Beyond Classification and Prediction: The Promise of Physics-Informed Machine Learning in Astronomy and Cosmology

Helen Meskhidze*

Written for: *Juan Durán and Giorgia Pozzi (Eds.): Philosophy of Science for Machine Learning: Core Issues and New Perspective, Synthese Library.*

Abstract

Though the use of machine learning (ML) is ubiquitous in astrophysics and cosmology, many still see the opacity of ML algorithms as a major issue to their scientific utility. One way of addressing this opacity is through an emerging trend in ML research of "teaching" ML algorithms physical laws and domain-specific knowledge. "Physics-informed machine learning" (PIML), as this methodology is called, promises to produce better predictions and yield more interpretable algorithms. It does so by using physical principles in the training process and/or by using physical principles to guide the development of the neural network architecture. In this chapter, I investigate two uses of PIML in astronomy/cosmology, each a representative example of the two PIML methods. In both cases, PIML provides improvements in terms of the predictions and efficiency of ML algorithms. However, I argue that only in the second case does PIML offer any improvement in terms of the interpretability of the algorithms.

Keywords: opacity, interpretability, understanding, machine learning, neural network, astronomy, cosmology

1 Introduction

"Physics-informed machine learning" (PIML) has been called "the next generation of artificial intelligence" by popular media (Andrzejczuk 2023) and the scientists who use it claim that it can "transform our modeling, simulation, and understanding of complex physical systems in various science and engineering disciplines" (Chen, Liu, and Sun 2021). PIML aims to incorporate physical laws and domain-specific knowledge into machine learning in order to produce better predictions and yield more interpretable algorithms. Domain-specific knowledge, in this context, can

*Black Hole Initiative, Harvard University, 20 Garden Street, Cambridge, MA 02138, emeskhidze@g.harvard.edu

include symmetry laws, conservation laws, or even the governing dynamics of a system.

Two means have been proposed for the incorporation of such physical laws and domain-specific knowledge: by "teaching" machine learning algorithms this information and/or by designing specialized network architectures. In either case, physics-informed machine learning is argued to yield better predictions in the presence of imperfect or noisy data because the algorithms are more robust to any small irregularities in the data (Karniadakis, Kevrekidis, Lu, et al. 2021, 423). More importantly, though, PIML is thought to be more interpretable because our knowledge of the physical world is incorporated from the start.

Since their introduction in Raissi, Perdikaris, and Karniadakis (2019), PIML methods have been developed and used in numerous domains across the sciences. Some examples of projects include performing parameter estimation for applications in systems biology (Daneker, Zhang, Karniadakis, and Lu 2023), modeling fluid dynamics (Cai, Wang, Fuest, Jeon, Gray, and Karniadakis 2021), characterizing a crack in the surface of a material (Shukla, Di Leoni, Blackshire, Sparkman, and Karniadakis 2020), predicting the many-electron wave equation for applications in quantum chemistry (Pfau, Spencer, Matthews, and Foulkes 2020), and forecasting weather/climate processes (Kashinath, Mustafa, Albert, et al. 2021). In astronomy and cosmology, PIML has been used to, e.g., model the formation of molecular clouds in the interstellar medium (Branca and Pallottini 2023), solve the radiative transfer equation for supernova simulations (Chen, Jeffery, Zhong, et al. 2022; Mishra and Molinaro 2021), investigate astrophysical shocks (Moschou, Hicks, Parekh, et al. 2023), and model the gravitational fields around small astrophysical objects (Martin and Schaub 2022a). These last two projects will be the case studies discussed in detail in the present chapter.

Given the rising use of PIML in various domains, questions about the prospects of the methodology and whether it fulfills its promise of increased scientific understanding are pressing. Are physics-informed ML algorithms truly novel in methodology? In what sense are they the "next generation of artificial intelligence"? Are they able to overcome the issues of opacity faced by standard ML algorithms that worry scientists and philosophers alike? These questions will be the focus of this chapter

I will begin by reviewing some standard uses of machine learning in astronomy and cosmology (§2). I then turn to philosophical analyses of the issues raised by ML (§3). With these issues in mind, I present two case studies (§4). These are representative of the two methods of PIML: incorporating physical principles into the training of the ML algorithm and designing a specialized network architecture. In both cases, PIML improves the predictions and efficiency of ML algorithms. However, I argue that improvement in the interpretability of the algorithms only occurs in the second case. This is because the inclusion of physical principles as part of the training of the ML algorithm is insufficient to ground claims of increased transparency.

2 Machine learning in Astronomy and Cosmology

The use of machine learning has a long history in astronomy and cosmology with some of the first uses dating back to the 1980s (see Miller 1993 for a review). Given the large amounts of data astronomers typically collect, early projects primarily used neural networks (NNs) for object classification on existent data sets. Occasionally, astronomers took advantage of the computational speeds offered by ML for real-time applications. For instance, another early application of ML in astronomy was for detector event filtering in high-energy telescopes (a kind of classification but used in real-time to filter events; see Meetre and Norris 1991). Another real-time application of ML during this time was for adaptive telescope optics, to quickly adjust telescope mirrors in response to atmospheric distortions (Angel, Wizinowich, Lloyd-Hart, and Sandler 1990). While projects today have broadened the scope of the use of ML/NNs, many still use NNs for image classification-related tasks. Below, I review a few significant, contemporary uses of ML in astronomy and cosmology. Though the projects detailed below constitute only a small sample of the various uses of ML in astronomy and cosmology today, they provide a sense of the scope of the use of such methods.[1] Understanding the applications of ML in astronomy/cosmology will help us understand what issues are relevant to these fields. Then, we will be able to situate PIML in this context and assess whether it resolves the pressing issues.

