Explainable Machine Learning for Scientific Insights and Discoveries

Ribana Roscher[*1,2], Bastian Bohn[3],
Marco F. Duarte[4], and Jochen Garcke[*3,5]

[1]Institute of Geodesy and Geoinformation, University of Bonn, Germany
[2]Institute of Computer Science, University of Osnabrueck, Germany
[3]Institute for Numerical Simulation, University of Bonn, Germany
[4]Dept. of Electrical and Computer Engineering, University of Massachusetts Amherst, USA
[5]Fraunhofer Center for Machine Learning and Fraunhofer SCAI, Sankt Augustin, Germany

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 which 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 incorporating machine learning, and in particular to the way that explainable machine learning is used in their respective application areas.

1 Introduction

Machine learning methods, especially with the rise of deep neural networks (DNNs), are nowadays used widely in commercial applications. This success has also led to a considerable uptake of machine learning (ML) in many scientific areas. Usually these models are trained with regard to high accuracy, but recently there is also a high demand for understanding the way a specific model operates and the underlying reasons for the produced decisions. One motivation behind this is that scientists increasingly adopt ML for optimizing and producing scientific outcomes, where explainability is a prerequisite to ensure the scientific value of the outcome. In this context, research directions such

[*]R. Roscher and J. Garcke contributed equally to this work

as explainable artificial intelligence (AI) [Samek et al., 2018], informed ML [von Rueden et al., 2019], or intelligible intelligence [Weld and Bansal, 2018] 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 and objectives, contextual standards of accuracy, and intended audiences. Regarding conditions on an adequate scientific representation we refer to the philosophy of science [Frigg and Nguyen, 2018].

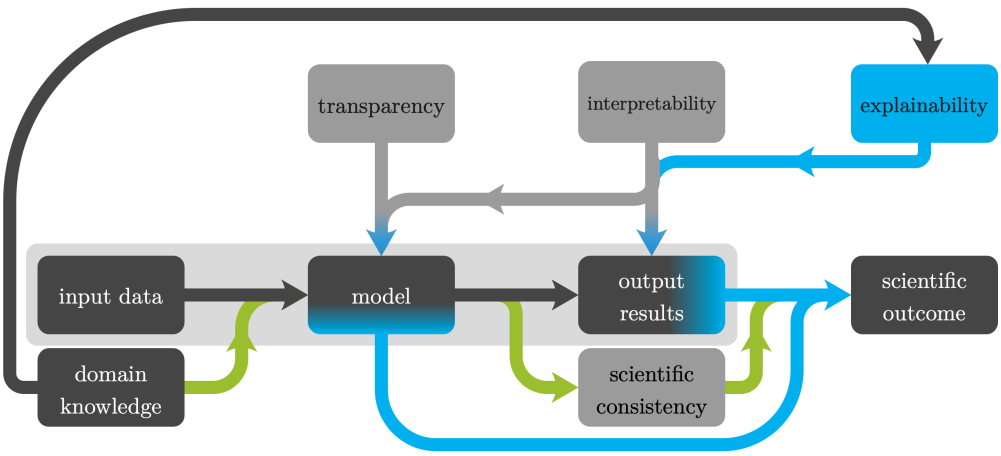

Figure 1: Major ML-based chains from which scientific outcomes can be derived: The commonly used, basic ML chain (light gray box) learns a black box model from given input data and provides an output. Given the black box model and input-output relations, a scientific outcome can be derived by explaining 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).

This article provides a survey of recent ML approaches which are meant to derive scientific outcomes, where we specifically focus on the natural sciences. Given the scientific outcomes, novel insights can be derived helping for a deeper understanding, or scientific discoveries can be revealed which 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.

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 Brunton and Kutz [2019],

> 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 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 which have been used in the natural sciences (see Fig. 1). A central role play the three elements *transparency*, *interpretability*, and *explainability*, which will be defined and discussed in detail in this survey. The core 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. A further essential part is *domain knowledge*, which is necessary to achieve explainability, but can also be used to foster *scientific consistency* of the model and the result. 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 information gained from domain insights such as laws of nature and chemical, biological, or physical models [von Rueden et al., 2019]. 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.

We will give diverse examples from the natural sciences for approaches which can be related to these topics. Our goal is to foster a better understanding and a clearer overview of ML algorithms applied to data from the natural sciences.

In the broader context, other properties that can be relevant when considering explainability of ML algorithms are safety/trust, accountability, reproducability, transferability, robustness and multi-objective trade-off or mismatched objectives, see e.g. [Doshi-Velez and Kim, 2017, Lipton, 2018]. 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 a 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” [Goodman and Flaxman, 2017]. 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 easier realized, if the model is interpretable [Lipton, 2018].

The paper is structured as follows. In Sec. 2 we discuss 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. In Sec. 3, we highlight several applications from the natural sciences which use these concepts to gain new scientific insights.

2 Terminology

It can be observed that in the literature about explainable ML several descriptive terms are used with diverse meanings, see e.g. Doshi-Velez and Kim [2017], Gilpin et al. [2018], Guidotti et al. [2018], Lipton [2018], Montavon et al. [2018], Murdoch et al. [2019]. Nonetheless, distinct ideas can be identified. For the purpose of this work, we differentiate 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.

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: This includes the overall model structure, the individual model components, the learning algorithm, and how the specific solution is obtained by the algorithm. We propose to differentiate between *model transparency*, *design transparency*, and *algorithmic transparency*. Generally, a fully transparent ML method in all aspects is rather doubtful; usually there will be different degrees of transparency.

