SAND2021-14093R

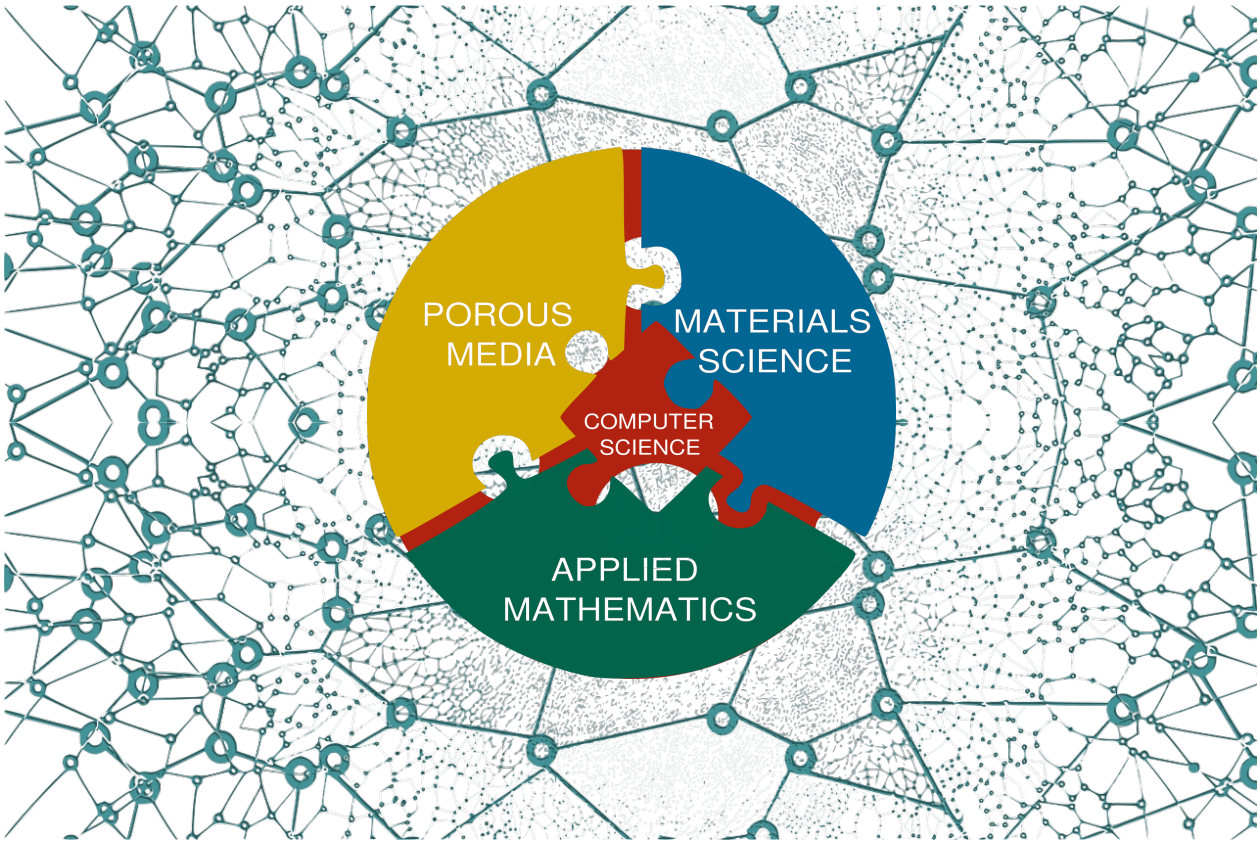
THE UNIVERSITY OF UTAH®

The National Academies of SCIENCES ENGINEERING MEDICINE

Machine Learning For Heterogeneous Porous Materials

AmeriMech Symposium Series

November 5, 2021



Introduction

The "**Workshop on Machine learning in heterogeneous porous materials**" was virtually held on October 4-6, 2020 at the University of Utah. The workshop was part of the AmeriMech Symposium series sponsored by in the National Academies of Sciences, Engineering and Medicine and the U.S. National Committee on Theoretical and Applied Mechanics.

The workshop for the first time brought together senior and early career international experts in the areas of heterogeneous materials, machine learning (ML) and applied mathematics to identify how machine learning can advance materials research.

There is no debate that machine learning has tackled many scientific and engineering problems in the past decade. The explosion of machine learning methods developed by industry (e.g. Google, Facebook, etc.) have been particularly successful in addressing data rich problems such as visualization (e.g. facial recognition) where machine learning techniques are very effective due to their ability to interpolate and fit using big data for training. However, its role in multi-physics, multi-scale problems which are often data sparse and require extrapolation (e.g. prediction, forward modeling) is less clear. For prediction and forward modeling, the underlying governing equations are often critical. Therefore, physics-informed machine learning approaches that combine the underlying equations and physical constraints with data-driven approaches are needed for many scientific problems. These approaches are far less developed by industry and is therefore a key knowledge gap that academia and national labs can fill. In addition, scientific discoveries, energy resiliency, and national security require an in-depth understanding of multi-scale, multi-physics and heterogeneous processes in order to predict and eventually control system behavior.

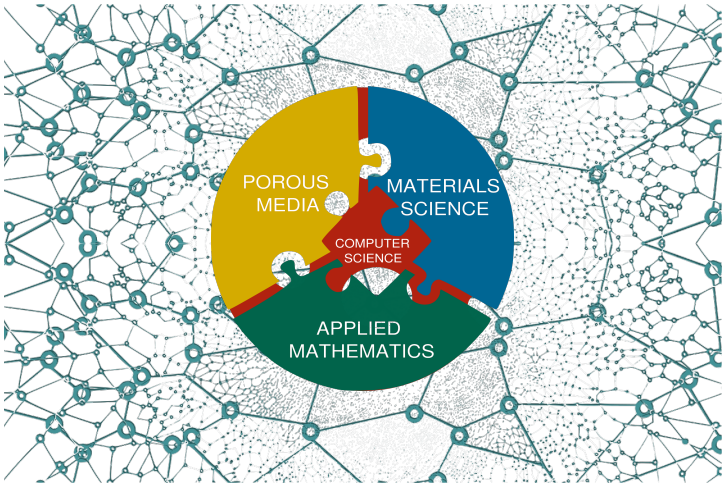


Figure 1: ML unites three scientific communities.