2.1 Imaging and data processing

Applications of machine learning to imaging are the most common use of ML in astronomy, especially to identify individual objects. For instance, ML has been used for the detection of strong gravitational lenses (Metcalf, Meneghetti, Avestruz, et al. 2019). These are astrophysical objects (like galaxies) that cause the light from even more distant objects to be bent as it travels to the Earth. Such objects can be used to probe dark matter or various cosmological parameters of interest. ML has also been used to comb through big data sets in search of regularities that indicate the presence of pulsars (a spinning neutron star emitting radiation from its poles; Lin, Li, and Luo 2020) or even exoplanets (planets that are part of other solar systems; Jin, Yang, and Chiang 2022). Such projects involve sifting through a large amount of astronomical data to look for regularities, a task that ML is particularly well-suited for.

A more complex application of machine learning to imaging has to do with astronomical catalogs. These catalogs are created by taking images of different parts of the sky and matching them up to create a larger image, a catalog. Typically, astronomers use a “best-fit” model of the constituents of each image to synthesize them into a catalog. This process involves a kind of source matching: a source that is identified in one image is found in another, allowing these images to be lined up. Machine learning has

[1] See George Stein’s GitHub page (2023) for a more comprehensive listing of various contemporary projects using ML in astronomy/cosmology.

the potential to allow probabilistic cataloging of such images, allowing one to consider a distribution across the possible constituents of an image (Brewer, Foreman-Mackey, and Hogg 2013). When images are crowded and noisy, this kind of probabilistic cataloging can be very valuable.

Along similar lines, machine learning has been used for data processing. Data processing is a complex undertaking for astronomers, especially given the large amount of atmospheric interference many telescopes are subject to. Some of the uses of ML for data processing in astronomy are similar to applications of ML for image analysis. For instance, algorithms like Source Extractor have been developed to detect and extract foreground sources in an image (i.e., sources that are between the detector and the object of interest in the image; Bertin and Arnouts 1996). Other algorithms have been developed to remove sensor artifacts, cosmic rays, etc. These kinds of algorithms often operate in the data processing "pipeline" between when data is collected by the instrument and when the scientist can use the data.

One emerging use of ML in the image analysis context is with multi-messenger astronomy. Here, astronomers conduct wide-field surveys, sweeping the sky for any objects of interest. If such an object is detected, the goal is then to use another instrument (often in another frequency) to analyze the object in more detail. Thus, speed is critical. One must be able to sift through the data collected by the wide-field survey quickly to then point another instrument at the object of interest (see, e.g., Narayan, Zaidi, Soraisam, et al. 2018 for an example of real-time classification of data from the Large Synoptic Survey Telescope). Much of the research in this area uses supervised or semi-supervised ML, but there is space to employ unsupervised methods to detect rare events (see e.g., Li, Ragosta, Clarkson and Bianco 2021).

2.2 Simulations

A particularly interesting use of machine learning that bridges data analysis and simulations is in cosmic parameter estimation. Here, the task is to determine the value for various cosmological parameters, usually by comparing simulations to observations of the large-scale structure of the universe. In particular, astronomers and cosmologists require a huge suite of simulations to perform parameter estimation. This is because one needs to compare the collected data with possible values of the various cosmological parameters. Running the suite of simulations is very computationally costly. Therefore, some astronomers and cosmologists have instead run a small suite of such simulations to train an "emulator," a machine learning algorithm aimed at recreating these results for many different parameter combinations but for a fraction of the computational cost (see, e.g., Heitmann, Higdon, White, et al. 2009). One then uses machine learning to pick out what combination/values of parameters best capture the data. In short, machine learning is being used to interpolate amongst the various combinations of parameters and the results are then compared to observations for scientifically interesting results.

In a recent project along these lines, Villaescusa-Navarro and collaborators investigate the scales of data required to extract cosmological information (2022). They study whether one can use ML methods to extract cosmological information from data collected on the scale of a single galaxy. To do so, they train an NN on the output of many simulations, produced using various values of cosmological parameters. Then, they provide the NN with data from a single galaxy and ask it to predict the cosmological model that is consistent with that particular data set. Given the limited data that Villaescusa-Navarro and collaborators provide to the trained NN, it must be very sensitive to the impacts of changes in the various cosmological parameters.

2.3 Methods of ML

Given the numerous astronomical projects that use ML for imaging-related tasks, convolutional NNs tend to be the most suitable and commonly used architecture.[2] However, in a review of the state of the field, Dvorkin and collaborators critique such methods. They note that "...many network architectures are sequential and recursive, which does not easily allow parallel computation (to account for the size of cosmological data sets) or they assume a Markov process, which means that they cannot easily learn long-range dependencies" (Dvorkin, Mishra-Sharma, Nord, et al. 2022, 13). Thus, they note, there is an opportunity to develop architectures specifically geared towards the context of cosmology. In particular, they highlight PIML, noting that such methods "have shown promise in reducing the dimensionality of the underlying latent space of a network with an associated reduction in the size of data sets needed to train the network" (Dvorkin, Mishra-Sharma, Nord, et al. 2022, 13). Besides offering advances in efficiency, PIML is also thought to improve the interpretability of ML algorithms. Thus, before considering PIML in more detail, I will first review the issues of interpretability/opacity that arise with standard uses of ML.

3 Interpretability of ML

Given the importance of machine learning methods for scientific progress, many have argued that greater attention needs to be given to better understanding these methods. Philosophers, scientists, and computer scientists alike have turned their attention to this task. The fields of explainable artificial intelligence (XAI) and interpretable machine learning (IML) both aim to tackle questions along these lines. Some propose novel ML algorithms for the task while others propose frameworks in which to situate such questions. Much progress has been made in clarifying what is at issue, outlining frameworks to think within, and defining key terms in the debate. I outline some themes from this literature next in order to ultimately ask whether physics-informed machine learning is responsive to the demands being made.

[2] An explanation of why CNNs are well-suited for imaging-related tasks generally is beyond the scope of this chapter (see Stewart 2019 for such an explanation). Put very briefly, one can imagine that the architecture of a CNN respects the relationships among the pixels in an image.