As an example, consider kernel-based ML approaches [Hofmann et al., 2008, Rasmussen and Williams, 2006]. The obtained model is accessible and transparent, and it is given as a sum of kernel functions. The individual design component is the chosen kernel. Choosing between a linear or non-linear kernel is typically a transparent design decision. However, the commonly used Gaussian kernel based on Euclidean distances can be a non-transparent design decision. In other words, it may not be clear why a given non-linear kernel is taken. Here, 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 [Rasmussen and Williams, 2006]. Thereby, design transparency goes 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 which can be reproducibly obtained, resulting in algorithmic transparency [Hofmann et al., 2008, Rasmussen and Williams, 2006]. 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 DNNs [Goodfellow et al., 2016]. 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 DNN, that are chosen based on domain knowledge can be considered as design transparent. Nonetheless, the layer parameters — be it their numbers, size, or involved nonlinearities — are often chosen in an ad-hoc or heuristic fashion and not motivated by 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, and others, has more a 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 with Lipton [2018], 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 DNNs appears to mimic a type of Rorschach test" according to Charles [2018].

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.

Interpretability We consider interpretability as about making sense of the obtained ML model. Generally, to interpret means "to explain the meaning of" or "present in understandable terms"[1]; see also Doshi-Velez and Kim [2017], Gilpin et al. [2018], Guidotti et al. [2018]. We consider explaining as a separate aspect, on top of an interpretation, and focus here on the second aspect. Therefore, the aim of interpretability is to present

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

some of the properties of an ML model in understandable terms to a human. Ideally, one could answer the question from Casert et al. [2019]: "Can we understand on what the ML algorithm bases its decision?" Somewhat formally, Montavon et al. [2018] state:

> 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 [Gilpin et al., 2018, Guidotti et al., 2018]. Longstanding approaches involve decision trees or rule extraction [Andrews et al., 1995] 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 which are relevant for a prediction, either globally or locally. For example, Ribeiro et al. [2016] 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. Sensitivity analysis can be used to inspect how a model output (locally) depends upon the different input parameters [Saltelli et al., 2004]. Such an extraction of information from the input and the output of a learned model is also called *post hoc interpretability* [Lipton, 2018] or *reverse engineering* [Guidotti et al., 2018]. Further details, types of interpretation, and specific realization can be found in recent surveys [Adadi and Berrada, 2018, Gilpin et al., 2018, Guidotti et al., 2018].

Visual approaches such as saliency masks or heatmaps show relevant patterns in the input based on feature importance or sensitivity analysis to explain model decisions, in particular employed for deep learning approaches for image classification [Hohman et al., 2018, Montavon et al., 2018, Olah et al., 2018]. Note that recently a formal and rigorous notion for interpreting neural networks was introduced, where a set of input features is deemed relevant for a classification decision if the expected classifier score remains nearly constant when randomising the remaining features [MacDonald et al., 2019]. 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 shows on the one hand the difficulty of algorithmically determining interpretations, and on the other hand 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 an example, by an interpretation of the obtained representation by linear or non-linear dimensionality reduction [Lee and Verleysen, 2007, Cichocki et al., 2009], or by inspecting the components of a low-rank tensor decomposition [Mørup, 2011].

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 DNNs exploits the known model layout [Montavon et al., 2018].

While the methods for interpretation allow the inspection of a single datum, Lapuschkin et al. [2019] observe that it becomes quickly 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.

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 [Mittelstadt et al., 2019].

While in philosophy and psychology explanations are in 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 [Miller, 2019], 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 Montavon et al. [2018]:

> 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 Guidotti et al. [2018], "[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."

We believe on the other hand, 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, but the important observation for an understanding of the overall predictive behavior could be that in a ranking with respect to the interpretation, different variable lists are determined for each data as being of relevance. 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 Adadi and Berrada [2018], 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 to differentiate between *algorithmic explanations* and *scientific explanations*. 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. For scientific explanations, Overton [2013] 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 whether there is a unifying general account of scientific explanation remains an open question.

One should also observe that explanations can be used to manipulate. For illustration, Baumeister and Newman [1994] distinguish 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, for example 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 [Koehler, 1991].

Domain knowledge As outlined, domain knowledge is an essential part of explainability, but also for treating small data scenarios or for performance reasons. A taxonomy for the explicit integration of knowledge into the ML pipeline, so called *informed ML*, is proposed in von Rueden et al. [2019]. 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 Karpatne et al. [2017], who use the term *theory-guided data science*, or physics-informed learning by [Raissi et al., 2017a]. For the purpose of this article, we follow von Rueden et al. [2019] and aim to 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 expression or differential equations, as relations between instances and/or classes in form of rules or constraints. It can be represented in the form of ontologies, by symmetries, or using similarity measures. Knowledge can be transformed by numerical simulations of models or through human interaction.

As ingredients of an ML approach one considers 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, while using numerical simulations to generate (additional) training data is a modern phenomena.

Integrating knowledge into the hypothesis space can be achieved by choosing the structure of the model. For example, by defining a specific architecture of a neural network or by choosing a structure of probability distributions which observes existing or non-existing links between variables. An example for the training phase is modifying the loss function according to additional knowledge, e.g., by adding a consistency term. Finally, the obtained model can be put in relation to existing knowledge, e.g., by checking known constraints for the predictions. This aspect we call scientific consistency and deem it especially important to obtain scientific outcomes.

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 way 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 von Rueden et al. [2019], 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. Reichstein et al. [2019] identify scientific consistency besides interpretability as one of the five major challenges we need to tackle to successfully adopt deep learning approaches in the geosciences. Karpatne et al. [2017] 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 following revised objective of model performance [...]: Performance $\propto$ Accuracy + Simplicity + Consistency.

They discuss several ways to restrict the solution space to physically consistent solutions, e.g., by (1) design of the model family such as specific network architectures, (2) guidance of a learning algorithm using, e.g., specific initializations, constraints, or (loss) regularizations, (3) refinement of the model output, e.g., using closed-form equations or model simulations, (4) 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 solutions is not a requirement to achieve valuable scientific outcomes. Neglecting this restriction, however, means that a consistent solution cannot be guaranteed, even if an optimal result has been achieved from a mathematical point of view.