The workshop's goal was to bring three scientific communities of applied mathematics,

porous media, and material sciences together to:

1. discuss the state-of-the-art in each community
2. promote and accelerate multi-disciplinary collaborative research
3. identify challenges and opportunities

The workshop identified four topic areas (**TAs**) that would benefit from machine learning enabled cross-disciplinary research between all three communities. These TAs are:

- **TA 1:** ML in predicting materials properties, and discovery and design of novel materials
- **TA 2:** ML in porous and fractured media and time-dependent phenomena
- **TA 3:** Multi-scale modeling in heterogeneous porous materials via ML
- **TA 4:** Discovery of materials constitutive laws and new governing equations

The workshop was attended by 70 participants in total. It featured 6 keynote lectures, four parallel sessions. The keynote lectures were:

- Combining Graph Theory and Machine Learning to Characterize Fractured Systems by Dr. Gowri Srinivasan from Los Alamos National Laboratory.
- Data-Driven Learning of Nonlocal models: Bridging Scales with Nonlocality by Dr. Marta D'Elia from Sandia National Laboratories.
- Applications of Machine Learning Techniques in Fracture Mechanics by Prof. Huajian Gao from Nanyang Technological University.
- Bioinspired AI towards Modeling, Design and Manufacturing of de novo Materials by Prof. Markus J. Buehler from MIT.
- Generative Design and Additive Manufacturing of Three-Dimensional architected metamaterials by Dr. Grace X. Gu from University of California, Berkeley.
- Learning Solution Operators In Continuum Mechanics by Prof. Andre Stuart from Caltech.

The abstracts and biographical information of the speakers are available in Appendix A.

Each morning and afternoon session started with the keynote speech, followed by break and discussion session. At the end of the day, a briefing session was held where all the chairs reported the highlight of their discussion to all the participants.
During the 7 discussion sessions (three in Day 1, three in Day 2 and one in Day 3), following questions were discussed in order to gain consensus state-of-the-art, challenges and future directions:

1. What is the state-of-the-art in your topic area?
2. What are the existing challenges and opportunities in your topic area?
3. How inclusive is the community associated with this topic area?
4. Where would we like this area to be in ten years?
5. List some of the common science questions in this area that overlaps with other communities (Applied Math, Porous material, Materials Science).
6. How can your community learn from others to overcome some of your existing challenges?
7. What technical advances must be made to overcome the existing challenges?
8. How can we make our communities more inclusive?

The symposium was also focused in promoting JEDI (Justice, Equity, Diversity, and Inclusion). To this end, a diverse group of keynote speakers, chairs and participants from different disciplines and at different stage of their careers were invited to this symposium. Furthermore, there were some discussion within each sessions with JEDI focus.
The following chapters summarize the discussions within each TA and provide a road-map for the future research directions within their area of research.

Pania Newell The University of Utah
George Karniadakis Brown University
Hari Viswanathan Los Alamos National Laboratory

Discovering new governing equations using ML

M. D'Elia[1], A. Howard[2], R. M. Kirby[3], N. Kutz[4], A. Tartakovsky[5], H. Viswanathan[6]

[1] Sandia National Laboratories, CA, USA.
[2] Pacific Northwest National Laboratory, WA, USA.
[3] University of Utah, UT, USA.
[4] University of Washington, WA, USA.
[5] University of Illinois, Urbana-Champaign, IL, USA.
[6] Los Alamos National Laboratory, NM, USA.

Abstract

A hallmark of the scientific process since the time of Newton has been the derivation of mathematical equations meant to capture relationships between *observables*. As the field of mathematical modeling evolved, practitioners specifically emphasized mathematical formulations that were predictive, generalizable, and interpretable. Machine learning's ability to interrogate complex processes is particularly useful for the analysis of highly heterogeneous, anisotropic materials where idealized descriptions often fail. As we move into this new era, we anticipate the need to leverage machine learning to aid scientists in extracting meaningful, but yet sometimes elusive, relationships between observed quantities.

Introduction

The derivation of governing equations for physical systems currently dominates the physical and engineering sciences. These governing equations are used for predictive physics-based models to forecast system behavior. Indeed, this is the dominant paradigm for the modeling and characterization of physical processes, engendering rapid and diverse technological developments in every application area of the sciences. Until the middle of the 20^{th} century, many models of natural and engineered systems relied on linear governing equations that are amenable to analytic solutions. With the invention of the computer and the rise of scientific computing, nonlinear problems could easily be explored through numerical simulation using techniques such as finite elements. Scientific computing allows one to emulate diverse and complex systems that are high-dimensional, multiscale, and potentially stochastic in nature. In modern times, the rapid evolution of sensor technologies and data-acquisition software/hardware, broadly defined, has opened new fields of exploration where governing equations are difficult to generate and/or produce. Biology and neuroscience, for instance, easily come to mind as application areas where first-principals derivations are

difficult to achieve, yet data is now becoming abundant and of exceptional quality. The coarse-grained macroscopic behavior of heterogeneous materials is also often difficult to derive or characterize from known microscopic descriptions. The ability to discover governing equations directly from data is thus of paramount importance in many modern scientific and engineering settings as they provide both interpretability and generalizability.

Here we first summarize the state-of-the-art in the area of machine learning (ML) dedicated to deducing governing equations from data, highlighting the importance of three fundamental features of such formulations (which we will define below): predictivity, generalizability, and interpretability. We then explore the current knowledge gaps and challenges to making this area fully realizable, and we provide some recommendations on where the field might focus its efforts to enable rapid advancements. We conclude with summary thoughts on promoting justice, equity, diversity, and inclusion.

State-of-the-art Only recently have mathematical architectures been developed to discover governing equations directly from data. But in a short time, the diversity of methods and their capabilities have had an impact across a number of application areas. In all cases, measurements for which the signal-to-noise ratio is low (e.g. noisy data) represent significant challenges for any analysis of the underlying signal, including model discovery. Two machine learning areas have dominated the discovery of parsimonious governing equations: symbolic regression with genetic programming (GPSR) and sparse regression. These two areas draw from the broader class of dictionary learning methods and have been employed in a wide range of applications outside the materials science field. Each is described below. We anticipate that both of these methods (as well as future methods that evolve from these) can be adapted in the future to heterogeneous anisotropic multiscale problems as found in materials science.

An early method for the discovery of governing equations was symbolic regression (SR), which was pioneered by Lipson and co-workers [5, 62]. SR is a method that aims to model an input dataset without assuming its form. Instead, candidate models are proposed and evaluated by the algorithm, and the only assumption is that the data can be modeled by some algebraic expression. This is in contrast to traditional regression methods in which model form selection is made first and the regression method then estimates the model parameters. From a general perspective, SR is an optimization problem that occurs over a non-numeric domain of mathematical operators (*e.g.,* $+$, exp, sin, d/dx, *etc.*) and numeric domain of model parameters. The SR model is characterized by a variable-length combination of operators and parameters and, therefore, poses an infinite space of possible model forms to search. In practice, the mathematical operator domain is limited by a finite set of operations and a limited model length (*i.e.,* stack size). SR is also computationally expensive, especially when used with genetic programming to search for models.

