideas can be identified. For the purpose of this work, we distinguish between transparency, interpretability, and explainability. Roughly speaking, transparency considers the ML approach, interpretability considers the ML model together with data, and explainability considers the model, the data, and human involvement.

A. TRANSPARENCY

An ML approach is transparent if the processes that extract model parameters from training data and generate labels from testing data can be described and motivated by the approach designer. We say that the transparency of an ML approach concerns its different ingredients: the overall model structure, the individual model components, the learning algorithm, and how the specific solution is obtained by the algorithm. We propose to distinguish between *model transparency*, *design transparency*, and *algorithmic transparency*. Generally, to expect an ML method to be completely transparent in all aspects is rather unrealistic; usually there will be different degrees of transparency.

As an example, consider kernel-based ML approaches [12], [13]. The obtained model is transparent as it is given as a sum of kernel functions. The individual design component is the chosen kernel. Choosing between a linear or nonlinear kernel is typically a transparent design decision. However, using the common Gaussian kernel based on Euclidean distances can be a non-transparent design decision. In other words, it may not be clear why a given nonlinear kernel was chosen. Domain specific design choices can be made, in particular using suitable distance measures to replace the Euclidean distance, making the design of this model component (more) transparent. In the case of Gaussian process (GP) regression, the specific choice of the kernel can be built into the optimization of the hyper-parameters using the maximum likelihood framework [13]. Thereby, design transparency crosses over to algorithmic transparency. Furthermore, the obtained specific solution is, from a mathematical point of view, transparent. Namely, it is the unique solution of a convex optimization problem that can be reproducibly obtained [12], [13], resulting in algorithmic transparency. In contrast, approximations in the specific solution method such as early stopping, matrix approximations, stochastic gradient descent, and others, can result in (some) non-transparency of the algorithm.

As another example, consider NNs [14]. The model is transparent since its input-output relation and structure can be written down in mathematical terms. Individual model components, such as a layer of a NN, that are chosen based on domain knowledge can be considered as design transparent. Nonetheless, the layer parameters — be it their numbers, size, or nonlinearities involved — are often chosen in an ad-hoc or heuristic fashion and not motivated by domain knowledge; these decisions are therefore not design transparent. The learning algorithm is typically transparent, e.g., stochastic gradient descent can be easily written down. However, the choice of hyper-parameters such as learning rate, batch size, etc., has a more heuristic, non-transparent algorithmic nature. Due to the presence of several local minima, the solution is usually not easily reproducible; therefore, the obtained specific solution is not (fully) algorithmically transparent.

Our view is closely related to Lipton [9], who writes:

> *Informally, transparency is the opposite of opacity or "black-boxness". It connotes some sense of understanding the mechanism by which the model works. Transparency is considered here at the level of the entire model (simulatability), at the level of individual components such as parameters (decomposability), and at the level of the training algorithm (algorithmic transparency).*

An important contribution to the understanding of ML algorithms is their mathematical interpretation and derivation, which help to understand when and how to use these approaches. Classical examples are the Kalman filter or principal component analysis, where several mathematical derivations exist for each and enhance their understanding. Note that although there are many mathematical attempts to a better understanding of deep learning, at this stage "the [mathematical] interpretation of NNs appears to mimic a type of Rorschach test," according to [15].

Overall, we argue that transparency in its three forms does to a large degree not depend on the specific data, but solely on the ML method. But clearly, the obtained specific solution, in particular the "solution path" to it by the (iterative) algorithm, depends on the training data. The analysis task and the type of attributes usually play a role in achieving design transparency. Moreover, the choice of hyper-parameters might involve model structure, components, or the algorithm, while in an algorithmic determination of hyper-parameters the specific training data comes into play again.

B. INTERPRETABILITY

For our purposes, interpretability pertains to the capability of making sense of an obtained ML model. Generally, to interpret means "to explain the meaning of" or "present in understandable terms;"[1] see also [6]–[8]. We consider explanation distinct from interpretation, and focus here on the second aspect. Therefore, the aim of interpretability is to present some of the properties of an ML model in terms understandable to a human. Ideally, one could answer the question from [16]: "Can we understand what the ML algorithm bases its decision on?" Somewhat formally, [10] states:

> *An interpretation is the mapping of an abstract concept (e.g., a predicted class) into a domain that the human can make sense of.*

Interpretations can be obtained by way of understandable proxy models, which approximate the predictions of a more complex approach [7], [8]. Longstanding approaches involve decision trees or rule extraction [17] and linear models. In prototype selection, one or several examples similar to the inspected datum are selected, from which criteria for the outcome can be obtained. For feature importance, the weights in a linear model are employed to identify attributes that are relevant for a prediction, either globally or locally. For example, [18] introduced the model-agnostic approach LIME (Local Interpretable Model-Agnostic Explanations), which gives interpretation by creating locally a linear proxy model in the neighborhood of a datum, while the scores in layer-wise relevance propagation (LRP) are obtained by means of a first-order Taylor expansion of the nonlinear function [10]. A sensitivity analysis can be used to inspect how a model output (locally) depends upon the different input parameters [19]. Such an extraction of information from the input and the output of a learned model is also called *post hoc interpretability* [9] or *reverse engineering* [8]. Further details, types of interpretation, and specific realization can be found in recent surveys [7], [8], [20].

Visual approaches such as saliency masks or heatmaps show relevant patterns in the input based on feature importance, sensitivity analysis, or relevance scores to explain model decisions, and are employed in particular with deep learning approaches for image classification [10], [21], [22]. [23] introduces a formal notion for interpreting NNs where a set of input features is deemed relevant for a classi-

[1] https://www.merriam-webster.com/dictionary/interpret

fication decision if the expected classifier score remains nearly constant when randomising the remaining features. The authors prove that under this notion, the problem of finding small sets of relevant features is NP-hard, even when considering approximation within any non-trivial factor. This, on the one hand, shows the difficulty of algorithmically determining interpretations; on the other hand, it justifies the current use of heuristic methods in practical applications.

In unsupervised learning, the analysis goal can be a better understanding of the data, for example, by an interpretation of the obtained representation by linear or nonlinear dimensionality reduction [24], [25], or by inspecting the components of a low-rank tensor decomposition [26].

Note that, in contrast to transparency, to achieve interpretability the data is always involved. Although there are model-agnostic approaches for interpretability, transparency or retaining the model can assist in the interpretation. Furthermore, method-specific approaches depend on transparency. For example, layer-wise relevance propagation for NNs exploits the known model layout [10].

While the methods for interpretation allow the inspection of a single datum, [27] observes that it quickly becomes very time consuming to investigate large numbers of individual interpretations. As a step to automate the processing of the individual interpretations for a single datum, they employ clustering of heatmaps of many data to obtain an overall impression of the interpretations for the predictions of the ML algorithm.

Finally, note that the interpretable and human level understanding of the performance of an ML approach can result in a different choice of the ML model, algorithm, or data pre-processing later on.

C. EXPLAINABILITY

While research into explainable ML is widely recognized as important, a joint understanding of the concept of explainability still needs to evolve. Concerning explanations, it has also been argued that there is a gap of expectations between ML and so-called explanation sciences such as law, cognitive science, philosophy, and the social sciences [28].

While in philosophy and psychology explanations have been the focus for a long time, a concise definition is not available. For example, explanations can differ in completeness or the degree of causality. We suggest to follow a model from a recent review relating insights from the social sciences to explanations in AI [29], which places explanatory questions into three classes: (1) what–questions, such as ‘‘What event happened?’’; (2) how–questions, such as ‘‘How did that event happen?’’; and (3) why–questions, such as ‘‘Why did that event happen?’’. From the field of explainable AI we consider a definition from [10]:

> *An explanation is the collection of features of the interpretable domain, that have contributed for a given example to produce a decision (e.g., classification or regression).*

As written in [8], ‘‘[in explainable ML] these definitions assume implicitly that the concepts expressed in the understandable terms composing an explanation are self-contained and do not need further explanations.’’

On the other hand, we believe that a collection of interpretations can be an explanation only with further contextual information, stemming from domain knowledge and related to the analysis goal. In other words, explainability usually cannot be achieved purely algorithmically. On its own, the interpretation of a model — in understandable terms to a human — for an individual datum might not provide an explanation to understand the decision. For example, the most relevant variables might be the same for several data; however, it is possible that the important observation for an understanding of the overall predictive behavior is that when ranking variables with respect to their interpretation, different lists of relevant variables are determined for each datum. Overall, the result will depend on the underlying analysis goal. ‘‘Why is the decision made?’’ will need a different explanation than ‘‘Why is the decision for datum A different to (the nearby) datum B?’’.

In other words, for explainability, the goal of the ML ‘‘user’’ is very relevant. According to [20], there are essentially four reasons to seek explanations: to justify decisions, to (enhance) control, to improve models, and to discover new knowledge. For regulatory purposes it might be fine to have an explanation by examples or (local) feature analysis, so that certain ‘‘formal’’ aspects can be checked. But, to attain scientific outcomes with ML one wants an understanding. Here, the scientist is using the data, the transparency of the method, and its interpretation to explain the output results (or the data) using domain knowledge and thereby to obtain a scientific outcome.

Furthermore, we suggest differentiating between *scientific explanations* and *algorithmic explanations*. For scientific explanations, [30] identifies five broad categories to classify the large majority of objects that are explained in science: data, entities, kinds, models, and theories. Furthermore, it is observed that the existence of a unifying general account of scientific explanation remains an open question. With an algorithmic explanation, one aims to reveal underlying causes to the decision of an ML method. This is what explainable ML aims to address. In recent years, a focus on applying interpretation tools to better explain the output of an ML model can be observed. This can be seen in contrast to symbolic AI techniques, e.g., expert or planning systems, which in contrast are often seen as explainable per se. Hybrid systems of both symbolic and, so-called, connectionist AI, e.g., artificial NNs, are investigated to combine advantages from both approaches. For example, [31] proposes ‘‘object-oriented deep learning’’ with the goal to convert a NN to a symbolic description to gain interpretability and explainability. They state that generally in NNs, there is inherently no explicit representation of symbolic concepts like objects or events, but rather a feature-oriented representation, which is difficult to explain. In their representation, objects could be formulated to have disentangled and interpretable properties. Although not commonly used so far, their work is one example of a promising direction towards a higher explainability of models.