3.1 Interpretability through transparency

In Lipton's influential paper, "The Mythos of Model Interpretability," he remarks

> ML-based systems do not know why a given input should receive some label, only that certain inputs are correlated with that label...As ML penetrates critical areas such as medicine, the criminal justice system, and financial markets, the inability of humans to understand these models seems problematic. (2018, 3)

Lipton reflects that though many propose "interpretability" as a remedy, few articulate precisely what they mean by interpretability.[3] With this in mind, he outlines five goals one might hold when demanding interpretability: increasing our trust in an ML model, helping support causal reasoning based on the model, transferring/generalizing the findings of the ML model to nearby contexts, being an informative component in human decision-making, and helping facilitate fair and ethical decision making (2018, 9-12).

Lipton then turns to considering how we might achieve these goals. The first method he considers for achieving interpretability is transparency.[4][5] Transparency, he argues, can be evaluated at various levels: at the level of the model, at the level of individual components, and at the level of the training algorithm (2018, 12). These levels each correspond to their own notion of transparency: simulatability, which captures the sense in which a model can be understood as a whole; decomposability, which captures the sense in which one can understand the various components of the model individually; and algorithmic transparency, which captures the sense in which one can understand the learning algorithm itself. Ultimately, Lipton argues that whether one algorithm is more interpretable than another will turn on what notion of interpretability one subscribes to and what kind of transparency that might demand. For instance, deep neural networks (DNNs) may be algorithmically more complex than linear models, but the high-dimensional or heavily engineered features of linear models mean that they may lose simulatability or decomposability respectively. Thus, in those domains, DNNs may exhibit more transparency and thus better interpretability.

The importance of specifying what kind of transparency one is interested in has also been highlighted in the recent philosophical literature. Creel (2020) argues that transparency is an important goal when considering opaque, complex computational systems. She presents three avenues for pursuing transparency: functional, structural, and run. Functional transparency has to do with the algorithmic functioning of the whole, i.e., understanding the high-level logical rules governing the system. Structural transparency recognizes that algorithms can be multiple realized so

[3] See also Chapters 7, 8, and 9 in this volume for more on interpretability.

[4] Lipton also discusses post-hoc methods for increasing interpretability, but they are not important for the present purposes.

[5] See Chapters 1 and 2 in this volume for a discussion of issues of opacity. For an alternative to transparency—computational reliabilism—see Chapter 4 in this volume.

it asks how the algorithm is realized in the code. Finally, run transparency asks whether one has knowledge of the program as it was actually run in a particular instance (Creel 2020, 572-582). With this framework in mind, let us now turn to evaluating the uses of ML in astronomy and cosmology and the kinds of transparency they may or may not exhibit.

3.2 Transparency of ML in Astronomy and Cosmology

ML algorithms are routinely referred to as "black boxes." This might lead us to think that the prospects for transparency when using ML are rather bleak. However, recall that in many of the uses of ML in astronomy/cosmology described above, the ML algorithms are just being used as statistical tools. It is precisely because they are used as statistical tools that we need not be worried about issues of transparency. Consider, for instance, the algorithms used to emulate the results of large-scale structure simulations. As mere interpolation devices, their results are neither surprising nor mysterious. They take a coarsely sampled parameter space as input and fill it in. In cases like this, a particular step in the process is being black-boxed using machine learning. However, since there is a well-understood, physically-motivated methodology one can always refer back to, the scientists could easily open up the black box. This means the machine learning algorithms can be made to have functional, structural, and run transparency. The overall functioning of the algorithm follows clear statistical rules appropriate for interpolation, the structure of the algorithm is relatively straightforward, and the result of the algorithm in any particular instance can be investigated. If one were to ask for a more physical explanation of any particular result, the simulations used to train the emulator can be appealed to. Put differently, we can distinguish between the "black-boxing" occurring in this instance—an (unproblematic) methodological step—and using a black box to provide understanding. This is a distinction I have argued for in more detail elsewhere (Meskhidze 2023). There, I argue that the case of emulators is an unproblematic use of methodological black-boxing.

Another way of distinguishing problematic from unproblematic uses of ML is through Sullivan's notion of "link uncertainty." In "Understanding from Machine Learning Models" (2022), Sullivan argues that the complexity or black-box nature of a model need not undercut the understanding that the model can provide. Instead, she argues that this understanding is undercut when "[t]here is a high level of *link uncertainty*, that is, a lack of scientific and empirical evidence supporting the link that connects the model to the target phenomenon" (2022, 6). To illustrate this point, she contrasts a DNN that identifies cases of melanoma from images of moles on skin to a DNN that uses facial recognition to identify an individual's sexual orientation. Since, as she argues, "The level of scientific justification and background knowledge linking the appearance of moles to instances of melanoma is extensive," (2022, 23) the link uncertainty is greatly reduced. But the link uncertainty between facial features and sexual orientation is much higher. The transparency of the DNNs in these two cases is largely the same and irrelevant to understanding. Thus, what matters for assessing understanding, on Sullivan's view, is the degree of link (un)certainty.

In their review of the recent philosophical discussion about the interpretability of machine learning, Beisbart and Räz discuss Sullivan's argument. They note the following consequence of her view: "Black boxes need not compromise the scientists' abilities to use a model if they merely black-box the implementation of the fulfillment of a known task (e.g. the calculation of a factorial)" (Beisbart and Räz 2022, 5). I agree with the claim here and would also refer to this process as unproblematic "black-boxing." Indeed, in the case of emulators, the link certainty is well-established; one can always run a simulation to check that the link between a particular cosmological parameter and the resultant large-scale structure of the universe accords with the predictions of the emulator. Therefore, we can conclude that issues of interpretability are not troubling in many of the above-described applications of machine learning in astronomy and cosmology.