Genetic programming is the most commonly used model-evolution algorithm for SR, termed

GPSR. Genetic algorithms (GA) are incredibly flexible and powerful optimization schemes proposed by Koza [32] that attempt to mimic the evolutionary selection processes observed in natural systems. Genetic programs are a type of GA in which models are represented as (nested) variable-length tree structures representing a program instead of a fixed-length list of operators and values. Within GPSR, genetic programs are used to generate random perturbations to models which are evaluated against a fitness function(s). This fitness is used to select models most likely to perform better, and then randomly recombine (*i.e.,* crossover) and permute (*i.e.,* mutate) them to generate new candidate models. At the same time, the candidates with the poorest fitness are evolved out of the population (*e.g.,* natural selection). The iterative exploration of the solution space is subject to both randomization and guidance from the particular fitness, crossover, and mutation procedures implemented.

As an example of the application of GPSR to multiscale systems, researchers have applied GPSR to multiscale constitutive models. In Bomarito *et al.* [4], GPSR was shown to learn the symbolic expression of the von Mises yield surface, provided corresponding training data from simulations of representative volume elements. Extension to learning the evolution equation for state variables (*e.g.,* plastic strain) was also demonstrated. While this demonstration "learned" an existing model, it demonstrated that GPSR can satisfy an engineering requirement: a means for verification of artificial intelligence (AI) and ML models and assessment for meaning and insight. More recently, researchers have used GPSR to learn microstructure-dependent plasticity models for additively-manufactured Inconel 718 [19]. These learned models were automatically parsed within a topology optimization code; a capability that can be included within various software environments.

Following on from the data-regression approaches mentioned above, the *sparse identification of nonlinear dynamics* (SINDy) method [9] is a SR method that leverages time-series data to discover the governing equations from a library of candidate models. The sparse regression procedure extracts the terms which best represent the time-series data. The SINDy algorithm has been broadly applied to a wide range of systems, including for reduced-order models of fluid dynamics [39, 40, 41, 23, 17, 11, 10] and plasma dynamics [15, 30], turbulence closures [2, 3, 61], nonlinear optics [67], numerical integration schemes [71], discrepancy modeling [28, 64], boundary value problems [63], identifying dynamics on Poincare maps [7, 8], tensor formulations [22], and systems with stochastic dynamics [6, 12]. In the work on plasmas [15], for instance, a predator-prey type dynamical system that approximates the underlying dynamics of the three energy state variables was discovered. Importantly, the model is amenable to a bifurcation analysis that reveals consistency between the bifurcation structures observed in the data. The integral formulation of SINDy [60, 44] has also proven to be powerful, enabling the identification of governing equations in a weak form without recourse to computing derivatives; this approach has recently been used to discover a hierarchy of fluid and plasma models [55, 24, 1, 56]. The open source software

package, PySINDy[1], has been developed in Python to integrate the various extensions of SINDy [65]. An attractive feature of SINDy is that it simply solves an over-determined $\mathbf{Ax} = \mathbf{b}$ by promoting sparsity and making it modular and amenable to enabling innovations. Moreover, it is exceptionally efficient computationally in comparison with SR, thus allowing for discovery on data with orders of magnitude less computational time. It can also be used with neural network (NN) architectures which provide automatic differentiation [49, 21] and learning coordinates and models jointly [13, 29].

In addition to the symbolic (dictionary) methods mentioned above, there is a strong surge of interest in physics-informed machine learning (PiML) [50] by means of deep neural networks (DNNs). This area evolved concurrently with the aforementioned dictionary learning methods. Interest in DNNs within the scientific community started in part due to their flexibility and expressiveness (see [59], for a first work in this direction, and [52]). While first being used as physics-constrained regressors such as physics-informed NNs (PINNs), their use has morphed into the topic of operator learning, as mentioned below. Given this trajectory, some anticipate that this line of work will transform what we as a community mean by "learning governing equations." In this context, a recent review article [31] summarizes the state of the art of PINNs and DeepONets [42], that we describe below. In PINNs-type approaches [51, 53, 54], the solution of a *known* PDE is modeled by a DNN whose parameters, together with other model parameters, are to be learned. We stress that, for these approaches, only constitutive relationships and model parameters are "discovered", whereas the fundamental underlying physics is established *a priori* (e.g., conservation laws). More recently, building on PINNs and other learning paradigms, new deep learning tools designed to discover governing relationships, have been considered. Among these we mention *neural-operator*-type approaches [35, 36, 37] and DeepONet [42]. This type of deep learning broadens the concept of "governing equations" to include complex mathematical forms such as DNNs, and, in doing so, challenges traditional notions of interpretability, as addressed below.

We summarize possible ways to embed or discover physics via machine learning in Figure 2. Here, several approaches are listed for decreasing levels of confidence in the knowledge of the underlying physics and, for each of them, we indicate where physics discovery is possible and what type of datasets are required. In the top-left box, *known* physics laws are embedded in the learning process via "strong" constraints, *i.e.,* via equality constraints to the optimization problem. In the top-middle box, physics constraints are embedded "weakly", as part of the cost functional (or loss functional) to be minimized. This approach is typical in PINNs-type algorithms [25, 48, 47, 57, 69, 70, 74] and in *nonlocal-kernel-regression* algorithms [75, 76, 77, 78]. Both approaches rely on some underlying physical knowledge in the form of an equation or a constitutive law. The top-right box represents those approaches for which physics knowledge is poor or absent. In this case, learning solely relies on data and, for this reason, it requires even richer datasets. Note that in this category we have

[1] `https://github.com/dynamicslab/pysindy`.

two types of learning: a pure data-regression approach, where given some dataset, a surrogate (*e.g.,* a NN) is trained in a least square sense, and an input-output approach, where, given input-output pairs for a specific system, the surrogate "discovers" new constitutive input-output relationships. In the latter class, we have both symbolic (dictionary) methods such as SR and SINDy and operator learning approaches such as neural operators and DeepONets.