In the broader context, other properties that can be relevant when considering explainability of ML algorithms are safety/trust, accountability, reproducibility, transferability, robustness and multi-objective trade-off or mismatched objectives, see e.g. [6], [9]. For example, in societal contexts, reasons for a decision often matter. Typical examples are (semi-)automatic loan applications, hiring decisions, or risk assessment for insurance applicants, where one wants to know why a model gives a certain prediction and how one might be affected by those decisions. In this context, and also due to regulatory reasons, one goal is that decisions based on ML models involve fair and ethical decision making. The importance to give reasons for decisions of an ML algorithm is also high for medical applications, where a motivation is the provision of trust in decisions such that patients are comfortable with the decision made. All this is supported by the General Data Protection Regulation, which contains new rules regarding the use of personal information. One component of these rules can be summed up by the phrase "right to an explanation" [32]. Finally, for ML models deployed for decision-support and automation, in particular in potentially changing environments, an underlying assumption is that robustness and reliability can be better understood, or more easily realized, if the model is interpretable [9].

One should also observe that explanations can be used to manipulate. For illustration, [33] distinguishes between the intuitive scientist, who seeks to make the most accurate or otherwise optimal decision, and the intuitive lawyer, who desires to justify a preselected conclusion. With that in mind, one often aims for human-centric explanations of black-box models. There are simple or purely algorithmic explanations, e.g., based on emphasising relevant pixels in an image. In so-called slow judgements tasks, an explanation might more easily enforce confirmation biases. For example, using human-centric explanations as evaluation baselines can be biased towards certain individuals. Further, a review of studies of experimental manipulations that require people to generate explanations or imagine scenarios indicates that people express greater confidence in a possibility, although false, when asked to generate explanations for it or imagine the possibility [34].

D. DOMAIN KNOWLEDGE

As outlined, domain knowledge is an essential part of explainability; it is also essential for treating small data scenarios or for performance reasons. A taxonomy for the explicit integration of knowledge into the ML pipeline, dubbed *informed ML*, is proposed by [2]. Three aspects are involved:

- type of knowledge,
- representation and transformation of knowledge, and
- integration of knowledge into the ML approach.

See also the related works of [35], who use the term *theory-guided data science*, or *physics-informed learning* by [36]. For the purpose of this article, we follow [2], who arrange different types of knowledge along their degree of formality, from the sciences, over (engineering or production) process flow to world knowledge and finally individual (expert's) intuition. Knowledge can be assigned to several of the types in this incomplete list.

In the sciences, knowledge is often given in terms of mathematical equations, such as analytic expressions or differential equations, or as relations between instances and/or classes in the form of rules or constraints. Its representation can, for example, be in the form of ontologies, symmetries, or similarity measures. Knowledge can also be exploited by numerical simulations of models or through human interaction.

As ingredients of an ML approach, one considers the training data, the hypothesis space, the training algorithm, and the final model. In each of these, one can incorporate additional knowledge. Feature engineering is a common and longstanding way to incorporate knowledge into the training data, whereas using numerical simulations to generate (additional) training data is a modern phenomena. One common way to integrate knowledge into the hypothesis space is by choosing the structure of the model. Examples of this include defining a specific architecture of a NN or by choosing a structure of probability distributions that observes existing or non-existing links between variables. An example for the training phase is modifying the loss function according to additional knowledge, for example by adding a consistency term. Finally, the obtained model can be put in relation to existing knowledge, for example by checking known constraints for the predictions.

E. SCIENTIFIC CONSISTENCY

A fundamental prerequisite for generating reliable outcomes for scientific applications is scientific consistency. This means that the result obtained is plausible and consistent with existing scientific principles. The selection and formulation of the scientific principles to be met is based on domain knowledge, where the manner of integration is the core research question in areas such as informed ML. In the chain of Fig. 1, scientific consistency can be considered *a priori* at the model design stage or *a posteriori* by analysing the output results. As pointed out by [2], scientific consistency at the design stage can be understood as the result of a regularization effect, where various ways exist to restrict the solution space to scientifically consistent solutions. Reference [37] identifies scientific consistency besides interpretability as one of the five major challenges we need to tackle to successfully adopt deep learning approaches in the geosciences. Reference [35] underlines the importance of consistency by defining it as an essential component to measure performance:

> *One of the overarching visions of [theory-guided data science] is to include [...] consistency as a critical component of model performance along with training accuracy and model complexity. This can be summarized in a simple way by the*

TABLE 1. Group 1 includes approaches without any means of interpretability. In Group 2, a first level of interpretability is added by employing domain knowledge to design the models or explain the outcomes. Group 3 deals with specific tools included in the respective algorithms or applied to their outputs to make them interpretable. Finally, Group 4 lists approaches where scientific insights are gained by explaining the machine learning model itself.

group	transparency		interpretability		integration of domain knowledge			
	design	alg.	model	in-out	explaining model	explaining outcome	design	post-hoc check
1a	-	-	-	-	-	-	-	-
1b	×	-	-	-	-	-	-	-
1c	×	-	-	-	-	-	×	-
2a	×	-	-	-	-	×	×	-
2b	×	×	-	-	-	×	×	-
2c	×	-	-	-	-	×	×	×
3a	-	-	-	×	-	×	-	-
3b	×	-	-	×	-	×	-	-
3c	×	-	-	×	-	×	×	-
3d	×	-	×	-	-	×	×	-
3e	×	-	×	-	-	×	×	×
4a	×	-	×	-	×	×	×	-
4b	×	×	×	-	×	×	×	-

> *following revised objective of model performance [...]: Performance* $\propto$ *Accuracy* $+$ *Simplicity* $+$ *Consistency.*

They discuss several ways to restrict the solution space to physically consistent solutions, by (1) designing the model family, such as specific network architectures; (2) guiding a learning algorithm using, e.g., specific initializations, constraints, or (loss) regularizations; (3) refining the model output, e.g., using closed-form equations or model simulations; (4) formulating hybrid models of theory and ML, and (5) augmenting theory-based models using real data such as data assimilation or calibration.

Overall, the explicit restriction of the solution space to scientifically consistent and plausible solutions is not a requirement to achieve valuable scientific outcomes. Neglecting this restriction, however, means that a consistent and plausible solution cannot be guaranteed, even if an optimal result has been achieved from a mathematical point of view.

III. SCIENTIFIC OUTCOMES FROM MACHINE LEARNING

In this section, we review examples that use ML and strive for different levels of transparency, interpretability, or explainability to produce scientific outcomes. To structure the different ML chains following Fig. 1, we define common groups and describe representative papers for each. The core ingredient is the basic ML chain, in which a model is learned from given input data and with a specific learning paradigm, yielding output results utilizing the learned model. In order to derive a scientific outcome, either the output results or the model is explained, where interpretability is the prerequisite for explainability. Moreover, transparency is required to explain a model. Generally, providing domain knowledge to an algorithm means to enhance the input data, model, optimizer, output results, or any other part of the ML algorithm by using information gained from domain insights such as laws of nature and chemical, biological, or physical models [2]. Besides the purpose of explainability, integrating domain knowledge can help with model tractability and regularization in scenarios where not enough data is available. It might also increase the performance of a model or reduce computational time.

In Table 1, we specify the four major groups and several subgroups in more detail. We expect that examples for additional subgroups can be found, but that will not affect our core observations. In particular, we distinguish between the following components:

Transparency We consider a model to be design-transparent if the model, or a part of it, was chosen for specific reasons, usually due to knowledge from the application domain. We call a model algorithmically transparent if the determination of the solution is obvious and traceable. In view of reproducible science, it is not surprising that essentially all the examples we found can be considered to be model-transparent.

Interpretability We take a closer look at two types of interpretability. First, we consider model components, such as neurons in a NN or obtained latent variables, to be interpretable if they are represented in a way that can be further explained, e.g., with domain knowledge. Second, the scientific outcome, i.e., the decision of the model, can be interpreted using the input, e.g., by using heatmaps.

Integration of domain knowledge We will look at several ways in which domain knowledge can be integrated. On the one hand, domain knowledge is needed to explain — either to explain the scientific outcome or to derive scientific findings from the model or individual model components. On the other hand, domain knowledge can be integrated to enforce scientifically plausible and consistent results. This can be done in different ways; cf. [2]. Besides the integration of domain knowledge during the learning process of the model, it can also be used for post-hoc checks, where the scientific plausibility and consistency of the results is checked and possibly invalid results are removed.

The following collection of research works is a non-exhaustive selection from the literature of the last few years, where we aim to cover a broad range of usages of ML with a variety of scientific outcomes. Furthermore, we focus on examples that utilize an extensive amount of scientific domain knowledge from the natural sciences. Due to the recent uptake of NNs in the sciences, these tend to be the dominating ML approach in current literature. Nonetheless, many of the described ML workflows or the approaches to integrate domain knowledge can be performed with other ML methods as well. Note that the choice of a group for a given article is not always a clear judgement, particularly in view of how and where domain knowledge is employed, and in what form, and to what extent, an explanation is derived.

A. SCIENTIFIC OUTCOMES BY EXPLAINING OUTPUT RESULTS

Many works address the derivation of outcomes by learning an ML model and generalizing from known input-output relations to new input-output pairs. This represents the lowest degree of explainability without the need for a transparent or interpretable model. In the case that a scientifically useful outcome is to be estimated, most of these approaches so far solely explain what the outcome is from a scientific point of view (scientific explanation), but cannot answer the question of why this specific outcome was obtained from an algorithmic point of view (algorithmic explanation). Other approaches attempt to scientifically explain the output in terms of the specific corresponding input, given a learned model. Here, interpretation tools are utilized, where the model is used only as a means to an end to explain the result and it is not explicitly analyzed itself.