There are some cases in which we do lack the necessary interpretability. Consider a simple case of image analysis. We may want to know, for instance, why the ML algorithm predicts an exoplanet from a set of input data. In this case, the ML algorithm is "making a decision" with respect to the image so the kind of statistical explanation proposed above for emulators might not be satisfactory. On Sullivan's account, this case seems akin to using moles for melanoma detection: we have clear reason to think that various aspects of the image correlate with the properties of interest. We have link certainty. Nonetheless, one might still worry whether the ML has selected some set of *real* indicators for the context/object of interest or whether it has captured some spurious correlations. In other words, though the link may be well-established, unless we are confident that the ML has captured that link, we might still worry about the interpretability and robustness of the results. If, following Lipton, we want our ML algorithms to support causal reasoning and allow generalizations, we need enough transparency to ensure that whatever "link" our ML algorithm is using to make predictions is the "link" we believe exists in the data. Put differently, while a lack of link certainty does undermine interpretability, having link certainty (without transparency) does not immediately grant interpretability.

The limitations of using link certainty to guide understanding can be made even more pressing with the following considerations. In "Two Dimensions of Opacity and the Deep Learning Predicament," Boge introduces a distinction between "h-opacity" and "w-opacity" where the former concerns *how* the machine is learning while the latter concerns *what* is learned by the machine. On his account, w-opacity is what sets deep learning apart from other scientific modeling. The features of the data picked out by the deep learning model correspond to "automatically discovered insights; complex, non-obvious features that can be abstracted from the data and allow the machine to discriminate" (Boge 2022, 61). This framing is helpful for what I am highlighting above: though some features of interest may exist in the data and the link may exist in the world, w-opacity undermines our ability to say whether these are the features the algorithm is leveraging. This is, of course, worrisome: there may or may not be link certainty but we have no way of determining which is the case. However, Boge presents a more worrisome possibility: that the ML leverages real links in the data that are, at present, inconceivable to humans (see also Boge 2023). In such a case, there would be link certainty in the sense that the ML would be leveraging real links. But we would not be able to say this as we could not conceive of the very link being exploited. These are cases in which, perhaps if we had greater transparency, we would be able to use the ML to make novel discoveries. The implication is that, in some cases, requiring link certainty would undermine our ability to make novel discoveries with ML.[6]

Creel's discussions of opacity/transparency can help pinpoint where we want greater clarity. It may not completely resolve the more troubling cases Boge is worried about, but it will assist in the cases of interest for the present chapter. On Creel's taxonomy, what we are worried about is the functional transparency of the algorithm. An example Creel gives nicely parallels this discussion. She writes:

> ...in all but the smallest networks it would be difficult to predict the outcome without tracing each step or to understand the behavior of the network, especially if the network includes feedback loops. More importantly, without further analysis it would be unclear to the observer why this neural net successfully classified an image and to what extent each of the neurons contributed to the result or why different neural nets might have different patterns of classification. In this sense, although we know how the learning algorithm works and what formal guarantees (if any) we have about its performance, we do not know how the learned 'algorithm' brings about the classification result. Thus, we lack functional transparency. (2020, 579-580)

To summarize: there are many frameworks for discussing interpretability, some highlight transparency while others highlight link certainty. Here, I have shown that, while appealing, link certainty is insufficient for interpretability. This is because, without transparency, we cannot be sure that the link captured by the ML algorithm is the actual link we expect to exist. Though many uses of ML in astronomy/cosmology are unproblematic because the ML algorithm is just being used as a statistical tool, some uses still highlight the need for interpretability. With all this in mind, we can now turn to physics-informed ML in order to ultimately ask whether such methods yield more interpretable ML.

[6] My worries in the present chapter are confined to the first category. For discussion of the second (i.e., cases when scientists do not have the resources to conceptualize the link being exploited), see Chapter 19 in this volume. There, Boge and de Regt discuss the prospects of using ML for discovering novel concepts and phenomena in particle physics and the problems that arise.

4 Physics-Informed Machine Learning

4.1 General methodology of PIML

The first projects and papers introducing physics-informed machine learning claim that the methodology is an improvement over standard ML approaches. When standard approaches are trained on huge collections of data, they may make physically inconsistent or implausible predictions due to biases that might be in the data or because the ML is being asked to extrapolate beyond the domain represented by the data. Authors advocating for PIML claim,

> ...there is a pressing need for integrating fundamental physical laws and domain knowledge by 'teaching' ML models about governing physical rules, which can, in turn, provide 'informative priors'—that is, strong theoretical constraints and inductive biases on top of the observational ones. (Karniadakis, Kevrekidis, Lu, et al. 2021, 423)

To understand this claim it will be essential to understand the methodology of such algorithms. In general, there are two methods for introducing physics into ML. The first is by including physical constraints or principles into the learning process, creating a learning bias. Projects in this vein might train the NN on both the data and a partial differential equation (PDE) known to model the situation. This is accomplished by having terms in the loss function for both the alignment of the NN with the data and the PDE. More specifically, the general form of the solution is fed into the network and soft penalties ensure that the network finds the relevant parameters. The second method uses the architecture of the network itself to incorporate the physical principles. Projects in this vein tend to be more diverse than those in the previous group. One salient example involves using graph neural networks to learn the structure of chemical bonds. There, the success of the architecture for the particular purpose relies on "their ability to pick up on structures in the graph at multiple different scales, while satisfying the crucial requirement that the output be invariant to permutations of the vertices" (Hy, Trivedi, Pan, Anderson, and Kondor 2018, 1). To better understand these methodologies, I now turn to describing two case studies, each of which highlights the use of one of the two methods of PIML in astronomy/cosmology.

4.2 Astrophysical shocks

PIML methods have been used to study astrophysical shocks. Here, I consider the use of PIML to study the solar termination shock in particular. The solar termination shock occurs at the edge of a solar system (at around 100 astronomical units for our solar system, where 1 astronomical unit is the distance between the Sun and Earth). There, streams of charged particles being released from a sun (i.e., solar winds) interact with the matter and radiation that exists between star systems in a galaxy (i.e., the interstellar medium). This causes the particles to slow down suddenly. The termination shock is formed as the solar wind goes from super to subsonic speeds causing compression and heating in the plasma.

In a recent project, Moschou and collaborators use a physics-informed neural network (PINN)[7] for modeling the solar termination shock in the presence of a gravitational field (2023). A nice feature of this context is that a simulation code already exists (PLUTO) that effectively models astrophysical gas dynamics. The computational expense of using PLUTO for the full analysis, however, is quite high. Therefore, the authors turn to machine learning, using PLUTO to generate a synthetic database that the PINN is trained on.

