Causal scientific explanations from Machine Learning

Stefan Buijsman*

Accepted for publication in *Synthese*

Abstract

Machine learning is used more and more in scientific contexts, from the recent breakthroughs with AlphaFold2 in protein fold prediction to the use of ML in parametrization for large climate/astronomy models. Yet it is unclear whether we can obtain scientific explanations from such models. I argue that when machine learning is used to conduct causal inference we can give a new positive answer to this question. However, these ML models are purpose-built models and there are technical results showing that standard machine learning models cannot be used for the same type of causal inference. Instead, there is a pathway to causal explanations from predictive ML models through new explainability techniques; specifically, new methods to extract structural equation models from such ML models. The extracted models are likely to suffer from issues though: they will often fail to account for confounders and colliders, as well as deliver simply incorrect causal graphs due to ML models tendency to violate physical laws such as the conservation of energy. In this case, extracted graphs are a starting point for new explanations, but predictive accuracy is no guarantee for good explanations.

1 Introduction

Machine learning models[1] are quickly gaining ground in scientific practice. A particular success is the use of deep learning model AlphaFold 2 to predict protein folding (Jumper et al., 2021), but examples abound. There is, for example, usage of deep learning in climate models (Rasp et al., 2018), astronomy (Agarwal et al., 2012), and materials science (Schmidt et al., 2019). These machine learning models are primarily predictive models: they are used because they can very accurately predict outcomes (e.g. protein folds, or climate parameters). An important exception is the case of neuroscience(Milkowski, 2013; Stinson, 2018; Piccinini, 2010) because of the direct modeling offered by machine learning models

*TU Delft, Jaffalaan 5, 2628BX, Delft, The Netherlands. Email: s.n.r.buijsman@tudelft.nl

[1]It should be noted here that the term 'machine learning' has both narrow and broad interpretations. In a broad interpretation it is any computer method that solves a problem by fitting a function to data. In that case, simple models such as those based on linear regression count as machine learning. I follow the narrower definition of machine learning common in the literature discussed here, where the term is only applied to methods such as deep neural networks and random forest algorithms, which are distinguished by their use of a large number of parameters and non-linearity.

in that area. I will set aside that field here to focus on the other areas of science. There, it is not clear that opaque machine learning models, which predict well but where we do not understand why they arrive at a certain prediction (Das and Rad, 2020) can be used to arrive at scientific explanations (López-Rubio and Ratti, 2021; Srećković et al., 2021). The most advanced machine learning tools, deep neural networks, present us with two difficulties: they are very complex (easily containing millions of parameters that are fine-tuned based on large data sets) and they lack internal representations that are legible to us (other models have explicit variables standing for e.g. physical quantities, deep neural networks have activation functions that respond to complex combinations of input features which we cannot interpret). As a result, it is difficult to see how we can acquire causal explanations when using these complex models. There are worries that "if you do molecular biology with machine learning techniques, and if you want to have the best machine learning performances, then you cannot even in principle elaborate fully-fledged mechanistic explanations." (López-Rubio and Ratti, 2021, p.3152)

Recently, Sullivan (2019), Knüsel and Baumberger (2020), Jebeile et al. (2021) and Meskhidze (2021) have argued that there is in fact a possibility to get more from these models than just predictions. They focus precisely on the predictive machine learning models just mentioned, and follow the idea that under certain conditions (primarily that link uncertainty is low, meaning that there is "scientific and empirical evidence supporting the link connecting the model to the target phenomenon" (Sullivan, 2019, p.30)) this is possible without the models being explainable. This approach has been extensively criticized in a recent paper by Räz and Beisbart (2022). I agree with those criticisms, but will leave that debate to the side for the current paper. Instead, my goal here is to offer a different positive answer to the question whether we can arrive at scientific explanations from machine learning models.

First, I argue in sections 2 and 3 that machine learning models that are used in causal inference can be seen as generating scientific explanations, at the very least if we take a causal account of scientific explanation (Halpern and Pearl, 2005; Woodward, 2005). On such an account, we can explain a target phenomenon by giving a cause for that phenomenon, which will cover both the actual case and answer a range of what-if-things-had-been-different questions. The general approach of causal inference (described in section 2) aims to determine such causal relations, and if successful, gives us the cause(s) for the target phenomenon. My argument, then, is that if machine learning models can provide us with causal relations for scientific phenomena, then they can thus provide us with (causal) scientific explanations. After all, causal relations are precisely what we should be providing according to Woodward (2005) and others in order to explain scientific phenomena. The challenge thus is to show that machine learning models can, in some cases, provide these causal relations. In this, I extend the work of Pietsch (2016) who argued that causal modeling is possible with big data. The added contribution here is a close connection to causal inference techniques and a link to specific ML models

that can be used for causal inference.