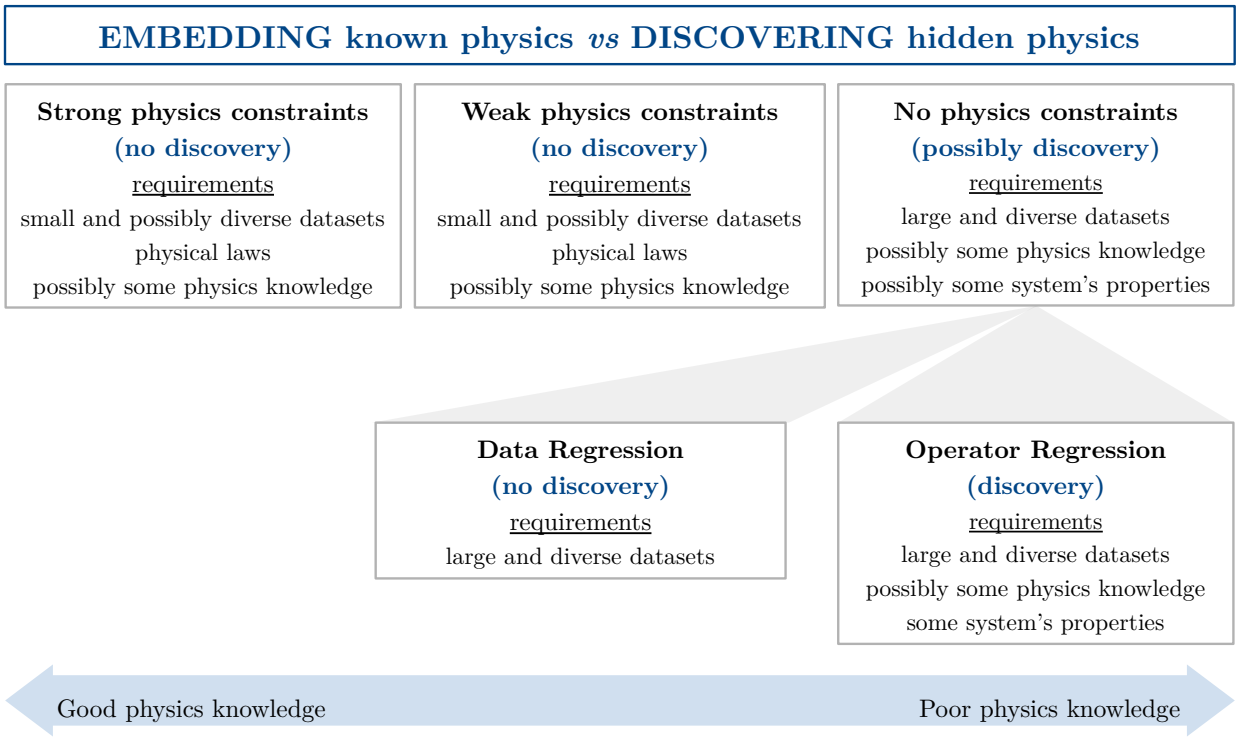


Figure 2: Different approaches to learn surrogates, from regression to discovering new physics.

Prediction, Generalization and Interpretability As the use of AI/ML technologies grows, there is an increasing number of people who believe that AI/ML representations will surpass model-based prediction capabilities. This enthusiasm has been bolstered by computer science success stories around DeepMind's AlphaGo and Tesla's self-driving car technologies. Critics of this view normally point to three desired features of models; they desire models that are predictive, generalizable, and interpretable. In addition, for these success stories, large training data sets exist whereas they often do not for heterogeneous materials problems. Since there is still debate about the precise definitions of these terms (and the debate is sometimes a consequence of the field or subfield of science in which they are used), we give here conceptual definitions that inform our discussion below. A model is

considered *predictive* if it can forecast likely future outcomes. A model is considered *generalizable* if the model can adapt to new, previously unseen data (*i.e.,* things outside the training set).[2] A model is considered *interpretable* if relevant (to the problem at hand) understanding of the relationships either contained in data or learned by the model can be inferred [45].

The AI/ML advocates argue that the hallmark of science is not "understanding" *per se* in a reductionist sense, but rather "trust" in generalization and predictive capability. When we are able to more faithfully generalize AI/ML results, might their use with materials science systems become more ubiquitous? With more generalizable AI/ML methods, difficult problems such as multi-scale, heterogeneous problems that lack traditional governing equations may be within reach. We acknowledge that at present, there is a tension between these two camps: those that stress predictive power while allowing a sacrifice of interpretability, and those that emphasize interpretability over predictive power. Both camps appear to agree on generalizability; however, they disagree as to whether predictive power or interpretability most naturally leads to capturing this constraint.

Knowledge gap and challenges

Despite the recent advances in Scientific Machine Learning (SciML), and in particular in the use of ML techniques such as dictionary learning, to learn governing equations, several modeling and computational challenges hinder their general usability in practical contexts, including the simulation of porous, heterogeneous materials. We list and discuss seven challenges, some of which lead directly to recommendations for action. We have highlighted these directly in this section; several recommended areas for action are further discussed in the following section.

The first two challenges are 1) the *interpretability* of the machine-learned surrogates and 2) their *generalization* to settings (boundary conditions, environment conditions such as temperatures, body loadings, etc.) that are substantially different from the ones used during training. These are not necessarily related to the application of interest in this report but are common gaps in the discovery of new governing equations for several engineering and scientific applications. Other challenges, strongly related to the simulation of materials, are 3) addressing the *multiscale* nature of the physical systems of interest and of the available data, 4) dealing with the presence of high degrees of *heterogeneity* at different scales, and 5) incorporating model or data *uncertainty* in the learning algorithms. We also briefly report on two additional challenges that are not directly related to the ML strategy, but highly affect its outcome, *i.e.,* 6) the need for efficient optimization techniques and 7) the availability, quality, and fidelity of the dataset, and how to incorporate multi-fidelity/multi-

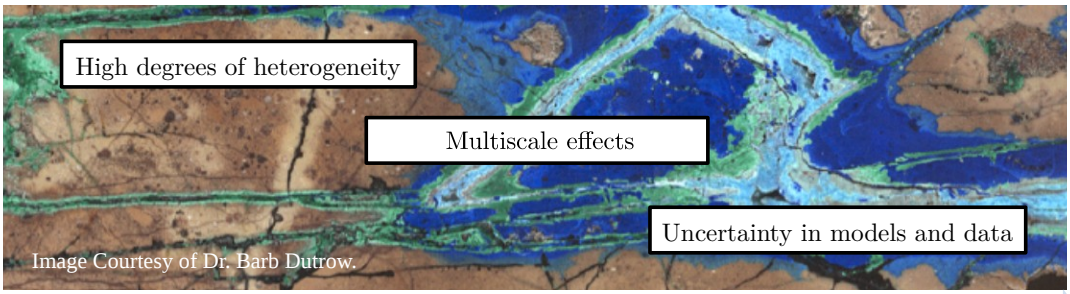
[2]We acknowledge that there is overlap between predictive power and generalization, but leave them as separate but complementary concepts.

modality data into training NNs.

A summary of the discussion that follows can be found, for the case of subsurface transport through porous media, in Figure 3. This specific application was chosen as an outstanding representative of all of scenarios considered in this report. The complexity of the subsurface environment and the difficulty in accessing such an environment, are such that dealing with multiscale effects, heterogeneity of the medium, and uncertainty in the measurements and in the models is a nontrivial task. In this figure, we list three possible learning approaches for increasing levels of abstraction of the resulting surrogate, $\mathcal{S}$, and we highlight their properties in terms of interpretability and generalization capability. Specifically, $\mathcal{S}_q$ represents a surrogate for a quantity q in the form of, *e.g.,* , a NN, a GP, a polynomial expansion, or a symbolic expression composed from a dictionary. The models listed here are by no means a fully representative set of state-of-the-art methods, but provide valuable examples of possible approaches characterized by different degrees of interpretability and generalizability. In the first column the solutions of *known* PDEs, such as Darcy and advection-dispersion equations, are modeled as surrogates together with other model parameter fields (e.g., permeability). Representatives of this technique are PINNs-type approaches [25, 69]. Clearly, the resulting surrogate is a solution of a PDE whose terms have an understood physical interpretation. On the opposite side of the spectrum of physics knowledge, we have the approach illustrated in the last column where the resulting model is an operator (or map) from some of the system's inputs to the solution, the pressure in this case. Among this class of techniques, we mention neural-operator-type approaches as well as symbolic (dictionary) methods. In this case, the interpretability depends on the method: as explained above, neural-operator-type approaches are harder to interpret whereas the interpretability of the components of a symbolic method is more straightforward.