The inclusion of physical principles in this context comes with the form of the loss function used. The researchers add two terms to the standard loss function. One of the new terms captures how close the model is to satisfying the partial differential equation thought to describe the system. The other new term corresponds to the initial and boundary conditions.

The PINN is then trained on the synthetic data created by PLUTO and put to work. Generally, the PINN performs very well and models shocks effectively. Additionally, it requires only a fraction of the data a standard NN would (Moschou, Hicks, Parekh, et al. 2023, see §3.2). The capability of the PINN goes beyond merely speeding up computations though, as the authors argue. They claim that the PINN can be "used to discover the underlying physics from data" (Moschou, Hicks, Parekh, et al. 2023, 6). But what do they mean by this? The "underlying physics" learned by the PINN in this context is the prediction of a particular parameter in the PDE being used to model the system. This parameter is known as the effective polytropic index. The polytropic index describes the exchange of energy between a gas and the environment.[8] In this case, the researchers are using an effective polytropic index that "mimics the effects of adding heating in the system" (Moschou, Hicks, Parekh, et al. 2023, 3) but avoids the computational expense of actually calculating the heating and cooling. The "ground truth" value of this parameter is found in the PLUTO simulation and the value predicted by the PINN is then compared to this "ground truth" value. In sum, the discovery of underlying physics by the PINN amounts to being able to predict the value of an effective parameter occurring in the PDE that is used as part of the loss function used to train the PINN.

Despite their success, PINNs still face challenges. One discussed by Moschou and collaborators has to do with the convergence of the PINN. The inclusion of two additional terms in the overall loss function used to train the PINN makes the convergence of the NN trickier. Whereas the convergence of standard NN is well-studied, the convergence of these PINNs with multifaceted loss functions is not. As Moschou and collaborators write, "The neural network optimizer might have to deal with losses that differ by several orders of magnitude which makes the minimization task and reaching a unique solution challenging" (2023, 15).[9] Issues with

[7] I will be using PIML and PINN interchangeably going forward; PINNs are a subset of PIML just as NNs are a subset of ML.

[8] For natural fluids, the polytropic index (γ) falls between 1 and 5/3 (where 5/3 is the index for an ideal gas). Here, they use an effective polytropic index $\gamma < 5/3$.

[9] They note that some (Jin, Cai, Li, and Karniadakis 2021, in particular) have tried to address this issue by using adaptive weighting algorithms to more effectively minimize the loss function and find convergence. Adaptive weighing does introduce additional model hyper-

the convergence of the cost function and questions about what to include in the cost function will arise with the next case study as well.

4.3 Gravitational fields around small bodies

Another context in which PINNs have been applied is to study gravitational fields around small astronomical bodies. Here, I consider a three-part project by Martin and Schaub. They began by first applying these methods to predict the gravitational field around the Earth and Moon (Martin and Schaub 2022b) before turning to small astrophysical bodies (Martin and Schaub 2022a; Martin and Schaub 2023). Their approach goes beyond the previous example of PINN: it alters the network architecture as part of the "physics-informed" aspect of the model. I begin by detailing the problem at issue and then outline the implementation of the PINN.

Suppose you want to take soil samples from the surface of an asteroid. Doing so requires that you land a spacecraft on its surface. Landing a spacecraft on the surface of an asteroid, though, is no easy feat. It requires being able to accurately model the gravitational field around an irregularly-shaped object, a field whose potential is described by Laplace's equation. One common method of modeling the gravitational field around objects like asteroids and comets—of solving Laplace's equation—is to use spherical harmonic functions (see, e.g., Bucha and Sanso 2021). While spherical harmonics may provide an exact, analytic solution to the problem, they are not suitable when one is close to the body of interest (e.g., approaching the surface) because they diverge. Indeed, as noted by Martin and Schaub, the divergence of the spherical harmonics "poses a problem for irregularly shaped objects, like asteroids, where spacecraft may operate at substantially lower radii than the Brillouin sphere" (2022a, 46).[10][11]

Traditional ML, NNs in particular, has been applied to this problem. However, the nature of the problem makes the use of NNs difficult: one would need a large amount of data at various altitudes for each astronomical body to capture any surface irregularities. Additionally, there are various physical constraints the researchers know that could be used to guide solutions to the problem. For instance, any derived gravitational potential must be a solution to Laplace's equation. However, there is no

parameters that have to be set up by the researchers. However, these hyper-parameters have to do with the convergence of the network, not the physical problem at issue. Nonetheless, Moschou and collaborators are not convinced that this method will be sufficient. As evidence, they cite another study (Fuks and Tchelepi 2020) that aimed to solve a non-linear hyperbolic equation. Fuks and Tchelepi explored many NN algorithms and architectures for their PINN but could not find convergence until they switched the equation form to a parabolic one. This leads Moschou and collaborators to believe that no matter what algorithm is being used to minimize the loss function, it cannot "address the fundamental problems in the PINNs optimization procedure" (2023, 15).

[10] The "Brillouin sphere" is the sphere encompassing all the field-generating mass. It is within this sphere that the spherical harmonic functions diverge.

[11] At low altitudes, such as those used for touch and go or landing maneuvers, one can use alternative representations (like the polyhedral gravity representation). However, these have their own issues (e.g., they assume uniform density and are typically computationally expensive).

way of incorporating this kind of knowledge into a traditional NN. As Martin and Schaub put it,

> Traditional neural networks are not trained with these physics properties in mind. Instead, they prioritize predicting an accurate acceleration from a position vector, irrespective of the more fundamental properties. In this sense, the network will be trained agnostic to the fact that the gravity field it represents produces conservative forces, and the underlying potential must be sufficiently smooth and continuous for sensible dynamics. (2022a, 9)

With these kinds of considerations in mind, Martin and Schaub turn to a physics-informed approach. As mentioned above, they begin by applying their methods to the gravitation field around the Earth and Moon (Martin and Schaub 2022b) and then generalize (Martin and Schaub 2022a; Martin and Schaub 2023). They begin with a relatively simple model that is similar to the first study. Their methodology evolves over the course of the three papers though, allowing this to serve as an example of how to incorporate physics-informed principles in the architecture of the NN.