3 Scientific Outcomes From Machine Learning

In this section, we will review several examples that use ML and strive for different levels of transparency, interpretability, and explainability to produce scientific outcomes. We will focus on examples which utilize an extensive amount of scientific domain knowledge from the natural sciences.

We define two general categories: The first one is the derivation of scientific outcomes by explaining output results. Many works address the derivation of scientific outcomes by learning an ML model and generalizing from known input-output relations to new input-output pairs. 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 why this specific outcome was arrived from an algorithmic point of view (algorithmic explanation). Other approaches attempt to scientifically explain the output in terms of the specific corresponding input. 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. This states the lowest degree of explainability with no necessity of a transparent or interpretable model.

The other approach is to derive scientific outcomes by explaining models. Here, interpretation tools are used to project processes in the model into a space which is interpretable, which can then be explained utilizing domain knowledge. Both scientific and algorithmic explanations are used to derive a scientific outcome. This means that even if the scientific outcome is more specifically defined by domain experts, transparency and interpretability of the models are not a prerequisite for these approaches. Note that the following collection of research works is a non-exhaustive selection from recent literature, where we aim to cover a broad range of usages of ML with a variety of scientific outcomes.

3.1 Scientific Outcomes by Explaining Output Results

3.1.1 Prediction of Intuitive Scientific Outcomes

The works described in this subsection have been developed in the physical domain, where generally two kind of outcomes are derived. The first is the derivation of intuitive physics: everyday-observed rules of nature which help us to predict the outcome of events even with a relatively untrained human perception, e.g., whether a tower will collapse [McCloskey, 1983]. The other one is concerned with the estimation of specific physical parameters from which static properties or object behavior can be derived. Chang et al. [2017] denote these respective approaches as bottom-up, where observations are directly mapped to an estimate of some object behavior or the physical outcome of a scene, and as top-down, where parameters are inferred to explain a scene. In both cases, only the scientific explanation is aspired.

A task often considered is the prediction of whether a certain construction collapses in an image or a video. Lerer et al. [2016] and Li et al. [2016] use video simulations to learn intuitive physics, for example about the stability of wooden block towers. Lerer et al. [2016] use ResNet-34 [He et al., 2016] and Googlenet [Szegedy et al., 2015] to predict the fall of towers of wooden blocks, as well as DeepMask [Pinheiro et al., 2015] and a custom network called PhysNet to predict the trajectory of the wooden blocks in case the tower is collapsing. The first task is formulated as a binary classification task and the second task is formulated as a semantic segmentation, where each wooden block is defined as one class. In both tasks, PhysNet outperforms human subjects on synthetic data and achieves comparable results on real data. The construction of PhysNet is made design transparent in the sense that the network layers are chosen such that the arrangement of the wooden blocks is determined via a local and translation-invariant image upscaling before their inherent physics are analyzed on a coarse scale. From experiments with occluded images, the authors were able to gain interpretability for the binary classification task by conducting a heatmap analysis. Similar experiments with more complex scenes or differently shaped objects were conducted by Li et al. [2016] and Groth et al. [2018] using various popular convolutional neural networks (CNNs). While the generic CNN choices there do not seem to be transparent per se, Groth et al. [2018] provide a first step towards an interpretable and physics-aware model by training their algorithm to actively counterbalance instabilities by placing new objects on top of unstable stacks. Tompson et al. [2017] and Jeong et al. [2015] use similar 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 Tompson et al. [2017] 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 Jeong et al. [2015] to predict a fluid particle's velocity can be viewed as a transparent choice per se due to its simple nature.

3.1.2 Prediction of Scientific Parameters and Properties

Although the approaches just described set up scientific outcome prediction as supervised learning problems, 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. The methods presented so far do not learn a model that is able to capture and derive the physical properties and dynamics of objects and their environment, as well as their interactions. Therefore, the model cannot inherently explain why a specific outcome was obtained from a scientific viewpoint. Several classification and regression frameworks have been formulated to tackle this challenge.

Stewart and Ermon [2017], for example, detect and track objects in videos in an unsupervised way. For this, they use a regression CNN and introduce terms which 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, their design process becomes interpretable and explainability is gained due to comparisons to the underlying physical process. However, the model and algorithms are not completely transparent since standard CNNs with an ADAM minimizer are employed. Wu et al. [2016] introduce Physics101, a dataset which contains over 17000 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 [LeCun et al., 1998] 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. Monszpart et al. [2016] introduce SMASH, which extracts physical collision parameters from videos of colliding objects, such as pre- and post collision velocities, to use them as input 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 and interpretable. Individual outcomes become explainable due to the direct relation of the computations to the underlying physical laws.

Also other disciplines use ML to help guide new scientific insights and discoveries. Regression, in particular, has often been leveraged to explain phenomena. Mauro et al. [2016] present an approach for the design of new functional glasses which comprises the prediction of characteristics relevant for manufacturing as well as end-use properties of glass. Among others, they utilize neural networks to estimate the liquidus temperatures for various silicate compositions comprising up to 8 different components. For this, they learn from several hundred composites with known output properties and apply the model to novel, unknown composites. 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. While transparency or interpretability is lacking in the mere process of training a neural network based on a least-squares loss to learn corresponding liquidus temperatures, the authors also introduce more physics-driven models for different quantities of interest, which also need to be estimated to aid the design process of functional glasses in the end.

For organic photovoltaics material, a related approach utilizing quantum chemistry calculations and ML techniques to calibrate theoretical results to experimental data was presented in [Pyzer-Knapp et al., 2016, Lopez et al., 2017]. The authors consider already performed existing experiments as current knowledge, which is embedded within a probabilistic non-parametric mapping. In particular, Gaussian processes were used to learn the deviation of properties calculated by computational models from the experimental analogues. By employing the chemcial Tanimoto similarity measure and building a prior based on experimental observations, model transparency and interpretability is attained. Furthermore, since the prediction results involve 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 [Pyzer-Knapp et al., 2016]. In Lopez et al. [2017], 838 high-performing candidate molecules have been identified within the explored molecular space, due to the now possible efficient screening of over 51,000 molecules.