Interpretability By interpretability, many scientists mean the ability of identifying in the resulting surrogate model a physical behavior such as diffusion, advection, reaction, etc. While this is particularly straightforward when learning coefficients of a known PDE, it is much less intuitive when learning, *e.g.,* NNs, as surrogates of a solution operator. In fact, while equations are highly interpretable, the architecture of a NN is not directly connected to a physical phenomenon. When little is known about the physics of the system at hand, the lack of interpretability affects the extent of trust in the ML surrogate.

The idea of interpretability stated above is not incorrect, but incomplete. We hold that the lens through which one should view interpretability is *decision-making*. For purposes of example, consider the following. The actions supporting decision-making often involve directly or implicitly enumerating all possible outcomes that come as a consequence of a particular choice. Prior to the modern age, this enumeration step often involved speculation and extrapolation built upon a combination of historical pattern analysis and human



increasing levels of abstraction

Good physics knowledge — Poor physics knowledge

Learning approach	Learning PDE model parameters	Learning constitutive laws or closure terms	Learning surrogates as input-output maps
Resulting Model	input: $\mathbf{x}$, output: permeability $\kappa(\mathbf{x})$, $p(\mathbf{x})$ $\nabla \cdot \mathbf{q} = f$, $\mathbf{q} = \frac{\mathcal{S}_\kappa}{\nu} \nabla \mathcal{S}_p$	input: permeability $\kappa(\mathbf{x})$ output: ith flux $q_i(\mathbf{x})$ $\nabla \cdot \mathcal{S}_\mathbf{q} = f$	input: permeability $\kappa(\mathbf{x})$ output: pressure $p(\mathbf{x})$ $p = \mathcal{S}(\kappa)$
Properties	Highly interpretable May or may not generalize	Somewhat interpretable More likely to generalize	Interpretation is method-dependent May or may not generalize
Notation	$\mathcal{S}_q$: surrogate for a quantity q in the form of, *e.g.*, a NN, a GP, a polynomial expansion, or a symbolic expression composed from a dictionary		

Figure 3: For subsurface flow, we list three possible learning approaches, their corresponding surrogates, and their properties. ML techniques are listed for increasing levels of abstraction and physics knowledge.

experience. The decision-maker's level of trust in the predictions was in part determined by their trust in the person from whom they were receiving their information and upon the *resonance* or *dissonance* of those predictions with the decision-maker's own experiences and intuitions. In the modern age, both simulation science and data science have augmented the outcome prediction step. Although the use of mathematical and statistical predictors based upon a combination of first-principle modeling and data-driven science is now ubiquitous and has often replaced human intuition as the generator of possible future outcomes, how the decision-maker engages with these tools has not changed – it is predicated on *trust*. As AI/ML tools move to the forefront of options used by decision-makers to create data-driven outcome predictions, it is important to develop strategies and tools that enable *explainable* AI/ML: tools that encourage and support trust-building.

There are a variety of ways by which decision-makers build trust in the tools that they use. In the case of simulation science, the most common trust enabler is an agreement on what are *first principles* and on how those first principle components are assembled to

yield a prediction. Many decision-makers struggling with the use of AI/ML in outcome prediction either explicitly or implicitly assume that this strategy is the only strategy for interpretability. However, this bottom-up approach to interpreting predictions is not the sole standard for interpretability. There are many systems for which, although we might vaguely understand the various first-principle building blocks that are being used, the complexity of the assembling of the system makes it impossible for us to create a cause-and-effect chain in our minds that justifies our trust[3].

An alternative way by which we gain experience and trust in complex systems is through *interrogation*. Through an understanding of the purposes of the model and the assumptions upon which it was built, we build our trust in its predictions by presenting it with scenarios for which we believe we understand what the prediction should be. We build confidence in the model each time the model, under known circumstances, reacts as we anticipate. Once this base level of trust is established, we begin to interactively interrogate the model with situations for which we may or may not have a complete idea as to what the prediction might be; however, we want to see and reason about the outcomes provided by our AI/ML recommender system. Like in the social context, this interactive session allows the decision-maker to build trust in their tools while at the same time building an appreciation of its biases and limitations.

Recommendation: We believe that research into interactive visualization and interrogation tools may play an important role in enabling explainable AI/ML for material systems. For example, the authors in [73] introduced the use of AI/ML tools for uncovering interpretable shared "hidden" structures across data spaces for design space analysis and exploration. Their work demonstrated how AI/ML tools could be used within an interactive framework (dSpaceX) to first build trust in the tools themselves, and then later to interrogate new topological optimization designs. Similarly, the authors in [38] introduced *NLIZE* – a perturbation-driven visual interrogation tool for analyzing and interpreting natural language inference models. They introduced a visualization system that, through a tight yet flexible integration between visualization elements and the underlying model, allows a user to interrogate the model by perturbing the input, internal state, and prediction while observing changes in other parts of the pipeline. They used the natural language inference problem as an example to illustrate how a perturbation-driven paradigm can help domain experts assess the potential limitation of a model, probe its inner states, and interpret and form hypotheses about fundamental model mechanisms such as attention.

Generalization ML algorithms are particularly effective in "interpolation" tasks, *i.e.,* in generating surrogates that well-represent the dataset used during training. However, sim-

[3]Although one might claim that an understanding of Newton's laws of motion combined with an understanding of thermodynamics aids someone in appreciating how an automobile transmission works, it would be an exaggeration to say that a person has built trust in the performance of their transmission due to this knowledge.

ple interpolation that only predicts the regime captured by available data is not sufficient for reliable predictions. A predictive surrogate must extrapolate (or generalize) to regimes that are different from the ones used for training. Ideally, a reliable surrogate would only embed the material's constitutive behavior and be independent of the system's inputs such as environment conditions, boundary and initial conditions, loadings, *etc.* Even more desirable would be a surrogate that generalizes beyond a specific material (*i.e.,* a surrogate for subsurface transport that provides reliable predictions regardless of the composition of the subsurface, as long as the system's inputs are available). Several SciML algorithms learn surrogates for the state of the system rather than for the constitutive behavior; this approach may lead to surrogates that are tied to specific inputs such as boundary conditions or that only represent solutions that belong to the training set. Recent works focused on learning constitutive laws or, more in general, a surrogate for the solution operator itself, are more likely to be independent of the system's inputs and, hence, to generalize better.