Martin and Schaub begin with minimal physical constraints: they only include one additional term into the PINN loss function. This term corresponds to the equation describing the relationship between the gravitational potential and the acceleration ($\mathbf{a} = -\nabla U$). They limit the modifications they make to the loss function for two reasons. First, further constraints undermine the convergence of the PINN. Second, other constraints would likely not be of the same order-of-magnitude and would require rescaling to be incorporated into the loss function (discussed in more detail below).

In the second paper of the series, Martin and Schaub aim to incorporate further physical principles in the PINN. To do so, they adopt a modified network architecture. This change also helps with the issue regarding the convergence of the loss function mentioned in the previous case study. Instead of a fully connected network (in which every neuron in one layer is connected to every neuron in the previous and the next layer), Martin and Schaub adopt an architecture proposed Wang, Teng, and Perdikaris (2021) that uses recent developments in neural attention mechanisms. Essentially, they introduce transformers that project the inputs into a higher-dimensional feature space. As Wang, Teng, and Perdikaris note, this: "(i) explicitly accounts for multiplicative interactions between different input dimensions, and (ii) enhances the hidden states with residual connections" (2021, A3069). These architectural modifications alleviate issues of convergence that otherwise arise with multi-faceted loss functions.

The final iteration of Martin and Schaub's PINN is presented in (Martin and Schaub 2023). There, they reflect that previous iterations of their algorithm "use a cost function which inadvertently leads models to prioritize low-altitude field points, where the accelerations are largest" (2023, 2). In order to address this issue and various other performance issues, they have again redesigned the architecture of the network, rescaling the output of the NN, changing the loss function, and including more physically-informed principles. I detail some of these changes below. The important takeaway here is the degree of control they have over the components of the PINN architecture and how this enables them to develop understanding.

One of the reasons low-altitude field points were prioritized is because the gravitational field is small at high altitudes, so small that the field values encroach on machine precision. To address this issue and deprioritize the low-altitude predictions, Martin and Schaub rescale the output of the NN before it is used in the loss function (2023, 10). Another effect of the gravitational field values being so much larger at low altitudes than at high altitudes is that the error values are also much larger. Therefore, the standard error used in these contexts (a percentage error) prioritizes low-altitude predictions. To account for this, they switch from using mean percent error to using root mean squared error (2023, 7). Finally, Martin and Schaub re-introduce the spherical harmonic functions for performance, "fus[ing] an a-priori gravity model with the neural network solution" (2023, 9). In terms of the final performance, the first iteration of the PINN already performed better than the spherical harmonic representation. However, it still required a large data set. In comparison, the third generation converges reliably and quickly with a small training set (2023, 13-14).

In sum, this series of three papers introduces and develops a PINN for predicting the gravitational field around small astrophysical bodies. As Martin and Schaub improve their PINN, they increase the number of physical principles incorporated in their PINN, alter the network architecture to better suit the problem at hand, and reintroduce the form of the solution they expect.

5 Discussion

The two case studies presented above are characteristic examples of the two approaches taken in physics-informed machine learning: incorporating physical principles into the training of the ML algorithm and using the network architecture to capture some physical properties of the system. We can now ask whether these algorithms are in fact more interpretable than standard ML methods.

Let us begin with the first case study. One can imagine training a standard NN to learn on the same synthetic data that Moschou and collaborators used to train their PINN. Such an NN would be able to similarly predict the polytropic index of some new data. Now, the question is whether the PINN is in a substantially better position than such an NN.

Most simply, one might argue that, because of the input physics, the improved efficiency of the PINN ought to correspond to an improvement in interpretability. In particular, one might claim that because PINN requires fewer nodes, hidden layers, etc. compared to NN for the same performance, it is easier to interpret. I would argue, however, that when we have tens of thousands of nodes or more, even an order of magnitude fewer does not seem significant to transparency. More importantly, recalling Creel's taxonomy of the types of transparency from §3.1, it is clear that the kind of transparency at stake when considering the number of nodes, the weights, etc. of these algorithms is run transparency. Though run transparency may be important for some purposes, it does not help with the overall interpretability of the network, i.e., its functional transparency.

Do the physical principles used to train the algorithm themselves provide improvements to the functional transparency? While the addition of physical principles improves the performance of the algorithm, I argue that they have no effect on the functional transparency, and thus interpretability, of the algorithm. To understand why, consider the difficulties both teams of researchers encountered when adding terms to the loss functions that represented the physical principles. They worried weighting of the various terms and the convergence (or lack thereof) of the network. If one wanted to argue that the functional transparency is increased with the addition of physical principles, understanding how those principles are being leveraged would be critical. However, the developers of these algorithms do not (yet) have a clear sense of how the algorithm is prioritizing the various components of the loss function. Though improvements are being made with further research, it seems too early to say whether PINNs are leveraging and respecting these physical principles.

However, I do not believe the situation is worrisome. We may not have the transparency required to tell whether PINNs are levering these physical principles in the way we expect, but we also do not need it. To understand this claim, recall the discussion of emulators presented in §3.2. There, a particular step of the methodology was being "black-boxed" when scientists used ML methods like emulators. The first case study of PIML is akin to the use of emulators. Here, like in the case of emulation, a particular step of the method is being "black-boxed." However, one knows the physical principles underlying the problem and, if needed, could use a first-principles-style simulation for the same goal. Put differently, one has an interpretable simulation to fall back on for explanatory purposes. We need not require the ML algorithm to serve explanatory purposes; like emulators, its purpose is to make computations faster and more efficient.

We can also analyze the situation with respect to link certainty. Should we be worried about whether the PINNs are using the "right" link? Since what we are predicting is the value of the effective polytropic index—not some further detail of the underlying physics—, it does not matter whether our ML algorithm is leveraging the right link or not. The PINN is just a statistical tool. And, again, because the background physical principles are well-understood, if needed, we could always revert to more of a first-principles kind of analysis.