In Ling et al. [2016b], a deep learning approach for Reynolds-averaged Navier–Stokes (RANS) turbulence modelling was presented. Here, domain-knowledge led to the constructions of a network architecture that embedded invariance using a higher-order multiplicative layer. This was shown to have significantly more accurate predictions compared to a generic, less interpretable, neural network 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 Wang et al. [2019], which uses a systematic approach with basis tensor invariants proposed by Ling et al. [2016a]. Additionally, a metric of prediction confidence and a nonlinear dimensionality reduction technique are employed to provide a priori assessment of the prediction confidence.

In Raissi et al. [2017b], 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 Gaussian process priors that are tailored to the corresponding and known type of differential operators. Therefore, the combination of rather generic ML models with domain knowledge in 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. A related information-based ML approach to solve an inverse problem in biomechanical applications was presented in Hoerig et al. [2017]. Here, in mechanical property imaging of soft biological media under quasi-static loads, elasticity imaging parameters are computed from estimated stresses and strains. Physics-aware GP models in remote sensing were studied in Camps-Valls et al. [2018]. In particular, a latent force model that incorporates ordinary differential equations was 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.

A tensor-based approach to ML for uncertainty quantification problems can be found in Eigel et al. [2018]. Here, the solutions to parametric convection-diffusion partial differential equations 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. Raissi [2018] proposes a nonlinear regression approach employing DNNs to learn closed form representations of partial differential equations 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, including 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 PDE and the dynamics themselves, the rather general neural network model does not allow to draw direct scientific conclusions on the structure of the underlying process.

Mottaghi et al. [2016] introduce Newtonian neural networks 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 which best describes the current dynamics in the image. Newtonian neural networks are two parallel CNNs, where one encodes the images and 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 (partly) allows for explaining the classification results of a single input image. Zhu et al. [2015] introduces a framework which calculates physical concepts from color-depth videos that explains 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. Their transparent SVM-like learning procedure allows to work with rather small sample sets.

3.1.3 Interpretation Tools for Scientific Outcomes

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, direct approaches are undertaken to present this information via 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 from the learned model during the post-hoc analysis.

Kailkhura et al. [2019] discusses explainable ML for scientific discoveries in material sciences. They identify challenges when using ML for material science applications such as the reliability-explainability trade-off. They point out that many works see interpretability and explainability as the inverse of complexity, leading to an increase in accuracy and reliability when reducing the complexity. In the worst case, this may lead to misunderstanding or incorrect interpretations. In their work, they propose an ensemble of simple models to predict material properties along with a novel evaluation metric focusing on trust by quantifying generalization performance. Moreover, their pipeline contains a rationale generator which 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 which are analyzed and explained by an expert, as well as global interpretations by estimating feature importance for material sub-classes.

In many domains an increased interest in using automatic approaches for estimating feature importances can be observed. While handcrafted and manually selected features are typically easier to understand, automatically determined features can reveal previously unknown scientific attributes and structures. Ginsburg et al. [2016], for example, proposes FINE (feature importance in nonlinear embeddings) for the analysis of cancer patterns in ER+ 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, non-linear 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. Hohman et al. [2018] give a survey of visual analytics in deep learning research, where such visualizations 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. Ghosal et al. [2018] use interpretation tools for image-based plant stress phenotyping. They 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 sum 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 reveal a high level of agreement between the automatic approach and human ratings. The goal of their approach is the analysis of the performance of their model, the provision of visual cues which are human-interpretable to support the prediction of the system, and a provision of important cues for the identification of plant stress. Abbasi-Asl et al. [2018] introduce DeepTune, a stability-driven visualization framework for CNNs, for applications in neuroscience. DeepTune consists of a battery 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 visual cortex are V4. The combination of the feature extraction and regression modules allows for accurate prediction of V4 neuron responses to additional visual stimuli. Representative visual stimuli for each neuron can then be generated from the trained modules via gradient optimization. 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 [Pinho et al., 2018, Varoquaux et al., 2018]. The biomarker design promotes spatially compact pixel selections, producing biomarkers for disease prediction that are focused on regions of the brain; these are then considered by expert physicians. As the analysis is based on high-dimensional linear regression approaches, transparency of the ML model is assured.

Interpretability methods have also been used for applications which utilize time-series data, often by way of highlighting features of the sequence data. For example, Deming et al. [2016] applied attention modules in neural networks trained on genomic sequences for the identification of important sequence motifs by visualizing the attention mask weights. Here, they propose a genetic architect that finds a suitable network architecture by iteratively searching over various neural network building blocks. In particular, they state that the choice of the neural network 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.

Additionally, Singh et al. [2017] use attention modules for genomics in their AttentiveChrome neural network. The network contains a hierarchy of attention modules to gain insights about where and what the network has focused and, thus, gaining interpretability of the results. Also Choi et al. [2016] 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 neural networks for multi-modal sensor-based activity recognition have been used by Chen et al. [2018]. Depending on the activity, their approach provides the most contributing body parts, modals, and sensors for the network's decision.

In certain cases, models can be interpreted by using them as a driver for an underlying design problem. For example, Brookes and Listgarten [2018] have proposed a data-centric approach for scientific design based on the combination of a generative model for the data being considered, e.g., 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. With that, one generates new synthetic data samples that optimize the value of the quantity or property by leveraging an adaptive sampling technique over the generative model.

Notwithstanding, classical tools such as confusion matrices are also used as interpretation tools on the way to scientific outcomes. In a bioacoustic application for the recognition of anurans using acoustic sensors, Colonna et al. [2018] use 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.

3.2 Scientific Outcomes by Explaining Models

So far, the presented approaches either treat the model as a black box or use it only indirectly by applying interpretation tools to better explain the output. Liao and Poggio [2017] propose a concept called 'object-oriented deep learning' with the goal to convert a DNN to a symbolic description to gain interpretability and explainability. They state that generally in DNNs 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 states a promising direction towards a higher explainability of models. The reviewed approaches in this section 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. We will further focus on recent ML approaches, which focus on the interpretation and explanation of single components of the model or the whole model structure.