1) PREDICTION OF INTUITIVE OUTCOMES

The derivation of intuitive physics is a task that is often considered in papers from the following group. Intuitive physics are everyday-observed rules of nature that help us to predict the outcome of events even with a relatively untrained human perception [38].

Group 1a (Basic ML Chain): A black-box approach, or at most model-transparent approach, is used to derive an outcome. It is not interpretable and cannot be explained. The outcome is not or only minimally explainable from a scientific point of view.

For example, [39] uses video simulations to learn intuitive physics, e.g., about the stability of wooden block towers. They use ResNet-34 and GoogLeNet and formulate a binary classification task to predict whether these towers will fall. In a similar way, but with more complex scenes or differently shaped objects, [40] predicts the physical stability of stacked objects using various popular convolutional neural network (CNN) architectures. Reference [41] predicts the spread of diseases on barley plants in microscopic hyperspectral images by generating highly probable image-based appearances over the course of several days. They use cycle-consistent generative adversarial networks to learn how an image will change from one day to the next or to the previous day, albeit without any biological parameters involved. Reference [42] presents an approach for the design of new functional glasses that comprises the prediction of characteristics relevant for manufacturing as well as end-use properties of glass. They utilize NNs to estimate the liquidus temperatures for various silicate compositions consisting of up to eight different components. Generally, the identification of an optimized composition of the silicates yielding a suitable liquidus temperature is a costly task and is oftentimes based on trial-and-error. For this, they learn from several hundred composites with known output properties and apply the model to novel, unknown composites. In their workflow, they also consider, outside of the ML chain, other quantities of interest that are derived by physics-driven models. Reference [43] proposes a nonlinear regression approach employing NNs to learn closed form representations of partial differential equations (PDEs) from scattered data collected in space and time, thereby uncovering the dynamic dependencies and obtaining a model that can be subsequently used to forecast future states. In benchmark studies, using Burgers' equation, nonlinear Schrödinger equation, or Navier-Stokes equation, the underlying dynamics are learned from numerical simulation data up to a specific time. The obtained model is used to forecast future states, where relative L_2-errors of up to the order of 10^{-3} are observed. While the method inherently models the PDEs and the dynamics themselves, the rather general network model does not allow the drawing of direct scientific conclusions on the structure of the underlying process.

Group 1b: These models are not only model- but also design-transparent to some extent, where the design is chosen and motivated with certain intentions.

Besides the simple prediction network presented by [39] in Group 1a, they also propose a network called PhysNet to predict the trajectory of the wooden blocks in case the tower is collapsing. It is formulated as a mask prediction network trained for instance segmentation, where each wooden block is defined as one class. The construction of PhysNet is made design-transparent in the sense that the network is constructed to capture the arrangement of blocks by using alternating upsampling and convolution layers, and an increased depth to reason about the block movement, as well. PhysNet outperforms human subjects on synthetic data and achieves comparable results on real data. Reference [44] designed a multi-source NN for exoplanet transit classification. They integrate additional information by adding identical information about centroid time-series data to all input sources, which is assumed to help the network learn important connections, and by concatenating the output of a hidden layer with stellar parameters, as it is assumed they are correlated with classification. Reference [45] introduces a framework that calculates physical concepts from color-depth videos that estimates tool and tool-use such as cracking a nut. In their work, they learn task-oriented representations for each tool and task combination defined over a graph with spatial, temporal, and causal relations. They distinguish between 13 physical concepts, e.g., painting a wall, and show that the framework is able to generalize from known to unseen concepts by selecting appropriate tools and tool-uses. A hybrid approach is presented by [46] to successfully model the properties of contaminant dispersion in soil. The authors extract temporal information from dynamic data using a long short-term memory network and combine it with static data in a NN. In this way, the network models the spatial correlations underlying the dispersion model, which are independent of the location of the contaminant. Reference [47] has proposed a data-centric approach for scientific design based on the combination of a generative model for the data being considered, e.g., an autoencoder trained on genomes or proteins, and a predictive model for a quantity or property of interest, e.g., disease indicators or protein fluorescence. For DNA sequence design, these two components are integrated by applying the predictive model to samples from the generative model. In this way, it is possible to generate new synthetic data samples that optimize the value of the quantity or property by leveraging an adaptive sampling technique over the generative model; see also [48].

Group 1c: Here, in addition to Group 1b, the design process is influenced by domain knowledge regarding the model, the cost function, or the feature generation process with the particular aim to enhance scientific consistency and plausibility.

For example, [49], [50] use physics-informed approaches for applications such as fluid simulations based on the incompressible Navier-Stokes equations, where physics-based losses are introduced to achieve plausible results. The idea in [49] is to use a transparent cost function design by reformulating the condition of divergence-free velocity fields into an unsupervised learning problem at each time step. The random forest model used in [50] to predict a fluid particle's velocity can be viewed as a transparent choice per se due to its simple nature. Reference [51] classifies land use and land cover and their changes based on remote sensing satellite timeseries data. They integrate domain knowledge about the transition of specific land use and land cover classes, such as forest or burnt areas, to increase the classification accuracy. They utilize a discriminative random field with transition matrices that contain the likelihoods of land cover and land use changes to enforce, for example, that a transition from burnt area to forest is unlikely.

2) PREDICTION OF SCIENTIFIC PARAMETERS AND PROPERTIES

Although the approaches just described set up prediction as a supervised learning problem, there is still a gap between common supervised tasks, e.g., classification, object detection, and prediction, and actual understanding of a scene and its reasoning. Like a layman in the corresponding scientific field, the methods presented so far do not learn a model that is able to capture and derive scientific properties and dynamics of phenomena or objects and their environment, as well as their interactions. Therefore, the model cannot inherently explain why an outcome was obtained from a scientific viewpoint. Reference [52] labels these respective approaches as bottom-up, where observations are directly mapped to an estimate of some behavior or some outcome of a scenario. To tackle the challenge of achieving a higher explainability and a better scientific usability, several so-called top-down classification and regression frameworks have been formulated that infer scientific parameters. In both cases, only the scientific explanation is sought.

Group 2a: In these ML models, domain knowledge is incorporated, often to enforce scientific consistency. Therefore, the design process is partly transparent and tailored to the application. The outcome is explainable from a scientific point of view since scientific parameters and properties are derived, which can be used for further processing.

For example, [53] detects and tracks objects in videos in an unsupervised way. The authors use a regression CNN and introduce terms during training that measure the consistency of the output when compared to physical laws which specifically and thoroughly describe the dynamics in the video. In this case, the input of the regression network is a video sequence and the output is a time-series of physical parameters such as the height of a thrown object. By incorporating domain knowledge and image properties into their loss functions, part of their design process becomes transparent and explainability is gained due to comparisons to the underlying physical process. However, the model and the algorithms are not completely transparent since standard CNNs with an ADAM minimizer are employed. Although these choices of model and algorithms are common in ML, they are usually motivated by their good performance, and not because there is any application-driven reasoning behind it; thus, there is no design transparency on this aspect. Furthermore, the reason why such choice works for this highly nonconvex problem is currently not well understood from a mathematical point of view; therefore, no algorithmic transparency is present. Reference [54] introduces Physics101, a dataset of over 17,000 video clips containing 101 objects of different characteristics, which was built for the task of deriving physical parameters such as velocity and mass. In their work, they use the LeNet CNN architecture to capture visual as well as physical characteristics while explicitly integrating physical laws based on material and volume to aim for scientific consistency. Their experiments show that predictions can be made about the behavior of an object after a fall or a collision using estimated physical characteristics, which serve as input to an independent physical simulation model. Reference [55] introduces SMASH, which extracts physical collision parameters from videos of colliding objects, such as pre- and post collision velocities, to use them as inputs for existing physics engines for modifications. For this, they estimate the position and orientation of objects in videos using constrained least-squares estimation in compliance with physical laws such as momentum conservation. Based on the determined trajectories, parameters such as velocities can be derived. While their approach is based more on statistical parameter estimation than ML, their model and algorithm building process is completely transparent. Individual outcomes become explainable due to the direct relation of the computations to the underlying physical laws.

Reference [56] introduces Newtonian NNs in order to predict the long-term motion of objects from a single color image. Instead of predicting physical parameters from the image, they introduce 12 Newtonian scenarios serving as physical abstractions, where each scenario is defined by physical parameters defining the dynamics. The image, which contains the object of interest, is mapped to a state in one of these scenarios that best describes the current dynamics in the image. Newtonian NNs are two parallel CNNs: one encodes the images, while the other derives convolutional filters from videos acquired with a game engine simulating each of the 12 Newtonian scenarios. The specific coupling of both CNNs in the end leads to an interpretable approach, which also (partially) allows for explaining the classification results of a single input image.

A tensor-based approach to ML for uncertainty quantification problems can be found in [57], where the solutions to parametric convection-diffusion PDEs are learned based on a few samples. Rather than directly aiming for interpretability or explainability, this approach helps to speed up the process of gaining scientific insight by computing physically relevant quantities of interest from the solution space of the PDE. As there are convergence bounds for some cases, the design process is to some extent transparent and benefits from domain knowledge.

An information-based ML approach using NNs to solve an inverse problem in biomechanical applications was presented in [58]. Here, in mechanical property imaging of soft biological media under quasi-static loads, elasticity imaging parameters are computed from estimated stresses and strains. For a transparent design of the ML approach, domain knowledge is incorporated in two ways. First, NNs for a material model are pre-trained with stress-strain data, generated using linear-elastic equations, to avoid non-physical behavior. Second, finite-element analysis is used to model the data acquisition process.

Group 2b: These ML models are highly transparent, which means that the design process as well as the algorithmic components are fully accessible. The outcome of the model is explainable and the scientific consistency of the outcome is enforced.