In sum, the methodology of the PINN in this first case does not present any novel strides in interpretability/transparency. But this is in part because one need not be worried about the interpretability/transparency of the ML algorithm in the first place. If one treats the original ML algorithms akin to how one treats emulators—as statistical tools—such questions about interpretability do not arise. Of course, there may be other reasons to adopt the PIML methods (e.g., they are better suited to noisy data and more efficient), but these reasons are irrelevant to considerations of the interpretability of the algorithms.

Let us now consider the second case study. I argue that this case study is an example in which the physics-informed nature of the ML helps with its interpretability. This is because 1) the goal is different and 2) the methodology the researchers adopt is different. Martin and Schaub's overall goal is to predict the gravitational field close to an astronomical body. In this case, at least in the first two generations, the PINN is not bound by any particular physics. Though a physical principle is introduced, it is not predicting some particular parameter (like the polytropic index). Instead, it is predicting the gravitational field across the surface of the body. While the spherical-harmonic representation might be the best for analytic calculations, their PINN is not subject to this representation. As Martin and Schaub write,

> Such an approach allows the PINN to efficiently learn custom and physically motivated basis functions which represent the natural features in the gravitational potential of a planetary body (e.g., craters on the Moon or mountain ranges on the Earth) rather than imposing basis functions like spherical harmonics which are inefficient at capturing these idiosyncratic and often discontinuous features. (2022a, 3)

Clearly, this is a different goal from the previous case study. This different goal guides the methodology. Like in the first case study, Martin and Schaub introduce a physical principle in their loss function. However, their iterative methodology—the fact that they consider how the loss function behaves, adjust the architecture to ensure convergence, and later, incorporate further physical principles—raises the likelihood that the PINN is leveraging that loss function and respecting the physical principles incorporated. And, as Martin and Schaub input more physical principles (a low-fidelity analytic model, boundary conditions, adjust the loss function to allow the network to be more sensitive at high-altitudes), the interpretability of the PINN improves further. Some of these changes are of course possible with standard NN models. However, the physics-informed NNs and the architectural flexibility they provide allow the researchers to have a better handle on the performance of the PINN overall.

We can again consider the situation in terms of link certainty and whether the PINN has leveraged the right link. In the previous case, the physical principles did not meaningfully impact the interpretability of the model. However, there, it did not matter whether the PINN leveraged the link we knew to exist between the data and the polytropic index. Here, there is a link between the data and the gravitational field and we want the NN to leverage that link. In adjusting the architecture of the network, our belief that the algorithm is leveraging that link increases. This is because our understanding of the algorithm is increasing as we investigate various components of it and its performance. Thus, the PIML methods, and the ability to iteratively adjust them, do present an improvement in transparency. This is important because, unlike in the first case, we do not have a first-principles style argument to fall back on. Our analytic models fail in the domains that the PIML is being used. To summarize, this second case study offers a case in which we have link certainty and we want functional transparency to ensure our ML algorithm is indeed leveraging that link. PIML methods—the inclusion of physical principles and iterative adjustment of the network architecture—increase our confidence that the ML algorithm is indeed leveraging that link.

6 Conclusion

Is PIML really the "next-generation of artificial intelligence"? Is it as "transformative" as proponents would have us believe? PINN algorithms likely do outperform standard ML techniques by introducing physical principles into their loss functions. In terms of their interpretability, there are many contexts in which we need not be worried about the interpretability of the algorithms in the first place; in such contexts, PINNs seem to be in just the same situation as traditional NN. In cases where greater transparency is desired, PINN methods—ones that adjust the network architecture to reflect physical principles—can offer improvements.

Biography: Helen Meskhidze is a Postdoctoral Fellow at the Black Hole Initiative at Harvard University. She has a Ph.D. in Philosophy from the Logic and Philosophy of Science Department. Her research concerns the epistemology of large-scale simulations and uses of machine learning in astronomy/cosmology. She is also interested in the foundations of spacetime theories.

Acknowledgments: My thanks to Florian Boge and Giorgia Pozzi for their comments on earlier drafts of this chapter.

This work was supported by the Black Hole Initiative, which is funded by grants from the John Templeton Foundation (Grant #62286) and the Gordon and Betty Moore Foundation (Grant GBMF-8273). The opinions expressed in this work are those of the author and do not necessarily reflect the views of these Foundations.

References

Angel, J. R. P., P. Wizinowich, M. Lloyd-Hart, and D. Sandler (1990). Adaptive optics for array telescopes using neural-network techniques. *Nature 348*(6298), 221–224.

Andrzejczuk, D. (2023). Physics Informed Machine Learning. Medium: Technology, Invention, App, and More. https://medium.com/the-quantum-data-center/physics-informed-machine-learning-the-next-generation-of-artificial-intelligence-solving-89ca4bb2e05b.

Beisbart, C. and T. Räz (2022). Philosophy of science at sea: Clarifying the interpretability of machine learning. *Philosophy Compass 17*(6), e12830.

Bertin, E. and S. Arnouts (1996). Sextractor: Software for source extraction. *Astronomy and astrophysics supplement series 117*(2), 393–404.

Boge, F. J. (2023). Functional concept proxies and the actually smart hans problem: What's special about deep neural networks in sciences. *Synthese*.

Boge, F. J. (2022). Two dimensions of opacity and the deep learning predicament. *Minds and Machines 32*(1), 43–75.

Branca, L. and A. Pallottini (2023). Neural networks: solving the chemistry of the interstellar medium. *Monthly Notices of the Royal Astronomical Society 518*(4), 5718–5733.

Brewer, B. J., D. Foreman-Mackey, and D. W. Hogg (2013). Probabilistic catalogs for crowded stellar fields. *The Astronomical Journal 146*(1), 7.

Bucha, B. and F. Sanso (2021). Gravitational field modelling near irregularly shaped bodies using spherical harmonics: a case study for the asteroid (101955) Bennu. *Journal of Geodesy 95*, 1–21.