In contrast to most of the works in Sec. 3.1, which rely on prior knowledge about relevant parameters, some other works derive characteristics of settings without any assumptions about the underlying scientific process. For example, Ehrhardt et al. [2017] derive physical parameters without assuming prior knowledge about the physical processes and without modelling the underlying physical models in order to make predictions in simple physical scenarios over time. Here, physically explainable parameters are not only derived as outcome, but also integrated in a recurrent end-to-end long-term prediction network. Therefore, a simulation software and the explicit modelling of the underlying physical laws is not necessary.

Another broad framework [Yair et al., 2017, Dsilva et al., 2018, Holidaya et al., 2019] 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.

Iten et al. [2018] introduces SciNet, a modified variational autoencoder which 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 and 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, Ye et al. [2018] construct the bottleneck layer in their neural network 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. Daniels et al. [2019] use their ML algorithm 'Sir Isaac' [Daniels and Nemenman, 2015] to infer a dynamical model of biological time series data to understand and predict dynamics of worm behavior. They model a system of differential equations, where the number of hidden variables is determined automatically from the system, and the meaning of them can be explained by an expert.

Feature selection schemes using embedded methods have been recently explored to establish or refine models in physical processes [Rudy et al., 2017] and material sciences [Ghiringhelli et al., 2017, Ouyang et al., 2018]. Using a sparsity-promoting penalty, they propose groups of variables that may explain a property of interest and promote the simplest model, that is, the model involving the fewest variables possible while achieving a target accuracy. Meila et al. [2018] propose a sparsity-enforcing technique to recover domain-specific meaning for the embedding coordinates obtained from unsupervised nonlinear dimensionality reduction approaches. 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. The application of sparsity has also proved fruitful in the broader class of problems leveraging partial differential equation and dynamical system models [Tran and Ward, 2017, Mangan et al., 2016, Schaeffer et al., 2013].

Complex ML methods such as DNNs, 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 plasma physics modeling for inversion, properties such as positivity and smoothness can be promoted by a modified deep learning network [Matos et al., 2018]. Similarly, the properties of contaminant dispersion in soil can be successfully modeled by a long-short-term memory network [Breen et al., 2018]. In [Adiga et al., 2018], 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. Ma et al. [2018] introduces visible neural networks, which encode the hierarchical structure of a gene ontology tree into an 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 was provided during model training, previously undocumented learned subsystem states could be confirmed by molecular measurements.

Understanding structures such as groups, relations and interactions is one of the main goals to achieve scientific outcomes. However, it states a core challenge and so far only a limited amount of works have been conducted in this area. Yan et al. [2019], for example, introduce a grouping layer in an interpretable neural network called GroupINN to identify subgroups of neurons in an end-to-end model. In their work, they build a network for the analysis of timeseries of functional magnetic resonance images of the brain, which are represented as functional graphs, with the goal to reveal 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.

Tsang et al. [2018] introduces neural interaction detection, a feedforward neural network for detecting statistical interactions. By examining the learned weight matrices of the hidden units, their framework was able to analyze feature interactions in the Higgs-Boson dataset [Adam-Bourdarios et al., 2014]. Specifically, they analyze feature interactions in simulated particle environments which originate from the decay of a Higgs Boson. Deep tensor networks are used by Schütt et al. [2017] 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 neural network 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.

Lusch et al. [2018] construct a DNN for computing Koopman eigenfunctions from data. Motivated by domain knowledge, they employ an auxiliary network to parameterize the continuous frequency. Thereby, a compact autoencoder model is obtained, which additionally is interpretable. For the example of the nonlinear pendulum, the two eigenfunctions, learned with a deep neural network, 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 which turned out to be consistent with recent theoretical derivations beforehand unknown to the authors. 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 to determine different aspects of cellular organization and function. Here, there is an emphasis on unsupervised learning approaches to cluster cells from single-cell profiles, and thereby to systematically detect beforehand unknown cellular subtypes, for which then defining markers are investigated in a second step, see [Wagner et al., 2016] for a review on key questions, progress, and open challenges in this application field.

3.3 Related Surveys about Machine Learning in the Natural Sciences

Butler et al. [2018] give 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.

Several ML approaches have been used in biology and medicine to derive new insights, as described in Ching et al. [2018] 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 Gazestani and Lewis [2019]. They describe how one can design DNNs 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.

Reichstein et al. [2019] give an overview of ML research in Earth system science. They 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. Also Karpatne et al. [2018] point out 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.

Acknowledgements

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. We are grateful for their financial support during the program. We cordially thank the participants of the long term program for fruitful discussions, in particular Keiko Dow, Longfei Gao, Pietro Grandinetti, Philipp Haehnel, Mojtaba Haghighatlari, and René Jäkel.

References

Reza Abbasi-Asl, Yuansi Chen, Adam Bloniarz, Michael Oliver, Ben DB Willmore, Jack L Gallant, and Bin Yu. The deeptune framework for modeling and characterizing neurons in visual cortex area v4. *bioRxiv*, page 465534, 2018.

Amina Adadi and Mohammed Berrada. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018. doi: 10.1109/ACCESS.2018.2870052.

Claire Adam-Bourdarios, Glen Cowan, Cecile Germain, Isabelle Guyon, Balazs Kegl, and David Rousseau. Learning to discover: the higgs boson machine learning challenge. *URL http://higgsml. lal. in2p3. fr/documentation*, page 9, 2014.

Abhijin Adiga, Chris J. Kuhlman, Madhav V. Marathe, Henning S. Mortveit, S. S. Ravi, and Anil Vullikanti. Graphical dynamical systems and their applications to bio-social systems. *International Journal of Advances in Engineering Sciences and Applied Mathematics*, Dec 2018. doi: 10.1007/s12572-018-0237-6.