For organic photovoltaics material, an approach utilizing quantum chemistry calculations and ML techniques to calibrate theoretical results to experimental data was presented by [59], [60]. The authors consider previously performed experiments as current knowledge, which is embedded within a probabilistic non-parametric mapping. In particular, GPs were used to learn the deviation of properties calculated by computational models from the experimental analogues. By employing the chemical Tanimoto similarity measure and building a prior based on experimental observations, design transparency is attained. Furthermore, since the prediction results involves a confidence in each calibration point being returned, the user can be informed when the scheme is being used for systems for which it is not suited [59]. In [60], 838 high-performing candidate molecules have been identified within the explored molecular space thanks to the newly possible efficient screening of over 51,000 molecules.

In [61], a data-driven algorithm for learning the coefficients of general parametric linear differential equations from noisy data was introduced, solving a so-called inverse problem. The approach employs GP priors that are tailored to the corresponding and known type of differential operators, resulting in design and algorithmic transparency. The combination of rather generic ML models with domain knowledge in the form of the structure of the underlying differential equations leads to an efficient method. Besides classical benchmark problems with different attributes, the approach was used on an example application in functional genomics, determining the structure and dynamics of genetic networks based on real expression data.

Group 2c: These ML models are similar to the models in Group 2a, but besides enforced scientific consistency and plausibility of the explainable outcome, an additional post-hoc consistency check is performed.

In [62], a deep learning approach for Reynolds-averaged Navier-Stokes (RANS) turbulence modelling was presented. Here, domain knowledge led to the construction of a network architecture that embedded invariance using a higher-order multiplicative layer. This was shown to result in significantly more accurate predictions compared to a generic, but less interpretable, NN architecture. Further, the improved prediction on a test case that had a different geometry than any of the training cases indicates that improved RANS predictions for more than just interpolation situations seem achievable. A related approach for RANS-modeled Reynolds stresses for high-speed flat-plate turbulent boundary layers was presented in [63], which uses a systematic approach with basis tensor invariants proposed by [64]. Additionally, a metric of prediction confidence and a nonlinear dimensionality reduction technique are employed to provide *a priori* assessment of the prediction confidence.

3) INTERPRETATION TOOLS FOR SCIENTIFIC OUTCOMES

Commonly used feature selection and extraction methods enhance the interpretability of the input data, and thus can lead to outcomes that can be explained by interpretable input. Other approaches use interpretation tools to extract information from learned models and to help to scientifically explain the individual output or several outputs jointly. Often, approaches are undertaken to present this information via feature importance plots, visualizations of learned representations, natural language representations, or the discussion of examples. Nonetheless, human interaction is still required to interpret this additional information, which has to be derived ante-hoc or with help of the learned model during a post-hoc analysis.

Group 3a: These ML approaches use interpretation tools to explain the outcome by means of an interpretable representation of the input. Such tools include feature importance plots or heatmaps.

While handcrafted and manually selected features are typically easier to understand, automatically determined features can reveal previously unknown scientific attributes and structures. Reference [65], for example, proposes FINE (feature importance in nonlinear embeddings) for the analysis of cancer patterns in breast cancer tissue slides. This approach relates original and automatically derived features to each other by estimating the relative contributions of the original features to the reduced-dimensionality manifold. This procedure can be combined with various, possibly intransparent, nonlinear dimensionality reduction techniques. Due to the feature contribution detection, the resulting scheme remains interpretable.

Arguably, visualizations are one of the most widely used interpretation tools. Reference [21] gives a survey of visual analytics in deep learning research, where such visualization systems have been developed to support model explanation, interpretation, debugging, and improvement. The main consumers of these analytics are the model developers and users as well as non-experts. Reference [66] uses interpretation tools for image-based plant stress phenotyping. The authors train a CNN model and identify the most important feature maps in various layers that isolate the visual cues for stress and disease symptoms. They produce so-called explanation maps as sums of the most important features maps indicated by their activation level. A comparison of manually marked visual cues by an expert and the automatically derived explanation maps reveals a high level of agreement between the automatic approach and human ratings. The goals of their approach are the analysis of the performance of their model, the provision of visual cues that are human-interpretable to support the prediction of the system, and a provision of important cues for the identification of plant stress. References [39] and [67] use attention heatmaps to visualize the stackability of multiple wooden blocks in images. They conduct a conclusion study by applying localized blurring to the image and collecting the resulting changes in the stability classification of the wooden blocks into a heatmap. Moreover, [67] provides a first step towards a physics-aware model by using their trained stability predictor and heatmap analysis to provide stackability scores for unseen object sets, for the estimation of an optimal placement of blocks, and to counterbalance instabilities by placing additional objects on top of unstable stacks.

As another example, ML has been applied to functional magnetic resonance imaging data to design biomarkers that are predictive of psychiatric disorders. However, only ''surrogate'' labels are available, e.g., behavioral scores, and so the biomarkers themselves are also ''surrogates'' of the optimal descriptors [68], [69]. The biomarker design promotes spatially compact pixel selections, producing biomarkers for disease prediction that are focused on regions of the brain. These are then assessed by expert physicians. As the analysis is based on high-dimensional linear regression approaches, transparency of the ML model is assured. Reference [70] introduces DeepTune, a visualization framework for CNNs, for applications in neuroscience. DeepTune consists of an ensemble of CNNs that learn multiple complementary representations of natural images. The features from these CNNs are fed into regression models to predict the firing rates of neurons in the visual cortex. The interpretable DeepTune images, i.e., representative images of the visual stimuli for each neuron, are generated from an optimization process and pooling over all ensemble members.

Classical tools such as confusion matrices are also used as interpretation tools on the way to scientific outcomes. In a bio-acoustic application for the recognition of anurans using acoustic sensors, [71] uses a hierarchical approach to jointly classify on three taxonomic levels, namely the family, the genus, and the species. Investigating the confusion matrix per level enabled for example the identification of bio-acoustic similarities between different species.

Group 3b: These models are design-transparent in the sense that they use specially tailored components such as attention modules to achieve increased interpretability. The output is explained by the input using the specially selected components.

In [72], [73] attention-based NN models are employed to classify and segment histological images, e.g., microscopic tissue images, magnetic resonance imaging (MRI), or computed tomography (CT) scans. Reference [72] found that the employed modules turned out to be very attentive to regions of pathological, cancerous tissue and non-attentive in other regions. Furthermore, [73] builds an attentive gated network that gradually fitted its attention weights with respect to targeted organ boundaries in segmenting tasks. The authors also used their attention maps to employ a weakly supervised object detection algorithm, which successfully created bounding boxes for different organs.

Interpretability methods have also been used for applications that utilize time-series data, often by way of highlighting features of the sequence data. For example, [74] applies attention modules in NNs trained on genomic sequences for the identification of important sequence motifs by visualizing the attention mask weights. They propose a genetic architect that finds a suitable network architecture by iteratively searching over various NN building blocks. In particular, they state that the choice of the NN architecture highly depends on the application domain, which is a challenge if no prior knowledge is available about the network design. It is cautioned that, depending on the optimized architecture, attention modules and expert knowledge may lead to different scientific insights. Reference [75] uses attention modules for genomics in their AttentiveChrome NN. The network contains a hierarchy of attention modules to gain insights about where and on what the network has focused on and, thus, gain interpretability of the results. Also [76] developed a hierarchical attention-based interpretation tool called RETAIN (REverse Time AttentIoN) in healthcare. The tool identifies influential past visits of a patient as well as important clinical variables during these visits from the patient's medical history to support medical explanations. Attention modules in recurrent NNs for multi-modal sensor-based activity recognition have been used by [77]. Depending on the activity, their approach provides the most contributing body parts, modals, and sensors for the network's decision.

Group 3c: As in Group 3b, these ML approaches use interpretation tools for a better understanding of the model's decision. Moreover, they integrate domain knowledge to enhance the scientific consistency and plausibility, for example, in combination with the outcome of interpretation tools.

For example, [78] discusses explainable ML for scientific discoveries in material sciences. In their work, they propose an ensemble of simple models to predict material properties along with a novel evaluation focusing on trust by quantifying generalization performance. Moreover, their pipeline contains a rationale generator that provides decision-level interpretations for individual predictions and model-level interpretations for the whole regression model. In detail, they produce interpretations in terms of prototypes that are analyzed and explained by an expert, as well as global interpretations by estimating feature importance for material sub-classes. Moreover, they use domain knowledge for the definition of material sub-classes and integrate it into the estimation process. Reference [79] proposes contextual decomposition explanation penalization, which constrains a classification or regression result to more accurate and more scientifically plausible results by leveraging the explained outcome of interpretation tools. They add an explanation term in the loss function, which compares the interpretation outcome (e.g., a heatmap indicating the important parts in the image) and an interpretation provided by the user. They determine a more accurate model from an International Skin Imaging Collaboration dataset whose goal is to classify cancerous and non-cancerous images, by learning that colorful patches present only in the benign data are not relevant for classification.

Group 3d: These approaches use the common feature-oriented representation with focus on the disentanglement of the underlying factors of variation in a system, which can be explained by an expert afterwards. Domain knowledge is employed in the design of the model and in the interpretation of the outcome.

A broad framework leverages unsupervised learning approaches to learn low-complexity representations of physical process observations. In many cases where the underlying process features a small number of degrees of freedom, it is shown that nonlinear manifold learning algorithms are able to discern these degrees of freedoms as the component dimensions of low-dimensional nonlinear manifold embeddings, which preserve the underlying geometry of the original data space [80]–[82]. It can be seen that the embedding coordinates relate to known physical quantities. At this stage, ongoing research is focused on obtaining new scientific outcomes in new situations using this promising approach. In a similar way, [83] and [84] use principal component analysis and the derived interpretable principal components for exploration of different phases, phase-transition, and crossovers in classical spin models. Embedded feature selection schemes have been recently explored to establish or refine models in physical processes. Using a sparsity-promoting penalty, they propose groups of variables that may explain a property of interest and promote the simplest model, i.e., the model involving the fewest variables possible while achieving a target accuracy. Domain knowledge is employed during the selection of the dictionary of candidate features. The application of sparsity has also proved fruitful in the broader class of problems leveraging PDEs and dynamical system models [85]–[89].