Cai, S., Z. Wang, F. Fuest, Y. J. Jeon, C. Gray, and G. E. Karniadakis (2021). Flow over an espresso cup: inferring 3-d velocity and pressure fields from tomographic background oriented schlieren via physics-informed neural networks. *Journal of Fluid Mechanics 915*, A102.

Chen, X., D. J. Jeffery, M. Zhong, et al. (2022). Using physics informed neural networks for supernova radiative transfer simulation. *arXiv preprint arXiv:2211.05219*.

Chen, Z., Y. Liu, and H. Sun (2021). Physics-informed learning of governing equations from scarce data. *Nature communications 12*(1), 6136.

Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science 87*(4), 568–589.

Daneker, M., Z. Zhang, G. E. Karniadakis, and L. Lu (2023). Systems biology: Identifiability analysis and parameter identification via systems-biology-informed neural networks. In *Computational Modeling of Signaling Networks*, pp. 87–105. Springer.

Dvorkin, C., S. Mishra-Sharma, B. Nord, et al. (2022). Machine learning and cosmology. *arXiv preprint arXiv:2203.08056*.

Fuks, O. and H. A. Tchelepi (2020). Limitations of physics informed machine learning for nonlinear two-phase transport in porous media. *Journal of Machine Learning for Modeling and Computing 1*(1).

Heitmann, K., D. Higdon, M. White, et al. (2009). The Coyote Universe. II. Cosmological Models and Precision Emulation of the Nonlinear Matter Power Spectrum. *The Astrophysical Journal 705*, 156–174.

Hy, T. S., S. Trivedi, H. Pan, B. M. Anderson, and R. Kondor (2018). Predicting molecular properties with covariant compositional networks. *The Journal of chemical physics 148*(24).

Jin, X., S. Cai, H. Li, and G. E. Karniadakis (2021). NSFnets (Navier-Stokes flow nets): Physics-informed neural networks for the incompressible navier-stokes equations. *Journal of Computational Physics 426*, 109951.

Jin, Y., L. Yang, and C.-E. Chiang (2022). Identifying exoplanets with machine learning methods: a preliminary study. *International Journal on Cybernetics & Informatics 11.*

Karniadakis, G. E., I. G. Kevrekidis, L. Lu, et al. (2021). Physics-informed machine learning. *Nature Reviews Physics 3*(6), 422–440.

Kashinath, K., M. Mustafa, A. Albert, et al. (2021). Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A 379*(2194), 20200093.

Li, X., Ragosta, F., Clarkson, W.I. and Bianco, F.B. (2021). Preparing to Discover the Unknown with Rubin LSST: Time Domain. *The Astrophysical Journal Supplement Series 258*(1), 2.

Lin, H., X. Li, and Z. Luo (2020). Pulsars detection by machine learning with very few features. *Monthly Notices of the Royal Astronomical Society 493*(2), 1842–1854.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue 16*(3), 31–57.

Martin, J. and H. Schaub (2022a). Physics-informed neural networks for gravity field modeling of small bodies. *Celestial Mechanics and Dynamical Astronomy 134*(46), 1–28.

Martin, J. and H. Schaub (2022b). Physics-informed neural networks for gravity field modeling of the earth and moon. *Celestial Mechanics and Dynamical Astronomy 134*(2), 13.

Martin, J. and H. Schaub (2023). The physics-informed neural network gravity model revisited: model generation iii. In *33rd AAS/AIAA Space Flight Mechanics Meeting, Austin, United States.*

Meetre, C. and J. Norris (1991). Recognition of gamma-ray events in egret using neural networks. In *Bulletin of the American Astronomical Society, Vol. 23, p. 905*, Volume 23, pp. 905.

Meskhidze, H. (2023). Can machine learning provide understanding? how cosmologists use machine learning to understand observations of the universe. *Erkenntnis 88*(5), 1895–1909.

Metcalf, R. B., M. Meneghetti, C. Avestruz, et al. (2019). The strong gravitational lens finding challenge. *Astronomy & Astrophysics 625*, A119.

Miller, A. (1993). A review of neural network applications in astronomy. *Vistas in Astronomy 36*, 141–161.

Mishra, S. and R. Molinaro (2021). Physics informed neural networks for simulating radiative transfer. *Journal of Quantitative Spectroscopy and Radiative Transfer 270*, 107705.

Moschou, S., E. Hicks, R. Parekh, et al. (2023). Physics-informed neural networks for modeling astrophysical shocks. *Machine Learning: Science and Technology 4*(3), 035032.

Narayan, G., T. Zaidi, M. D. Soraisam, et al. (2018). Machine-learning-based Brokers for Real-time Classification of the LSST Alert Stream. *Astrophysical Journal Supplement Series 236*(1), 9.

Pfau, D., J. S. Spencer, A. G. Matthews, and W. M. C. Foulkes (2020). Ab initio solution of the many-electron schrödinger equation with deep neural networks. *Physical Review Research 2*(3), 033429.

Raissi, M., P. Perdikaris, and G. E. Karniadakis (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics 378*, 686–707.

Shukla, K., P. C. Di Leoni, J. Blackshire, D. Sparkman, and G. E. Karniadakis (2020). Physics-informed neural network for ultrasound nondestructive quantification of surface breaking cracks. *Journal of Nondestructive Evaluation 39*(3), 1–20.

Stein, G. (2023). https://github.com/georgestein/ml-in-cosmology.

Stewart, M. (2019). https://towardsdatascience.com/simple-introduction-to-convolutional-neural-networks-cdf8d3077bac.

Sullivan, E. (2022). Understanding from Machine Learning Models. *The British Journal for the Philosophy of Science*.

Villaescusa-Navarro, F., J. Ding, S. Genel, et al. (2022). Cosmology with one galaxy? *The Astrophysical Journal 929*(2), 132.

Wang, S., Y. Teng, and P. Perdikaris (2021). Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM Journal on Scientific Computing 43*(5), A3055–A3081.