Finally, one has to keep in mind that with higher levels of generalization and abstraction, the price to pay might be interpretability. This is the case of *e.g.*, *neural operators*, *i.e.,* NNs that reproduce the system's behavior, where the surrogate is a NN itself and little can be said about its connection to a physical phenomenon, as anticipated above.

Recommendation: A particularly complex task is the prediction of emergent phenomena; most of the current SciML algorithms are still not able to capture anomalies that may arise in a system and that are not accounted for in the training set, such as bifurcations. This indicates an area of future research.

Multiscale nature, heterogeneity, and stochasticity Modeling and simulation of deterministic, homogeneous physical systems characterized by a single time and length scale have experienced significant progress. In the absence of multiscale effects, uncertainties (in the model and data), and heterogeneities, we are now in the position of delivering accurate, efficient, and predictive simulation tools in the area of materials science. In this context, ML is useful when some model parameters are unknown or uncertain. However, in the presence of multiscale effects, heterogeneities, and uncertainties, current PDE-based models may be insufficient to appropriately capture the system's behavior. Thus, current SciML algorithms may fail to be predictive because they were designed and tested on a single-scale, homogeneous toy problem whose physical behavior is well-known and understood. In the presence of heterogeneities and when the small scale behavior affects the system's global behavior, there is the need for new surrogates that are able to capture the effects of the small scales at the continuum level. A typical example of this situation is subsurface modeling where high degrees of heterogeneity and small scale effects that cannot be captured at large scales compromise the ability to describe the system using classical models. In this case, standard PDE-based ML algorithms may fail to be predictive by addressing one scale at a time. In such cases, a way to circumvent this challenge is to learn a NN as

a surrogate of the solution map; this approach exploits the ability of NNs to capture complex behavior thanks to their *compositional nonlinearity*. On the other hand, high degrees of heterogeneities, especially at the small scales, cannot be accurately detected due to difficulties in measuring material properties at very fine scales or in reaching specific locations (*e.g.,* deep subsurface layers). When this is the case, heterogeneity can be treated as uncertainty; as such, material properties and system's input are treated as random fields for which prior statistical information is available. Paper [43] represents the first work in this direction.

Recommendation: There is the need for additional ML tools that can learn models while embedding prior stochastic information and deliver probability distributions, rather than deterministic surrogates. In this context, as for any other uncertainty-quantification algorithm, a ML learning tool might suffer from the so-called curse of dimensionality. This creates the need for new tools that can embed uncertainty while featuring a cost that scales linearly (or sub-linearly) with the number of parameters.

Efficient optimization algorithms The latest successes in ML, in general, and deep learning, in particular, are in large part due to the recent advances in optimization algorithms, including stochastic gradient descent. However, the stochastic nature of such algorithms in combination with the non-convexity of the optimization problems in deep learning makes ML predictions uncertain even when the underlying physical system is deterministic (*e.g.,* a system with fully known governing equations and parameters). In addition, ML predictions might strongly depend on hyper-parameters; this fact further complicates their interpretability.

Data availability, quality, and fidelity As of now, the literature does not offer ML algorithms that are able to handle multiscale, multi-source, multi-resolution data in an efficient and automatic way.

Recommendation: Two additional obstacles are the ability to deal with datasets that are large in size, unstructured, and non-regular and the ability to use datasets that are small, sparse, and irregular in sampling. The former is particularly problematic when learning PDEs, as the solution is expected to feature some degree of smoothness. This indicates an area where further research investment is needed.

Recommendations to advance the field in ten years

We envisage a day when the field of mathematical modeling regularly uses AI/ML tools to aid in finding and/or refining new governing equations for complex, multiscale hetero-

geneous systems. To arrive at such a place in the future, we hold that several foundations areas must be addressed, as summarized below.

Advancing the understanding of generation and limitations of training data In comparison with many current applications where an abundance of data is available, such as computer vision problems, the problems considered here have limited data and/or measurement availability. Relevant problems in the field require dataset curation for complex, time-dependent problems. Large datasets cause slower training, while small datasets may not sufficiently capture key features. Multi-fidelity or multi-modality data presents challenges in how to incorporate or weight different types and sources of data. In data constrained modeling, knowledge of the underlying physics may be used to restrict the solution to within a given space. In all cases, data is absolutely the most important asset for leveraging the various approaches advocated.

Developing a theory of the rigorous mathematics behind ML There is also a call to develop a better understanding of the mathematical theory behind machine learning tools so we can design trustworthy computational capabilities. In particular, there are community members that stress that the success of the finite element method was not merely its applicability, but the hand-in-glove nature of theory and application that developed along its evolutionary path. There is a need to increase our confidence in ML models analogous to the theoretical understanding of FEM methods, where rigorous theory guarantees and implementations convergence. Such developments are necessary in order to provide rigorous bounds and uncertainty in modeling efforts.

Developing standardized benchmark datasets Many fields have standard benchmarks used to compare and build confidence in results from numerical methods. For machine learning, an early standard benchmark was the MNIST handwritten digit database [33]. More sophisticated data sets, including the ImageNet challenge [16], and the Penn Machine Learning Benchmark for classification [46], have provided a platform for the proper evaluation of methods. Many more benchmarking suites are currently being established for new machine learning applications, such as the Open Graph Benchmarks for machine learning on graphs [26]. Benchmark problems allow comparisons between methods by providing a standard database. Currently, no such benchmark databases exist for discovering governing equations. This puts the onus on researchers to have to spend time and resources developing their own datasets and does not allow for standardized comparisons between methods. We suggest that there is a need for a rigorous hierarchy of challenge problems to be developed for heterogeneous, anisotropic materials. The benchmark problems should cover the range of regimes from data-rich to data-poor, include applications with and without knowledge of the underlying physics, and range from toy problems to very

complex problems.

Recommendations to promote JEDI (Justice, Equity, Diversity, and Inclusion)

When discussing justice, equity, diversity, and inclusion, we start by noting that a diverse group addressing a problem is always more successful. This is especially true in the context of ML for heterogeneous materials, where a wide range of knowledge and perspectives across fields is needed to understand the underlying physics and to implement efficient ML models. We have found that the ML communities in which we participate are strongly lacking in all forms of diversity. Less than 14% of AI papers on arXiv were written by women [68], and a 2018 report estimated that just 12% of AI researchers are women [14].