Robert Andrews, Joachim Diederich, and Alan B. Tickle. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*, 8(6):373–389, 1995. doi: 10.1016/0950-7051(96)81920-4.

Roy F. Baumeister and Leonard S. Newman. Self-Regulation of Cognitive Inference and Decision Processes. *Personality and Social Psychology Bulletin*, 20(1):3–19, feb 1994. doi: 10.1177/0146167294201001.

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 Abstracts*, Dec. 2018.

David H. Brookes and Jennifer Listgarten. Design by adaptive sampling. arXiv preprints arXiv:1810.03714, 2018.

Steven L. Brunton and J. Nathan Kutz. *Data-Driven Science and Engineering*. Cambridge University Press, 2019. doi: 10.1017/9781108380690.

Keith T. Butler, Daniel W. Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 2018. doi: 10.1038/s41586-018-0337-2.

Gustau Camps-Valls, Luca Martino, Daniel Svendsen, Manuel Campos-Taberner, Jordi Muñoz, Valero Laparra, David Luengo, and Javier García-Haro. Physics-aware gaussian processes in remote sensing. *Applied Soft Computing*, 68, 03 2018. doi: 10.1016/j.asoc.2018.03.021.

C. Casert, T. Vieijra, J. Nys, and J. Ryckebusch. Interpretable machine learning for inferring the phase boundaries in a nonequilibrium system. *Physical Review E*, 99(2): 023304, feb 2019. doi: 10.1103/PhysRevE.99.023304.

Michael B Chang, Tomer Ullman, Antonio Torralba, and Joshua B Tenenbaum. A compositional object-based approach to learning physical dynamics. In *ICLR*, 2017.

Adam S Charles. Interpreting deep learning: The machine learning rorschach test? *SIAM News July/August*, 2018. arXiv preprint arXiv:1806.00148.

Kaixuan Chen, Lina Yao, Xianzhi Wang, Dalin Zhang, Tao Gu, Zhiwen Yu, and Zheng Yang. Interpretable parallel recurrent neural networks with convolutional attentions for multi-modality activity modeling. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.

Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141):20170387, 2018.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, pages 3504–3512, 2016.

Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shun Ichi Amari. *Nonnegative matrix and tensor factorization*. John Wiley & Sons, 2009.

Juan G. Colonna, João Gama, and Eduardo F. Nakamura. A comparison of hierarchical multi-output recognition approaches for anuran classification. *Machine Learning*, 107 (11):1651–1671, 2018. doi: 10.1007/s10994-018-5739-8.

Bryan C Daniels and Ilya Nemenman. Automated adaptive inference of phenomenological dynamical models. *Nature communications*, 6:8133, 2015.

Bryan C Daniels, William S Ryu, and Ilya Nemenman. Automated, predictive, and interpretable inference of caenorhabditis elegans escape dynamics. *Proceedings of the National Academy of Sciences*, page 201816531, 2019.

Laura Deming, Sasha Targ, Nate Sauder, Diogo Almeida, and Chun Jimmie Ye. Genetic architect: Discovering genomic structure with learned neural architectures. *arXiv preprint arXiv:1605.07156*, 2016.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608, 2017.

Carmeline J. Dsilva, Ronen Talmon, Ronald R. Coifman, and Ioannis G. Kevrekidis. Parsimonious representation of nonlinear dynamical systems through manifold learning: A chemotaxis case study. *Applied and Computational Harmonic Analysis*, 44(3): 759 – 773, May 2018.

S. Ehrhardt, A. Monszpart, N. J. Mitra, and A. Vedaldi. Learning a physical long-term predictor. arXiv e-prints arXiv:1703.00247, 2017.

Martin Eigel, Reinhold Schneider, Philipp Trunschke, and Sebastian Wolf. Variational Monte Carlo - bridging concepts of machine learning and high dimensional partial differential equations. arXiv e-prints arXiv:1810.01348, 2018.

Roman Frigg and James Nguyen. Scientific representation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2018 edition, 2018.

Vahid H. Gazestani and Nathan E. Lewis. From Genotype to Phenotype: Augmenting Deep Learning with Networks and Systems Biology. *Current Opinion in Systems Biology*, apr 2019. doi: 10.1016/j.coisb.2019.04.001.

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 Journal of Physics*, 19(2), Feb. 2017.

Sambuddha Ghosal, David Blystone, Asheesh K Singh, Baskar Ganapathysubramanian, Arti Singh, and Soumik Sarkar. An explainable deep machine vision framework for plant stress phenotyping. *Proceedings of the National Academy of Sciences*, 115(18): 4613–4618, 2018.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining Explanations: An Overview of Interpretability of Machine Learning. arXiv preprints arXiv:1806.00069, may 2018.

Shoshana B Ginsburg, George Lee, Sahirzeeshan Ali, and Anant Madabhushi. Feature importance in nonlinear embeddings (fine): applications in digital pathology. *IEEE transactions on medical imaging*, 35(1):76–88, 2016.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, 2017.

Oliver Groth, Fabian B. Fuchs, Ingmar Posner, and Andrea Vedaldi. Shapestacks: Learning vision-based physical intuition for generalised object stacking. In *The European Conference on Computer Vision (ECCV)*, pages 702–717, 2018.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5):1–42, aug 2018. doi: 10.1145/3236009.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Cameron Hoerig, Jamshid Ghaboussi, and Michael F. Insana. An information-based machine learning approach to elasticity imaging. *Biomechanics and Modeling in Mechanobiology*, 16(3):805–822, 2017. doi: 10.1007/s10237-016-0854-6.

Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220, 2008. doi: 10.1214/009053607000000677.

Fred Matthew Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):1–20, 2018. doi: 10.1109/TVCG.2018.2843369.

A. Holidaya, M. Kooshkbaghib, J. M. Bello-Rivas, C. W. Geara, A. Zagarisc, and I. G. Kevrekidis. Manifold learning for parameter reduction. *Journal of Computational Physics*, 392:419 – 431, 2019.