The combination of parse trees with ML is investigated in [90], [91]. A so-called syntax-direct variational autoencoder is introduced in [90], where syntax and semantic constraints are used in a generative model for structured data. As an application, the drug properties of molecules are predicted. The learned latent space is visually interpreted, while the diversity of the generated molecules is interpreted using domain expertise. The work in [91] uses a NN during a Monte Carlo tree search to guide its finding of an expression for symbolic regression that conforms to a set of data points and has the desired leading polynomial powers of the data. The NN learns the relation between syntactic structure and leading powers. As a proof-of-concept application, the authors are able to learn a physical force field, where the leading powers in the short and long ranges are known by domain experts and can be used as asymptotic constraints. Reference [92] proposes a sparsity-enforcing technique to recover domain-specific meaning for the abstract embedding coordinates obtained from unsupervised nonlinear dimensionality reduction approaches in a principled fashion. The ansatz is to explain the embedding coordinates as nonlinear compositions of functions from a user-defined dictionary. As an illustrative example, the ethanol molecule is studied, where the approach identifies the bond torsions that explain the torus obtained from the embedding method, which reflects the two rotational degrees of freedom.

Group 3e: In addition to the works in Group 3d, domain knowledge is employed to perform a posteriori consistency checks on feature-oriented representations.

Feature selection schemes using embedded methods, similar to the previous group, have been used in areas such as material sciences [93], [94]. In contrast to the preceding works, additional consistency checks on the outcome of the predictive model are performed based on domain expertise, including the robustness of the model and in particular their extrapolation capability for predicting new materials.

B. SCIENTIFIC OUTCOMES BY EXPLAINING MODELS

In the following examples, either interpretation tools are used to project processes in the model into a space that is interpretable or the model is designed inherently to be interpretable. In this way, models and their components can be explained utilizing domain knowledge.

Group 4a: These models are designed in a transparent way and the model design enforces that model components are interpretable and scientifically explainable. Due to their design, scientific consistency and plausibility is enforced, even if not as a primary goal. The explanation of specific model components is meant to lead to novel scientific discoveries or insights.

Complex ML methods such as NNs, for example, can be customized to a specific scientific application so that the used architecture restricts or promotes properties that are desirable in the data modeled by the network. For example, in [95], an application of ML for epidemiology leverages a networked dynamical system model for contagion dynamics, where nodes correspond to subjects with assigned states; thus, most properties of the ML model match the properties of the scientific domain considered. A complex NN is reduced by [96] to understand processes in neuroscience. By reducing the number of units in the complex model by means of a quantified importance utilizing gradients and activation values, a simple NN with one hidden layer is derived that can be easily related to neuroscientific processes. Reference [97] constructs a NN for computing Koopman eigenfunctions from data. Motivated by domain knowledge, the authors employ an auxiliary network to parameterize the continuous frequency. Thereby, a compact autoencoder model is obtained, which, in addition, is interpretable. For the example of the nonlinear pendulum, two eigenfunctions are learned with a NN, which can be mapped into magnitude and phase coordinates. In this interpretable form, it can be observed that the magnitude traces level sets of the Hamiltonian energy, a new insight that turned out to be consistent with recent theoretical derivations previously unknown to the authors. Reference [98] introduces visible NNs, which encode the hierarchical structure of a gene ontology tree into a NN, either from literature or inferred from large-scale molecular data sets. This enables transparent biological interpretation, while successfully predicting effects of gene mutations on cell proliferation. Furthermore, it is argued that the employed deep hierarchical structure captures many different clusters of features at multiple scales and pushes interpretation from the model input to internal features representing biological subsystems. In their work, despite no information about subsystem states being provided during model training, previously undocumented learned subsystem states could be confirmed by molecular measurements.

Beside NNs, other ML algorithms can also be used to derive scientific outcomes from an interpretable model. Reference [99] use the ML algorithm ‘Sir Isaac’ to infer a dynamical model of biological time-series data to understand and predict dynamics of worm escape behavior. They model a system of differential equations, where the number of hidden variables is determined automatically from the system, and their meaning can be explained by an expert.

Reference [100] introduces SciNet, a modified variational autoencoder that learns a representation from experimental data and uses the learned representation to derive physical concepts from it rather than from the experimental input data. The learned representation is forced to be much simpler than the experimental data, for example by being captured in a few neurons, and it contains the explanatory factors of the system, such as the physical parameters. This is proven by the fact that physical parameters and the activations of the neurons in the hidden layers have a linear relationship. Additionally, [101] constructs the bottleneck layer in their NN to represent physical parameters to predict the outcome of a collision of objects from videos. However, the architecture of the bottleneck layer is not learned, but designed with prior knowledge about the underlying physical process.

Understanding structures such as groups, relations, and interactions is one of the main goals to achieve scientific outcomes. However, it constitutes a core challenge, and so far only limited amount of work has been conducted in this area. Reference [102], for example, introduces a grouping layer in a graph-based NN called GroupINN to identify subgroups of neurons in an end-to-end model. In their work, they build a network for the analysis of time-series of functional magnetic resonance images of the brain, which are represented as functional graphs, with the goal of revealing relationships between highly predictive brain regions and cognitive functions. Instead of working with the whole functional graph, they exploit a grouping layer in the network to identify groups of neurons, where each neuron represents a node in the graph and corresponds to a physical region of interest in the brain. The grouped nodes in the coarsened graph are assigned to regions of interest, which are useful for prediction of cognitive functions, and the connections between the groups are defined as functional connections.

Reference [103] introduces neural interaction detection, a framework with variants of feedforward NNs for detecting statistical interactions. By examining the learned weight matrices of the hidden units, their framework was able to analyze feature interactions in a Higgs boson dataset. Specifically, they analyze feature interactions in simulated particle environments that originate from the decay of a Higgs boson. Deep tensor networks are used by [104] in quantum chemistry to predict molecular energy up to chemical accuracy, while allowing interpretations. A so-called local chemical potential, a variant of sensitivity analysis where one measures the effect on the NN output of inserting a charge at a given location, can be used to gain further chemical insight from the learned model. As an example, a classification of aromatic rings with respect to their stability can be determined from these three-dimensional response maps.

Group 4b: These ML models are designed with a high degree of transparency and with the goal to derive scientifically plausible results. Due to this, the outcome of the model and the model components themselves are interpretable and *can be scientifically explained. In contrast to the works presented in group 4a, the following examples employ methods that are also algorithmically transparent.*

Different types of physics-aware GP models in remote sensing were studied by [105] with the goal to estimate bio-physical parameters such as leaf area index. In one case, a latent force model that incorporates ordinary differential equations is used in inverse modelling from real in-situ data. The learned latent representation allowed an interpretation in view of the physical mechanism that generated the input-output observed relations, i.e., one latent function captured the smooth and periodic component of the output, while two other focus on the noisier part with an important residual periodical component. So-called order parameters in condensed matter physics are analysed in [106], [107]. Using domain knowledge, a kernel is introduced to investigate $O(3)$-breaking orientational order. A two-class and a multi-class setting are tackled with support vector machines (SVM). The decision function is physically interpreted as an observable corresponding to an order parameter curve, while the bias-term of the SVM can be exploited to detect phase transitions. Furthermore, nontrivial blocks of the SVM kernel matrices can be identified with so-called spin color indices. In these works, the analytical order parameters for spin and orbital systems could be extracted.

C. RELATED SURVEYS ABOUT MACHINE LEARNING IN THE NATURAL SCIENCES

Reference [108] gives on overview on recent research using ML for molecular and materials science. Given that standard ML models are numerical, the algorithms need suitable numerical representations that capture relevant chemical properties, such as the Coulomb matrix and graphs for molecules, and radial distribution functions that represent crystal structures. Supervised learning systems are in common use to predict numerical properties of chemical compounds and materials. Unsupervised learning and generative models are being used to guide chemical synthesis and compound discovery processes, where deep learning algorithms and generative adversarial networks have been successfully employed. Alternative models exploiting the similarities between organic chemistry and linguistics are based on textual representations of chemical compounds.

A review on the manifold recent research topics in the physical sciences is given by [109], with applications in particle physics and cosmology, quantum many-body physics, quantum computing, and chemical and material physics. The authors observe a surge of interest in ML, while noting that the research is starting to move from exploratory efforts on toy models to the use of real experimental data. It is stressed that an understanding of the potential and the limitations of ML includes insight into the breaking point of these methods, but also the theoretical justification of the performance in specific situations, be it positive or negative.

In single-cell genomics, computational data-driven analysis methods are employed to reveal the diverse simultaneous facets of a cell's identity, including a specific state on a developmental trajectory, the cell cycle, or a spatial context. The analysis goal is to obtain an interpretable representation of the dynamic transitions a cell undergoes that allows a determination of different aspects of cellular organization and function. There is an emphasis on unsupervised learning approaches to cluster cells from single-cell profiles, and thereby to systematically detect previously unknown cellular subtypes. Defining markers for these subtypes are then investigated in a second step. See [110] for a review on key questions, progress, and open challenges in this application field.

Several ML approaches have been used in biology and medicine to derive new insights, as described in [111] for the broad class of deep learning methods. Supervised learning mostly focuses on the classification of diseases and disease types, patient categorization, and drug interaction prediction. Unsupervised learning has been applied to drug discovery. The authors point out that in addition to the derivation of new findings, an explanation of these is of great importance. Furthermore, the need in deep learning for large training datasets poses a limit to its current applicability beyond imaging (through data augmentation) and so-called 'omics' studies. An overview of deep learning approaches in systems biology is given in [112]. The authors describe how one can design NNs that encode the extensive, existing network- and systems-level knowledge that is generated by combing diverse data types. It is said that such designs inform the model on aspects of the hierarchical interactions in the biological systems that are important for making accurate predictions but are not available in the input data. Reference [113] discusses the difference between explainability and causality for medical applications, and the necessity of a person to be involved. For the successful application of ML for drug design, [114] identifies five ''grand challenges'': obtaining appropriate datasets, generating new hypotheses, optimizing in a multi-objective manner, reducing cycle times, and changing the research culture and mindset. These underlying themes should be valid for many scientific endeavours.