In the context of ML, with its strong applications in industry, it is especially hard to retain early career, and later career scientists. In fact, the combination of benefits, salary, and lack of control over location can make careers in academic research less appealing than industry positions. These problems particularly confound the issues of recruiting and retaining a diverse staff or faculty. We note that female faculty members are 33% more likely to have full-time working spouses than male faculty members, greatly increasing the pressure on female faculty members [27]. This pressure can make higher salaries and benefits of industrial careers more appealing. One recommendation is to pay specific attention to benefits that can aid early career researchers with families, such as the childcare support provided with some NIH grants [18]. Childcare support can increase accessibility of workshops, conferences, and meetings. Many postdocs report being discouraged from taking, or not having access, to maternity leave (one study found 44% of externally funded postdocs have no access to leave for birth parents [34].) The issue is compounded for those in historically underrepresented groups, who are more likely to be discouraged from taking any parental leave [34]. When parental leave is available, the decrease in academic output due to leave can increase the pressure to move to a career viewed more compatible with having a family [79].

This pressure has been compounded by the care-taking responsibilities necessitated by the COVID-19 pandemic, reducing the number of publications by female authors in 2021 compared to 2019 [58, 72, 20]. This reduction in publishing can and will have long-term impacts on the careers of female researchers. Simultaneously, the COVID-19 pandemic has increased the prevalence of flexible work environments, remote work, and virtual and hybrid collaborations and conferences. In particular, virtual conferences have increased the participation of students and women significantly [66]. Those who have family obligations or have limited ability to travel due to disabilities or funding can participate in virtual conferences while in-person conferences may be inaccessible. We encourage the field to examine

the role of virtual and hybrid work and events in increasing inclusivity within the field. As the field begins to come out of the COVID-19 pandemic and adopt new modes of work, we encourage consideration of the lessons learned from the pandemic to be continued forward.

Acknowledgments

Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This report, SAND2021-XXXX, describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

Pacific Northwest National Laboratory is operated by Battelle for the DOE under Contract DE-AC05-76RL01830.

References

[1] E Paulo Alves and Frederico Fiuza. "Data-driven discovery of reduced plasma physics models from fully-kinetic simulations". In: *arXiv preprint arXiv:2011.01927* (2020).

[2] S Beetham and J Capecelatro. "Formulating turbulence closures using sparse regression with embedded form invariance". In: *Physical Review Fluids* 5.8 (2020), p. 084611.

[3] Sarah Beetham, Rodney O Fox, and Jesse Capecelatro. "Sparse identification of multiphase turbulence closures for coupled fluid–particle flows". In: *Journal of Fluid Mechanics* 914 (2021).

[4] GF Bomarito et al. "Development of interpretable, data-driven plasticity models with symbolic regression". In: *Computers & Structures* 252 (2021), p. 106557.

[5] Josh Bongard and Hod Lipson. "Automated reverse engineering of nonlinear dynamical systems". In: *Proceedings of the National Academy of Sciences* 104.24 (2007), pp. 9943–9948.

[6] Lorenzo Boninsegna, Feliks Nüske, and Cecilia Clementi. "Sparse learning of stochastic dynamical equations". In: *The Journal of Chemical Physics* 148.24 (2018), p. 241723.

[7] Jason J Bramburger and J Nathan Kutz. "Poincaré maps for multiscale physics discovery and nonlinear Floquet theory". In: *Physica D: Nonlinear Phenomena* 408 (2020), p. 132479.

[8] Jason J Bramburger, J Nathan Kutz, and Steven L Brunton. "Data-driven stabilization of periodic orbits". In: *IEEE Access* 9 (2021), pp. 43504–43521.

[9] S. L. Brunton, J. L. Proctor, and J. N. Kutz. "Discovering governing equations from data by sparse identification of nonlinear dynamical systems". In: *Proceedings of the National Academy of Sciences* 113.15 (2016), pp. 3932–3937.

[10] Jared L Callaham, Steven L Brunton, and Jean-Christophe Loiseau. "On the role of nonlinear correlations in reduced-order modeling". In: *arXiv preprint arXiv:2106.02409* (2021).

[11] Jared L Callaham et al. "An empirical mean-field model of symmetry-breaking in a turbulent wake". In: *arXiv preprint arXiv:2105.13990* (2021).

[12] Jared L Callaham et al. "Nonlinear stochastic modelling with Langevin regression". In: *Proceedings of the Royal Society A* 477.2250 (2021), p. 20210092.

[13] K. Champion et al. "Data-driven discovery of coordinates and governing equations". In: *Proceedings of the National Academy of Sciences* 116.45 (2019), pp. 22445–22451.

[14] Casey Chin. "AI is the Future—But Where Are the Women?" In: *WIRED. August* 17 (2018).

[15] Magnus Dam et al. "Sparse identification of a predator-prey system from simulation data of a convection model". In: *Physics of Plasmas* 24.2 (2017), p. 022310.

[16] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[17] Nan Deng et al. "Galerkin force model for transient and post-transient dynamics of the fluidic pinball". In: *Journal of Fluid Mechanics* 918 (2021).

[18] "Family-Friendly Initiatives". In: *National Institutes of Health* (). `https://grants.nih.gov/grants/policy/nih-family-friendly-initiative.htm`.

[19] K Garbrecht et al. "Interpretable machine learning for texture-dependent constitutive models with automatic code generation for topological optimization". In: *Integrating Materials and Manufacturing Innovation* Accepted.IMMJ-D-21-00080R1 (2021).

[20] Angèle Gayet-Ageron et al. "Female authorship of covid-19 research in manuscripts submitted to 11 biomedical journals: cross sectional study". In: *bmj* 375 (2021).

[21] Maximilian Gelbrecht, Niklas Boers, and Jürgen Kurths. "Neural partial differential equations for chaotic systems". In: *New Journal of Physics* 23.4 (2021), p. 043005.

[22] Patrick Gelß et al. "Multidimensional approximation of nonlinear dynamical systems". In: *Journal of Computational and Nonlinear Dynamics* 14.6 (2019).

[23] Yifei Guan, Steven L Brunton, and Igor Novosselov. "Sparse nonlinear models of chaotic electroconvection". In: *Royal Society Open Science* 8.8 (2021), p. 202367.

[24] Daniel R Gurevich, Patrick AK Reinbold, and Roman O Grigoriev. "Robust and optimal sparse regression for nonlinear PDE models". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 29.10 (2019), p. 103113.

[25] QiZhi He et al. "Physics-Informed Neural Networks for Multiphysics Data Assimilation with Application to Subsurface Transport". In: *Advances in Water Resources* (2020), p. 103610.

[26] Weihua Hu et al. "Open graph benchmark: Datasets for machine learning on graphs". In: *arXiv preprint arXiv:2005.00687* (2020).

[27] Jerry A Jacobs and Sarah E Winslow. "Overworked faculty: Job stresses and family demands". In: *The ANNALS of the American academy of Political and Social Science* 596.1 (2004), pp. 104–129.

[28] Kadierdan Kaheman et al. "Learning Discrepancy Models From Experimental Data". In: *CDC [arXiv preprint arXiv:1909.08574]* (2019).