Raban Iten, Tony Metger, Henrik Wilming, Lidia del Rio, and Renato Renner. Discovering physical concepts with neural networks. *arXiv preprint arXiv:1807.10300*, 2018.

SoHyeon Jeong, Barbara Solenthaler, Marc Pollefeys, Markus Gross, et al. Data-driven fluid simulations using regression forests. *ACM Transactions on Graphics (TOG)*, 34 (6):199, 2015.

Bhavya Kailkhura, Brian Gallagher, Sookyung Kim, Anna Hiszpanski, and T Han. Reliable and explainable machine learning methods for accelerated material discovery. *arXiv preprint arXiv:1901.02717*, 2019.

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 Transactions on Knowledge and Data Engineering*, 29(10):2318–2331, 2017.

Anuj Karpatne, Imme Ebert-Uphoff, Sai Ravela, Hassan Ali Babaie, and Vipin Kumar. Machine Learning for the Geosciences: Challenges and Opportunities. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–12, 2018. doi: 10.1109/TKDE.2018.2861006.

Derek J Koehler. Explanation, imagination, and confidence in judgment. *Psychological Bulletin*, 110(3):499–519, 1991. doi: 10.1037/0033-2909.110.3.499.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, 2019. doi: 10.1038/s41467-019-08987-4.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

John A. Lee and Michel Verleysen. *Nonlinear dimensionality reduction*. Information Science and Statistics. Springer, New York, 2007. doi: 10.1007/978-0-387-39351-3.

Adam Lerer, Sam Gross, and Rob Fergus. Learning physical intuition of block towers by example. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 430–438, New York, New York, USA, 2016. PMLR.

Wenbin Li, Seyedmajid Azimi, Aleš Leonardis, and Mario Fritz. To fall or not to fall: A visual approach to physical stability prediction. arXiv preprint arXiv:1604.00066, 2016.

Qianli Liao and Tomaso Poggio. Object-oriented deep learning. Technical report, Center for Brains, Minds and Machines (CBMM), 2017.

Julia Ling, Reese Jones, and Jeremy Templeton. Machine learning strategies for systems with invariance properties. *Journal of Computational Physics*, 318:22–35, 2016a. doi: 10.1016/j.jcp.2016.05.003.

Julia Ling, Andrew Kurzawski, and Jeremy Templeton. Reynolds averaged turbulence modelling using deep neural networks with embedded invariance. *Journal of Fluid Mechanics*, 807:155–166, 2016b. doi: 10.1017/jfm.2016.615.

Z. Lipton. The mythos of model interpretability. *Queue*, 16(3):1–28, 2018.

Steven A. Lopez, Benjamin Sanchez-Lengeling, Julio de Goes Soares, and Alán Aspuru-Guzik. Design Principles and Top Non-Fullerene Acceptor Candidates for Organic Photovoltaics. *Joule*, 1(4):857–870, 2017. doi: 10.1016/j.joule.2017.10.006.

Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9(1):4950, dec 2018. doi: 10.1038/s41467-018-07210-0.

Jianzhu Ma, Michael Ku Yu, Samson Fong, Keiichiro Ono, Eric Sage, Barry Demchak, Roded Sharan, and Trey Ideker. Using deep learning to model the hierarchical structure and function of a cell. *Nature Methods*, 15(4):290–298, apr 2018. doi: 10.1038/nmeth.4627.

Jan MacDonald, Stephan Wäldchen, Sascha Hauch, and Gitta Kutyniok. A rate-distortion framework for explaining neural network decisions. arXiv preprints arXiv:1905.11092, 2019.

N. M. Mangan, S. L. Brunton, J. L. Proctor, and J. N. Kutz. Inferring biological networks by sparse identification of nonlinear dynamics. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 2(1):52–63, June 2016.

F. Matos, F. Hendrich, F. Jenko, and T. Odstrcil. Deep learning for plasma diagnostics. In *Jahrestagung der DPG und DPG-Frühjahrstagung der Sektion AMOP*, Erlangen, Germany, Apr. 2018.

John C. Mauro, Adama Tandia, K. Deenamma Vargheese, Yihong Z. Mauro, and Morten M. Smedskjaer. Accelerating the Design of Functional Glasses through Modeling. *Chemistry of Materials*, 28(12):4267–4277, 2016. doi: 10.1021/acs.chemmater.6b01054.

Michael McCloskey. Intuitive physics. *Scientific american*, 248(4):122–131, 1983.

Marina Meila, Samson Koelle, and Hanyu Zhang. A regression approach for explaining manifold embedding coordinates. arXiv preprints arXiv:1811.11891, 2018.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. doi: 10.1016/j.artint.2018.07.007.

Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining Explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, pages 279–288, New York, New York, USA, 2019. ACM Press. ISBN 9781450361255. doi: 10.1145/3287560.3287574.

Aron Monszpart, Nils Thuerey, and Niloy J. Mitra. SMASH: Physics-guided reconstruction of collisions from videos. *ACM Trans. Graph.*, 35(6):199:1–199:14, 2016.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018. doi: 10.1016/j.dsp.2017.10.011.

Morten Mørup. Applications of tensor (multiway array) factorizations and decompositions in data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):24–40, 2011. doi: 10.1002/widm.1.

Roozbeh Mottaghi, Hessam Bagherinezhad, Mohammad Rastegari, and Ali Farhadi. Newtonian scene understanding: Unfolding the dynamics of objects in static images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3521–3529, 2016.

W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Interpretable machine learning: definitions, methods, and applications. arXiv preprint arXiv:1901.04592, 2019.

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 2018. doi: 10.23915/distill.00010.

Runhai Ouyang, Stefano Curtarolo, Emre Ahmetcik, Matthias Scheffler, and Luca M. Ghiringhelli. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Physical Review Materials*, 2(8):1–11, 2018. doi: 10.1103/PhysRevMaterials.2.083802.