Reference [37] gives an overview of ML research in Earth system science. The authors conclude that, while the general cycle of exploration, hypotheses generation and testing remains the same, modern data-driven science and ML can extract patterns in observational data to challenge complex theories and Earth system models, and thereby strongly complement and enrich geoscientific research. Moreover, [115] observes that a close collaboration with domain experts in the geoscientific area and ML researchers is necessary to solve novel and relevant tasks. They state that developing interpretable and transparent methods is one of the major goals to understand patterns and structures in the data and to turn it into scientific value.

IV. DISCUSSION

In this work, we reviewed the concept of explainable machine learning and discerned between transparency, interpretabil-

TABLE 2. Collection of all references regarding transparency (🟦: at most model transparent, without color: design transparent, 🟪: design + algorithmically transparent), interpretability, and integration of domain knowledge.

		integration of domain knowledge					
		-	design	explaining outcome	explaining outcome + design	explaining outcome + design + post-hoc check	explaining model + explaining outcome + design
interpretability	-	🟦 [39], [40], [41], [42], [43], [39], [44], [45], [46], [47], [48]	[49], [50], [51]	-	[53], [54], [55], [56], [57], [58], 🟪 [59], [60], [61]	[62], [63], [64]	-
	in-out	-	-	🟦 [21], [66], [39], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77]	[78], [79]	-	-
	model	-	-	-	[80], [81], [82], [83], [84], [85], [86], [87], [88], [89], [90], [91], [92]	[93], [94]	[95], [96], [97], [98], [99], [100], [101], [102], [103], [104], 🟪 [105], [106], [107]

ity, and explainability. We also discussed the possibility of influencing model design choices and the step of interpreting algorithmic outputs by domain knowledge and *a posteriori* consistency checks. We presented a more fine-grained characterization of different stages of explainability, which we briefly elaborated on by means of several recent exemplary works in the field of machine learning in the natural sciences; see Table 2 for a summary.

While machine learning is employed in uncountable scientific projects and publications nowadays, the vast majority is not concerned with aspects of interpretability or explainability. We argue that the latter is crucial for extracting truly novel scientific results and ideas from employing ML methods. Therefore, we hope that this survey provides new ideas and methodologies to scientists looking for means to explain their algorithmic results or to extract relevant insights on the corresponding study object.

Finally, we note that as an additional component in the scientific data analysis workflow of the future, we expect causal inference [116], [117] to play a role. Having said this, we believe that causal inference will require even more basic research than what is still needed for the uptake of explainable machine learning in the natural sciences.

ACKNOWLEDGMENT

Part of the work was performed during the long-term program on "Science at Extreme Scales: Where Big Data Meets Large-Scale Computing" held at the Institute for Pure and Applied Mathematics, University of California Los Angeles, USA. The authors would like to thank financial support during the program. They would also like to thank the participants of the long term program for fruitful discussions, in particular K. Dow, L. Gao, P. Grandinetti, P. Hähnel, M. Haghighatlari, and R. Jäkel. Additionally, they cordially thank L. von Rueden and C. Robinson. *(Ribana Roscher and Jochen Garcke contributed equally to this work.)*

REFERENCES

[1] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *ITU J., ICT Discoveries-Impact Artif. Intell. (AI) Commun. Netw. Services*, vol. 1, no. 1, pp. 39–48, 2018.

[2] L. von Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, J. Pfrommer, A. Pick, R. Ramamurthy, M. Walczak, J. Garcke, C. Bauckhage, and J. Schuecker, "Informed machine learning - a taxonomy and survey of integrating knowledge into learning systems," 2019, *arXiv:1903.12394*. [Online]. Available: http://arxiv.org/abs/1903.12394

[3] D. S. Weld and G. Bansal, "The challenge of crafting intelligible intelligence," *Commun. ACM*, vol. 62, no. 6, pp. 70–79, May 2019.

[4] R. Frigg and J. Nguyen, "Scientific representation," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. Stanford, CA, USA: Stanford Univ., 2018.

[5] S. L. Brunton and J. N. Kutz, *Data-Driven Science and Engineering*. Cambridge, U.K.: Cambridge Univ. Press, 2019.

[6] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*. [Online]. Available: http://arxiv.org/abs/1702.08608

[7] S. Ehrhardt, A. Monszpart, N. J. Mitra, and A. Vedaldi, "Learning a physical long-term predictor," 2017, *arXiv:1703.00247*. [Online]. Available: http://arxiv.org/abs/1703.00247

[8] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, Aug. 2018.

[9] Z. C. Lipton, "The mythos of model interpretability," *Commun. ACM*, vol. 61, no. 10, pp. 36–43, Sep. 2018.

[10] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digit. Signal Process.*, vol. 73, pp. 1–15, Feb. 2018.

[11] W. James Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Interpretable machine learning: Definitions, methods, and applications," 2019, *arXiv:1901.04592*. [Online]. Available: http://arxiv.org/abs/1901.04592

[12] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *Ann. Stat.*, vol. 36, no. 3, pp. 1171–1220, 2008.

[13] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[15] A. S. Charles, "Interpreting deep learning: The machine learning rorschach test?" 2018, *arXiv:1806.00148*. [Online]. Available: http://arxiv.org/abs/1806.00148

[16] C. Casert, T. Vieijra, J. Nys, and J. Ryckebusch, "Interpretable machine learning for inferring the phase boundaries in a nonequilibrium system," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 99, no. 2, Feb. 2019, Art. no. 023304.

[17] R. Andrews, J. Diederich, and A. B. Tickle, "Survey and critique of techniques for extracting rules from trained artificial neural networks," *Knowl.-Based Syst.*, vol. 8, no. 6, pp. 373–389, Dec. 1995.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?" in *Proc. KDD*, 2016, pp. 1135–1144.

[19] A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto, *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. Hoboken, NJ, USA: Wiley, 2004.

[20] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[21] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau, "Visual analytics in deep learning: An interrogative survey for the next frontiers," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 8, pp. 2674–2693, Aug. 2019.

[22] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev, "The building blocks of interpretability," *Distill*, vol. 3, no. 3, p. e10, Mar. 2018.

[23] J. Macdonald, S. Wäldchen, S. Hauch, and G. Kutyniok, "A rate-distortion framework for explaining neural network decisions," 2019, *arXiv:1905.11092*. [Online]. Available: http://arxiv.org/abs/1905.11092

[24] J. A. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction* (Information Science and Statistics). New York, NY, USA: Springer, 2007.

[25] A. Cichocki, R. Zdunek, A. H. Phan, and S. I. Amari, *Nonnegative Matrix Tensor Factorization*. Hoboken, NJ, USA: Wiley, 2009.

[26] M. Mørup, "Applications of tensor (multiway array) factorizations and decompositions in data mining," *WIREs Data Mining Knowl. Discovery*, vol. 1, no. 1, pp. 24–40, Jan. 2011.

[27] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever hans predictors and assessing what machines really learn," *Nature Commun.*, vol. 10, no. 1, p. 1096, Mar. 2019.

[28] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in AI," in *Proc. Conf. Fairness, Accountability, Transparency*, 2019, pp. 279–288.

[29] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, Feb. 2019.

[30] J. A. Overton, "Explain in scientific discourse," *Synthese*, vol. 190, no. 8, pp. 1383–1405, 2013.

[31] Q. Liao and T. Poggio, "Object-oriented deep learning," Center Brains, Minds Mach. (CBMM), Cambridge, MA, USA, Tech. Rep., 2017.

[32] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a 'right to explanation,'" *AI Mag.*, vol. 38, no. 3, pp. 50–57, 2017.

[33] R. F. Baumeister and L. S. Newman, "Self-regulation of cognitive inference and decision processes," *Personality Social Psychol. Bull.*, vol. 20, no. 1, pp. 3–19, Jul. 2016.

[34] D. J. Koehler, "Explanation, imagination, and confidence in judgment," *Psychol. Bull.*, vol. 110, no. 3, pp. 499–519, 1991.

[35] A. Karpatne, G. Atluri, J. H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar, "Theory-guided data science: A new paradigm for scientific discovery from data," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 10, pp. 2318–2331, Oct. 2017.

[36] M. Raissi, P. Perdikaris, and G. Em Karniadakis, "Physics informed deep learning (part II): Data-driven discovery of nonlinear partial differential equations," 2017, *arXiv:1711.10566*. [Online]. Available: http://arxiv.org/abs/1711.10566

[37] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, no. 7743, pp. 195–204, Feb. 2019.

[38] M. McCloskey, "Intuitive physics," *Sci. Amer.*, vol. 248, no. 4, pp. 122–130, Apr. 1983.

[39] A. Lerer, S. Gross, and R. Fergus, "Learning physical intuition of block towers by example," in *Proc. 33rd Int. Conf. Mach. Learn.*, M. F. Balcan and K. Q. Weinberger, Eds., 2016, pp. 430–438.

[40] W. Li, S. Azimi, A. Leonardis, and M. Fritz, "To fall or not to fall: A visual approach to physical stability prediction," 2016, *arXiv:1604.00066*. [Online]. Available: http://arxiv.org/abs/1604.00066

[41] A. Forster, J. Behley, J. Behmann, and R. Roscher, "Hyperspectral plant disease forecasting using generative adversarial networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 1793–1796.

[42] J. C. Mauro, A. Tandia, K. D. Vargheese, Y. Z. Mauro, and M. M. Smedskjaer, "Accelerating the design of functional glasses through modeling," *Chem. Mater.*, vol. 28, no. 12, pp. 4267–4277, Jun. 2016.

[43] M. Raissi and G. E. Karniadakis, "Hidden physics models: Machine learning of nonlinear partial differential equations," *J. Comput. Phys.*, vol. 357, pp. 125–141, Mar. 2018.