[29] Manu Kalia et al. "Learning normal form autoencoders for data-driven discovery of universal, parameter-dependent governing equations". In: *arXiv preprint arXiv:2106.05102* (2021).

[30] Alan A Kaptanoglu et al. "Physics-constrained, low-dimensional models for MHD: First-principles and data-driven approaches". In: *Physical Review E* 104.015206 (2021).

[31] George Em Karniadakis et al. "Physics-informed machine learning". In: *Nature Reviews Physics* 3 (2021), pp. 422–440.

[32] J.R. Koza. "Genetic programming as a means for programming computers by natural selection". In: *Statistics and Computing* 4 (1994), pp. 87–112.

[33] Yann LeCun and Corinna Cortes. "MNIST handwritten digit database". In: (2010). URL: `http://yann.lecun.com/exdb/mnist/`.

[34] Heidi Ledford. "US postdocs face steep challenges when starting families". In: *Nature. doi* 10 (2017).

[35] Zongyi Li et al. "Fourier neural operator for parametric partial differential equations". In: *arXiv preprint arXiv:2010.08895* (2020).

[36] Zongyi Li et al. "Multipole graph neural operator for parametric partial differential equations". In: *arXiv preprint arXiv:2006.09535* (2020).

[37] Zongyi Li et al. "Neural operator: Graph kernel network for partial differential equations". In: *arXiv preprint arXiv:2003.03485* (2020).

[38] Shusen Liu et al. "NLIZE: A Perturbation-Driven Visual Interrogation Tool for Analyzing and Interpreting Natural Language Inference Models". In: *IEEE Transactions on Visualization and Computer Graphics* 25 (1 2019).

[39] J.-C. Loiseau and S. L. Brunton. "Constrained Sparse Galerkin Regression". In: *Journal of Fluid Mechanics* 838 (2018), pp. 42–67.

[40] J.-C. Loiseau, B. R. Noack, and S. L. Brunton. "Sparse reduced-order modeling: sensor-based dynamics to full-state estimation". In: *Journal of Fluid Mechanics* 844 (2018), pp. 459–490.

[41] Jean-Christophe Loiseau. "Data-driven modeling of the chaotic thermal convection in an annular thermosyphon". In: *Theoretical and Computational Fluid Dynamics* 34.4 (2020), pp. 339–365.

[42] Lu Lu et al. "Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators". In: *Nature Machine Intelligence* 3.3 (2021), pp. 218–229.

[43] Xuhui Meng et al. "Learning Functional Priors and Posteriors from Data and Physics". In: *arXiv preprint arXiv:2106.05863* (2021).

[44] Daniel A Messenger and David M Bortz. "Weak SINDy for partial differential equations". In: *Journal of Computational Physics* (2021), p. 110525.

[45] W. James Murdoch et al. "Definitions, methods, and applications in interpretable machine learning". In: *Proceedings of the National Academy of Sciences* 116.44 (2019), pp. 22071–22080. ISSN: 0027-8424. DOI: 10.1073/pnas.1900654116. eprint: https://www.pnas.org/content/116/44/22071.full.pdf. URL: https://www.pnas.org/content/116/44/22071.

[46] Randal S Olson et al. "PMLB: a large benchmark suite for machine learning evaluation and comparison". In: *BioData mining* 10.1 (2017), pp. 1–13.

[47] Guofei Pang, Lu Lu, and George Em Karniadakis. "fPINNs: Fractional physics-informed neural networks". In: *SIAM Journal on Scientific Computing* 41.4 (2019), A2603–A2626.

[48] Guofei Pang et al. "nPINNs: nonlocal Physics-Informed Neural Networks for a parametrized nonlocal universal Laplacian operator. Algorithms and Applications". In: *Journal of Computational Physics* 422 (2020), p. 109760.

[49] Christopher Rackauckas et al. "Universal differential equations for scientific machine learning". In: *arXiv preprint arXiv:2001.04385* (2020).

[50] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. *Physics informed learning machine*. US Patent 10,963,540. Mar. 2021.

[51] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations". In: *Journal of Computational Physics* 378 (2019), pp. 686–707.

[52] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. "Multistep neural networks for data-driven discovery of nonlinear dynamical systems". In: *arXiv preprint arXiv:1801.01236* (2018).

[53] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. "Physics Informed Deep Learning (Part I): Data-driven Solutions of Nonlinear Partial Differential Equations". In: *arXiv preprint arXiv:1711.10561* (2017).

[54] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. "Physics Informed Deep Learning (Part II): Data-driven Discovery of Nonlinear Partial Differential Equations". In: *arXiv preprint arXiv:1711.10566* (2017).

[55] Patrick AK Reinbold, Daniel R Gurevich, and Roman O Grigoriev. "Using noisy or incomplete data to discover models of spatiotemporal dynamics". In: *Physical Review E* 101.1 (2020), p. 010203.

[56] Patrick AK Reinbold et al. "Robust learning from noisy, incomplete, high-dimensional experimental data via physically constrained symbolic regression". In: *Nature communications* 12.1 (2021), pp. 1–8.

[57] Brandon Reyes et al. "Learning unknown physics of non-Newtonian fluids". In: *Phys. Rev. Fluids* 6 (July 2021), p. 073301.

[58] Alana K Ribarovska et al. "Gender inequality in publishing during the COVID-19 pandemic". In: *Brain, behavior, and immunity* 91 (2021), p. 1.

[59] R Rico-Martinez, JS Anderson, and IG Kevrekidis. "Continuous-time nonlinear signal processing: a neural network based approach for gray box identification". In: *Proceedings of IEEE Workshop on Neural Networks for Signal Processing*. IEEE. 1994, pp. 596–605.

[60] Hayden Schaeffer and Scott G McCalla. "Sparse model selection via integral terms". In: *Physical Review E* 96.2 (2017), p. 023302.

[61] Martin Schmelzer, Richard P Dwight, and Paola Cinnella. "Discovery of Algebraic Reynolds-Stress Models Using Sparse Symbolic Regression". In: *Flow, Turbulence and Combustion* 104.2 (2020), pp. 579–603.

[62] Michael Schmidt and Hod Lipson. "Distilling free-form natural laws from experimental data". In: *Science* 324.5923 (2009), pp. 81–85.

[63] Daniel E Shea, Steven L Brunton, and J Nathan Kutz. "SINDy-BVP: Sparse identification of nonlinear dynamics for boundary value problems". In: *Physical Review Research* 3.2 (2021), p. 023255.

[64] Brian de Silva et al. "Discovery of physics from data: Universal laws and discrepancy models". In: *arXiv preprint arXiv:1906.07906* (2019).

[65] Brian M de Silva et al. "PySINDy: a Python package for the sparse identification of nonlinear dynamics from data". In: *Journal of Open Source Software* 5.49 (2020), p. 2104.

[66] Matthew Skiles et al. "Beyond the carbon footprint: Virtual conferences increase diversity, equity, and inclusion". In: (2020).