James A. Overton. "Explain" in scientific discourse. *Synthese*, 190(8):1383–1405, 2013. doi: 10.1007/s11229-012-0109-8.

P. O. Pinheiro, R. Collobert, and P. Dollár. Learning to segment object candidates. In *Advances in Neural Information Processing Systems*, pages 1990–1998, 2015.

Ana Luísa Pinho, Alexis Amadon, Torsten Ruest, Murielle Fabre, Elvis Dohmatob, Isabelle Denghien, Chantal Ginisty, Séverine Becuwe-Desmidt, Séverine Roger, Laurence Laurier, Véronique Joly-Testault, Gaëlle Médiouni-Cloarec, Christine Doublé, Bernadette Martins, Philippe Pinel, Evelyn Eger, G. Varoquaux, Christophe Pallier, Stanislas Dehaene, Lucie Hertz-Pannier, and Bertrand Thirion. Individual brain charting, a high-resolution fmri dataset for cognitive mapping. *Scientific Data*, 5:180105 EP –, 06 2018.

Edward O. Pyzer-Knapp, Gregor N. Simm, and Alán Aspuru Guzik. A Bayesian approach to calibrating high-throughput virtual screening results and application to organic photovoltaic materials. *Materials Horizons*, 3(3):226–233, 2016. doi: 10.1039/c5mh00282f.

M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics informed deep learning (Part II): data-driven discovery of nonlinear partial differential equations. arXiv preprint arXiv:1711.10566, 2017a.

Maziar Raissi. Deep hidden physics models: Deep learning of nonlinear partial differential equations. *The Journal of Machine Learning Research*, 19(1):932–955, 2018.

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Machine learning of linear differential equations using Gaussian processes. *Journal of Computational Physics*, 348:683–693, 2017b. doi: 10.1016/j.jcp.2017.07.050.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, et al. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195, 2019.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?". In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pages 1135–1144, New York, New York, USA, 2016. ACM Press. doi: 10.1145/2939672.2939778.

Samuel H. Rudy, Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3(4), 2017. doi: 10.1126/sciadv.1602614.

Andrea Saltelli, Stefano Tarantola, Francesca Campolongo, and Marco Ratto. *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. Wiley, 2004. ISBN 978-0-470-87093-8.

Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *ITU Journal: ICT Discoveries - Special Issue 1 - The Impact of Artificial Intelligence (AI) on Communication Networks and Services*, 1(1):39–48, 2018.

Hayden Schaeffer, Russel Caflisch, Cory D. Hauck, and Stanley Osher. Sparse dynamics for partial differential equations. *Proceedings of the National Academy of Sciences*, 110(17):6634–6639, 2013.

Kristof T Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature communications*, 8:13890, 2017.

Ritambhara Singh, Jack Lanchantin, Arshdeep Sekhon, and Yanjun Qi. Attend and predict: Understanding gene regulation by selective attention on chromatin. In *Advances in neural information processing systems*, pages 6785–6795, 2017.

R. Stewart and S. Ermon. Label-free supervision of neural networks with physics and domain knowledge. In *AAAI*, volume 1, pages 1–7, 2017.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

Jonathan Tompson, Kristofer Schlachter, Pablo Sprechmann, and Ken Perlin. Accelerating Eulerian fluid simulation with convolutional networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3424–3433, 2017.

G. Tran and R. Ward. Exact recovery of chaotic systems from highly corrupted data. *Multiscale Modeling & Simulation*, 15(3):1108–1129, 2017.

Michael Tsang, Dehua Cheng, and Yan Liu. Detecting statistical interactions from neural network weights. In *ICLR*, 2018.

Gaël Varoquaux, Yannick Schwartz, Russell A. Poldrack, Baptiste Gauthier, Danilo Bzdok, Jean-Baptiste Poline, and Bertrand Thirion. Atlases of cognition with large-scale human brain mapping. *PLOS Computational Biology*, 14(11):1–18, 11 2018. doi: 10.1371/journal.pcbi.1006565.

Laura von Rueden, Sebastian Mayer, Jochen Garcke, Christian Bauckhage, and Jannis Schuecker. Informed machine learning – towards a taxonomy of explicit integration of knowledge into machine learning. arXiv preprint arXiv:1903.12394, 2019.

Allon Wagner, Aviv Regev, and Nir Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology*, 34(11):1145–1160, 2016. doi: 10.1038/nbt.3711.

Jian Xun Wang, Junji Huang, Lian Duan, and Heng Xiao. Prediction of Reynolds stresses in high-Mach-number turbulent boundary layers using physics-informed machine learning. *Theoretical and Computational Fluid Dynamics*, 33(1):1–19, 2019. doi: 10.1007/s00162-018-0480-2.

D Weld and Gagan Bansal. The challenge of crafting intelligible intelligence. *Communications of ACM*, 2018.

J. Wu, J. J. Lim, H. Zhang, J. B. Tenenbaum, and W. T. Freeman. Physics 101: Learning physical object properties from unlabeled videos. In *BMVC*, volume 2, page 7, 2016.

Or Yair, Ronen Talmon, Ronald R. Coifman, and Ioannis G. Kevrekidis. Reconstruction of normal forms by learning informed observation geometries from data. *Proceedings of the National Academy of Sciences*, 114(38):E7865–E7874, 2017. ISSN 0027-8424. doi: 10.1073/pnas.1620045114.

Yujun Yan, Jiong Zhu, Marlena Duda, Eric Solarz, Chandra Sripada, and Danai Koutra. GroupINN: Grouping-based interpretable neural network for classification of limited, noisy brain data. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2019.

Tian Ye, Xiaolong Wang, James Davidson, and Abhinav Gupta. Interpretable intuitive physics model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 87–102, 2018.

Yixin Zhu, Yibiao Zhao, and Song Chun Zhu. Understanding tools: Task-oriented object modeling, learning and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2855–2864, 2015.