[44] M. Ansdell, Y. Ioannou, H. P. Osborn, M. Sasdelli, J. C. Smith, D. Caldwell, J. M. Jenkins, C. Räissi, and D. Angerhausen, "Scientific domain knowledge improves exoplanet transit classification with deep learning," *Astrophys. J.*, vol. 869, no. 1, p. L7, Dec. 2018.

[45] Y. Zhu, Y. Zhao, and S.-C. Zhu, "Understanding tools: Task-oriented object modeling, learning and recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2855–2864.

[46] K. Breen, S. C. James, and J. D. White, "Deep learning model integration of remotely sensed and SWAT-simulated regional soil moisture," *AGU Fall Meeting Abstr.*, Dec. 2018.

[47] D. H. Brookes and J. Listgarten, "Design by adaptive sampling," 2018, *arXiv:1810.03714*. [Online]. Available: http://arxiv.org/abs/1810.03714

[48] D. H. Brookes, H. Park, and J. Listgarten, "Conditioning by adaptive sampling for robust design," 2019, *arXiv:1901.10060*. [Online]. Available: http://arxiv.org/abs/1901.10060

[49] J. Tompson, K. Schlachter, P. Sprechmann, and K. Perlin, "Accelerating Eulerian fluid simulation with convolutional networks," in *Proc. ICML*, D. Precup and Y. W. Teh, Eds., 2017, pp. 3424–3433.

[50] L. Ladický, S. Jeong, B. Solenthaler, M. Pollefeys, and M. Gross, "Data-driven fluid simulations using regression forests," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–9, Oct. 2015.

[51] R. Hagensieker, R. Roscher, J. Rosentreter, B. Jakimow, and B. Waske, "Tropical land use land cover mapping in Pará (Brazil) using discriminative Markov random fields and multi-temporal TerraSAR-X data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 63, pp. 244–256, Dec. 2017.

[52] M. B. Chang, T. Ullman, A. Torralba, and J. B. Tenenbaum, "A compositional object-based approach to learning physical dynamics," in *Proc. ICLR*, 2017, pp. 1–7.

[53] R. Stewart and S. Ermon, "Label-free supervision of neural networks with physics and domain knowledge," in *Proc. AAAI*, vol. 1, 2017, pp. 1–7.

[54] J. Wu, J. Lim, H. Zhang, J. Tenenbaum, and W. Freeman, "Physics 101: Learning physical object properties from unlabeled videos," in *Proc. Brit. Mach. Vis. Conf.*, 2016. p. 7.

[55] A. Monszpart, N. Thuerey, and N. J. Mitra, "SMASH: Physics-guided reconstruction of collisions from videos," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–14, Nov. 2016.

[56] R. Mottaghi, H. Bagherinezhad, M. Rastegari, and A. Farhadi, "Newtonian image understanding: Unfolding the dynamics of objects in static images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3521–3529.

[57] M. Eigel, R. Schneider, P. Trunschke, and S. Wolf, "Variational Monte Carlo-bridging concepts of machine learning and high dimensional partial differential equations," 2018, *arXiv:1810.01348*. [Online]. Available: http://arxiv.org/abs/1810.01348

[58] C. Hoerig, J. Ghaboussi, and M. F. Insana, "An information-based machine learning approach to elasticity imaging," *Biomech. Model. Mechanobiol.*, vol. 16, no. 3, pp. 805–822, Nov. 2016.

[59] E. O. Pyzer-Knapp, G. N. Simm, and A. Aspuru Guzik, "A Bayesian approach to calibrating high-throughput virtual screening results and application to organic photovoltaic materials," *Mater. Horizons*, vol. 3, no. 3, pp. 226–233, 2016.

[60] S. A. Lopez, B. Sanchez-Lengeling, J. de Goes Soares, and A. Aspuru-Guzik, "Design principles and top non-fullerene acceptor candidates for organic photovoltaics," *Joule*, vol. 1, no. 4, pp. 857–870, Dec. 2017.

[61] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Machine learning of linear differential equations using Gaussian processes," *J. Comput. Phys.*, vol. 348, pp. 683–693, Nov. 2017.

[62] J. Ling, A. Kurzawski, and J. Templeton, "Reynolds averaged turbulence modelling using deep neural networks with embedded invariance," *J. Fluid Mech.*, vol. 807, pp. 155–166, Oct. 2016.

[63] J. X. Wang, J. Huang, L. Duan, and H. Xiao, "Prediction of Reynolds stresses in high-mach-number turbulent boundary layers using physics-informed machine learning," *Theor. Comput. Fluid Dyn.*, vol. 33, no. 1, pp. 1–19, 2019.

[64] J. Ling, R. Jones, and J. Templeton, "Machine learning strategies for systems with invariance properties," *J. Comput. Phys.*, vol. 318, pp. 22–35, Aug. 2016.

[65] S. B. Ginsburg, G. Lee, S. Ali, and A. Madabhushi, "Feature importance in nonlinear embeddings (FINE): Applications in digital pathology," *IEEE Trans. Med. Imag.*, vol. 35, no. 1, pp. 76–88, Jan. 2016.

[66] S. Ghosal, D. Blystone, A. K. Singh, B. Ganapathysubramanian, A. Singh, and S. Sarkar, "An explainable deep machine vision framework for plant stress phenotyping," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 18, pp. 4613–4618, Apr. 2018.

[67] O. Groth, F. B. Fuchs, I. Posner, and A. Vedaldi, "Shapestacks: Learning vision-based physical intuition for generalised object stacking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 702–717.

[68] A. L. Pinho *et al.*, "Individual brain charting, a high-resolution fMRI dataset for cognitive mapping," *Sci. Data*, vol. 5, no. 1, p. 180, Jun. 2018.

[69] G. Varoquaux, Y. Schwartz, R. A. Poldrack, B. Gauthier, D. Bzdok, J.-B. Poline, and B. Thirion, "Atlases of cognition with large-scale human brain mapping," *PLOS Comput. Biol.*, vol. 14, no. 11, Nov. 2018, Art. no. e1006565.

[70] R. Abbasi-Asl, Y. Chen, A. Bloniarz, M. Oliver, B. D. Willmore, J. L. Gallant, and B. Yu, "The deeptune framework for modeling and characterizing neurons in visual cortex area v4," *bioRxiv*, Jan. 2018, Art. no. 465534, doi: 10.1101/465534.

[71] J. G. Colonna, J. Gama, and E. F. Nakamura, "A comparison of hierarchical multi-output recognition approaches for anuran classification," *Mach. Learn.*, vol. 107, no. 11, pp. 1651–1671, Jul. 2018.

[72] N. Tomita, B. Abdollahi, J. Wei, B. Ren, A. Suriawinata, and S. Hassanpour, "Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides," *JAMA Netw. Open*, vol. 2, no. 11, Nov. 2019, Art. no. e1914645.

[73] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, Apr. 2019.

[74] L. Deming, S. Targ, N. Sauder, D. Almeida, and C. Jimmie Ye, "Genetic architect: Discovering genomic structure with learned neural architectures," 2016, *arXiv:1605.07156*. [Online]. Available: http://arxiv.org/abs/1605.07156

[75] R. Singh, J. Lanchantin, A. Sekhon, and Y. Qi, "Attend and predict: Understanding gene regulation by selective attention on chromatin," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6785–6795.

[76] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3504–3512.

[77] K. Chen, L. Yao, X. Wang, D. Zhang, T. Gu, Z. Yu, and Z. Yang, "Interpretable parallel recurrent neural networks with convolutional attentions for multi-modality activity modeling," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–8.

[78] B. Kailkhura, B. Gallagher, S. Kim, A. Hiszpanski, and T. Y.-J. Han, "Reliable and explainable machine-learning methods for accelerated material discovery," *NPJ Comput. Mater.*, vol. 5, no. 1, pp. 1–9, Nov. 2019.

[79] L. Rieger, C. Singh, W. James Murdoch, and B. Yu, "Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge," 2019, *arXiv:1909.13584*. [Online]. Available: http://arxiv.org/abs/1909.13584

[80] O. Yair, R. Talmon, R. R. Coifman, and I. G. Kevrekidis, "Reconstruction of normal forms by learning informed observation geometries from data," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 38, pp. E7865–E7874, Aug. 2017.

[81] C. J. Dsilva, R. Talmon, R. R. Coifman, and I. G. Kevrekidis, "Parsimonious representation of nonlinear dynamical systems through manifold learning: A chemotaxis case study," *Appl. Comput. Harmon. Anal.*, vol. 44, no. 3, pp. 759–773, May 2018.

[82] A. Holiday, M. Kooshkbaghi, J. M. Bello-Rivas, C. W. Gear, A. Zagaris, and I. G. Kevrekidis, "Manifold learning for parameter reduction," *J. Comput. Phys.*, vol. 392, pp. 419–431, Sep. 2019.

[83] W. Hu, R. R. P. Singh, and R. T. Scalettar, "Discovering phases, phase transitions, and crossovers through unsupervised machine learning: A critical examination," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 95, no. 6, Jun. 2017, Art. no. 062122.

[84] L. Wang, "Discovering phase transitions with unsupervised learning," *Phys. Rev. B, Condens. Matter*, vol. 94, no. 19, Nov. 2016, Art. no. 195105.

[85] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 15, pp. 3932–3937, Mar. 2016.

[86] N. M. Mangan, S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Inferring biological networks by sparse identification of nonlinear dynamics," *IEEE Trans. Mol., Biol. Multi-Scale Commun.*, vol. 2, no. 1, pp. 52–63, Jun. 2016.

[87] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Data-driven discovery of partial differential equations," *Sci. Adv.*, vol. 3, no. 4, Apr. 2017, Art. no. e1602614.

[88] H. Schaeffer, R. Caflisch, C. D. Hauck, and S. Osher, "Sparse dynamics for partial differential equations," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 17, pp. 6634–6639, 2013.

[89] G. Tran and R. Ward, "Exact recovery of chaotic systems from highly corrupted data," *Multiscale Model. Simul.*, vol. 15, no. 3, pp. 1108–1129, Jan. 2017.

[90] H. Dai, Y. Tian, B. Dai, S. Skiena, and L. Song, "Syntax-directed variational autoencoder for structured data," in *Proc. ICLR*, 2018, pp. 1–17.

[91] L. Li, M. Fan, R. Singh, and P. Riley, "Neural-guided symbolic regression with asymptotic constraints," 2019, *arXiv:1901.07714*. [Online]. Available: http://arxiv.org/abs/1901.07714

[92] M. Meila, S. Koelle, and H. Zhang, "A regression approach for explaining manifold embedding coordinates," 2018, *arXiv:1811.11891*. [Online]. Available: http://arxiv.org/abs/1811.11891

[93] L. M. Ghiringhelli, J. Vybiral, E. Ahmetcik, R. Ouyang, S. V. Levchenko, C. Draxl, and M. Scheffler, "Learning physical descriptors for materials science by compressed sensing," *New J. Phys.*, vol. 19, no. 2, Feb. 2017, Art. no. 023017.

[94] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, and L. M. Ghiringhelli, "SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates," *Phys. Rev. Mater.*, vol. 2, no. 8, pp. 1–11, Aug. 2018.

[95] A. Adiga, C. J. Kuhlman, M. V. Marathe, H. S. Mortveit, S. S. Ravi, and A. Vullikanti, "Graphical dynamical systems and their applications to bio-social systems," *Int. J. Adv. Eng. Sci. Appl. Math.*, vol. 11, no. 2, pp. 153–171, Dec. 2018.

[96] H. Tanaka, A. Nayebi, N. Maheswaranathan, L. McIntosh, S. Baccus, and S. Ganguli, "From deep learning to mechanistic understanding in neuroscience: The structure of retinal prediction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8535–8545.

[97] B. Lusch, J. N. Kutz, and S. L. Brunton, "Deep learning for universal linear embeddings of nonlinear dynamics," *Nature Commun.*, vol. 9, no. 1, p. 4950, Nov. 2018.

[98] J. Ma, M. K. Yu, S. Fong, K. Ono, E. Sage, B. Demchak, R. Sharan, and T. Ideker, "Using deep learning to model the hierarchical structure and function of a cell," *Nature Methods*, vol. 15, no. 4, pp. 290–298, Mar. 2018.

[99] B. C. Daniels, W. S. Ryu, and I. Nemenman, "Automated, predictive, and interpretable inference of caenorhabditis elegans escape dynamics," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 15, pp. 7226–7231, Mar. 2019.

[100] R. Iten, T. Metger, H. Wilming, L. del Rio, and R. Renner, "Discovering physical concepts with neural networks," 2018, *arXiv:1807.10300*. [Online]. Available: http://arxiv.org/abs/1807.10300

[101] T. Ye, X. Wang, J. Davidson, and A. Gupta, "Interpretable intuitive physics model," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 87–102.

[102] Y. Yan, J. Zhu, M. Duda, E. Solarz, C. Sripada, and D. Koutra, "GroupINN: Grouping-based interpretable neural network for classification of limited, noisy brain data," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2019, pp. 772–782.

[103] M. Tsang, D. Cheng, and Y. Liu, "Detecting statistical interactions from neural network weights," in *Proc. ICLR*, 2018, pp. 1–21.

[104] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks," *Nature Commun.*, vol. 8, no. 1, Jan. 2017, Art. no. 13890.

[105] G. Camps-Valls, L. Martino, D. Svendsen, M. Campos-Taberner, J. Muñoz-Marí, V. Laparra, D. Luengo, and J. García-Haro, "Physics-aware Gaussian processes in remote sensing," *Appl. Soft Comput.*, vol. 68, Mar. 2018.

[106] J. Greitemann, K. Liu, and L. Pollet, "Probing hidden spin order with interpretable machine learning," *Phys. Rev. B, Condens. Matter*, vol. 99, no. 6, pp. 1–6, Feb. 2019.

[107] K. Liu, J. Greitemann, and L. Pollet, "Learning multiple order parameters with interpretable machines," *Phys. Rev. B, Condens. Matter*, vol. 99, no. 10, pp. 1–15, Mar. 2019.

[108] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," *Nature*, vol. 559, pp. 547–555, Jul. 2018.

[109] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, "Machine learning and the physical sciences," *Rev. Mod. Phys.*, vol. 91, Dec. 2019, Art. no. 045002.

[110] A. Wagner, A. Regev, and N. Yosef, "Revealing the vectors of cellular identity with single-cell genomics," *Nature Biotechnol.*, vol. 34, no. 11, pp. 1145–1160, Nov. 2016.

[111] T. Ching *et al.*, "Opportunities and obstacles for deep learning in biology and medicine," *J. Roy. Soc. Interface*, vol. 15, no. 141, 2018, Art. no. 20170387.

[112] V. H. Gazestani and N. E. Lewis, "From genotype to phenotype: Augmenting deep learning with networks and systems biology," *Current Opinion Syst. Biol.*, vol. 15, pp. 68–73, Jun. 2019.

[113] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *WIREs Data Mining Knowl. Discovery*, vol. 9, no. 4, Apr. 2019.

[114] P. Schneider, W. P. Walters, A. T. Plowright, N. Sieroka, J. Listgarten, R. A. Goodnow, J. Fisher, J. M. Jansen, J. S. Duca, T. S. Rush, M. Zentgraf, J. E. Hill, E. Krutoholow, M. Kohler, J. Blaney, K. Funatsu, C. Luebkemann, and G. Schneider, "Rethinking drug design in the artificial intelligence era," *Nature Rev. Drug Discovery*, Dec. 2019, doi: 10.1038/s41573-019-0050-3.

[115] A. Karpatne, I. Ebert-Uphoff, S. Ravela, H. Ali Babaie, and V. Kumar, "Machine learning for the geosciences: Challenges and opportunities," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 8, pp. 1544–1554, Aug. 2019.

[116] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[117] B. Schölkopf, "Causality for machine learning," 2019, *arXiv:1911.10500*. [Online]. Available: http://arxiv.org/abs/1911.10500

RIBANA ROSCHER (Member, IEEE) received the Dipl.Ing. and Ph.D. degrees in geodesy from the University of Bonn, Bonn, Germany, in 2008 and 2012, respectively.

Until 2015, she was a Postdoctoral Researcher with the University of Bonn, the Julius-Kuehn Institute, Siebeldingen, Germany, Freie Universitaet Berlin, Berlin, Germany, and the Humboldt Innovation, Berlin. In 2015, she was a Visiting Researcher with the Fields Institute, Toronto, ON, Canada. She is currently an Assistant Professor of remote sensing with the Institute of Geodesy and Geoinformation, University of Bonn, and an Interims Professor of semantic technologies with the Institute of Computer Science, University of Osnabrueck, Osnabrueck, Germany. Her researches include pattern recognition and machine learning for remote sensing applications.

Prof. Roscher is a Convenor Team Member of Gesellschaft für Geodaesie, Geoinformation und Landmanagement, working group member of International Society for Photogrammetry and Remote Sensing, and a Technical Committee Member of the International Association of Pattern Recognition. She was a recipient of the IEEE International Geoscience and Remote Sensing Symposium Prize Paper Award (with C. Roemer, B. Waske, and L. Pluemer), in 2016. She is a Reviewer of the IEEE Transactions on Geoscience and Remote Sensing, the IEEE Transactions on Neural Networks and Learning Systems, and the IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

BASTIAN BOHN received the Diploma and Ph.D. degrees in mathematics from the University of Bonn, Germany, in 2010 and 2016, respectively.

He is currently a Postdoctoral Researcher with the Institute for Numerical Simulation, University of Bonn, Germany. His focus lies on the numerical analysis of machine learning algorithms employed in high-dimensional engineering and natural sciences tasks. He was an Invited Research Fellow with the Institute for Pure and Applied Mathematics at UCLA, CA, USA, for the long program Science at Extreme Scales: Where Big Data Meets Large-Scale Computing, in Fall 2018. His research interests include machine learning, the mathematics of data science, numerical algorithms in high-dimensions, and approximation theory.

MARCO F. DUARTE (Senior Member, IEEE) received the B.Sc. degree (Hons.) in computer engineering and the M.Sc. degree in electrical engineering from the University of Wisconsin–Madison, Madison, WI, USA, in 2002 and 2004, respectively, and the Ph.D. degree in electrical and computer engineering from Rice University, Houston, TX, USA, in 2009.

He was an NSF/IPAM Mathematical Sciences Postdoctoral Research Fellow of the Program of Applied and Computational Mathematics, Princeton University, Princeton, NJ, USA, from 2009 to 2010, and the Department of Computer Science, Duke University, Durham, NC, USA, from 2010 to 2011. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, University of Massachusetts Amherst, MA, USA. His research interests include machine learning, compressed sensing, sensor networks, and computational imaging.

Dr. Duarte is also a member of Tau Beta Pi. He received the Presidential Fellowship and the Texas Instruments Distinguished Fellowship, in 2004, and the Hershel M. Rich Invention Award, in 2007, from Rice University. He was a recipient of the IEEE Signal Processing Society Overview Paper Award (with Y. C. Eldar), in 2017. He is also an Associate Editor of the IEEE Signal Processing Letters.

JOCHEN GARCKE received the Diploma degree in mathematics and the Ph.D. degree in mathematics from the Universität Bonn, in 1999 and 2004, respectively.

He was a Postdoctoral Fellow at the Australian National University, from 2004 to 2006, and a Postdoctoral Researcher, from 2006 to 2008 and a Junior Research Group Leader, from 2008 to 2011, at Technische Universität Berlin. Since 2011, he has been Professor of numerics with the Universität Bonn and a Department Head with Fraunhofer SCAI, Sankt Augustin. His research interests include machine learning, scientific computing, reinforcement learning, and high-dimensional approximation.

Prof. Garcke is also a member of DMV, GAMM, and SIAM. He is also a Reviewer of the IEEE Transactions on Industrial Informatics, the IEEE Transactions on Neural Networks, and the IEEE Transactions on Pattern Analysis and Machine Intelligence.