I follow the influential work of Pearl (2009) to structure the discussion on causal inference, and whether machine learning can be used to arrive at causal explanations of scientific phenomena. On his framework causal inference happens in essentially two steps: first, a directed acyclic causal graph (or, equivalently, structural equations) is formulated for the specific case. This specifies which variables influence which other variables, and is crucial for the causal inference process. Without a causal graph in place, causal inference is not possible, and with different causal graphs the same data will give different results in the second step. As such, these assumptions are crucial (and are often verified afterwards by seeing how results respond to changes to the causal graph that should not have an effect, such as introducing a random extra cause). The second step is to estimate the treatment effect(s) in the causal graph. This comes down to estimating the strength along the arrows in the causal graph, or the coefficients in the structural equations and effectively shows what the exact causal influence is of a cause on the target phenomenon.

Interestingly, both causal inference steps can be accomplished using machine learning models, and even using deep neural networks. Causal graphs can be learned from data with the help of machine learning (section 2) and specific double machine learning techniques allow the calculation of treatment effects using neural networks when given a causal graph (section 3). Together, this shows (so I argue) that there are definitely some machine learning models that can provide causal explanations. In section 4 I then expand on these results to consider the prospects of scientific explanations from predictive ML models. There, I argue that recent techniques to extract structural equation models from deep neural networks give us good reason to think that through better explainability of such models we can get (candidate) scientific explanations. However, these predictive models lack some of the properties of the ML models considered in sections 2 and 3 and so these candidate explanations are likely to be incorrect when adopted directly. Still, they can provide a good starting point for new causal scientific explanations and thus explainable AI can help us extract explanations from predictive ML models as well.

2 Learning causal graphs

Causal inference requires assumptions about the direction in which variables influence each other. Where correlations go both ways and can be directly calculated from observational data, causal relations are not so easily inferred. After all, there are plenty of correlations that do not correspond to causal relations (take the example of a correlation between drownings and ice cream sales, which is the result of a common cause: nice weather). To arrive at causes, correlations thus fall short. As Pearl

says: "Causal analysis goes one step further; its aim is to infer not only beliefs or probabilities under static conditions, but also the dynamics of beliefs under changing conditions, for example, changes induced by treatments or external interventions." (Pearl, 2009, p.99) In order to get to these causal relations assumptions need to be made. Again: "behind any causal conclusion there must be some causal assumption, untested in observational studies." (Pearl, 2009, p.100) These assumptions are the causal graphs, or (equivalently) structural equations that describe our choices about the direction of causation. In other words, these graphs/models describe whether X causes Y, or Y causes X (or that the two are independent). They also tell us what variables are confounders (variables that influence both the outcome and the treatment) and which are colliders (variables influenced by both the outcome and the treatment), and how we can control for these to arrive at correct estimations of the strength of causal relations (the treatment effect). An example causal graph with a confounder X for treatment T and outcome Y is shown in figure 1.

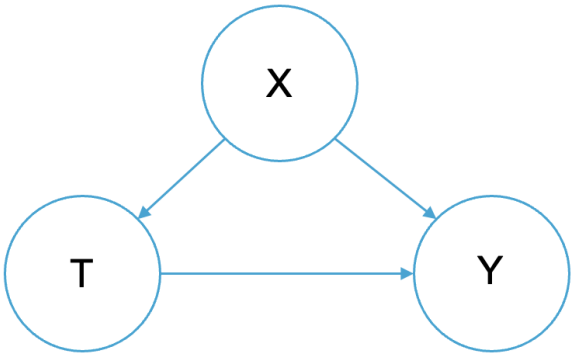

Figure 1: A simple causal graph with a confounder X, treatment T and outcome Y. Change the direction of both arrows from X and it becomes a collider.

How does machine learning come in here? Well, it turns out that to some extent causal graphs can be learned from observational data (for a review see Glymour et al. (2019)). And (purpose-built) machine learning models are one way to do so. To see how, I'll discuss two methods to learn causal graphs: constraint-based methods and scoring methods. For both I show how machine learning can be applied and how they yield new causal graphs.

First, one can follow a constraint-based method (cf. Spirtes et al., 2000): starting with a maximally connected graph one estimates conditional independencies between variables to remove edges (one direction at a time) from that graph. Ideally this results in a graph where enough edges are removed to fill in the direction of causal inference. This happens using two inference rules:

1. "For each triple of variables (A, B, C) such that A and B are adjacent, B and C are adjacent,

and A and C are not adjacent, orient the edges A – B – C as A → B ← C, if B was not in the set conditioning on which A and C became independent and the edge between them was accordingly eliminated. We call such a triple of variables a *v-structure*" (Glymour et al., 2019, p.4)

2. "For each triple of variables such that A → B – C, and A and C are not adjacent, orient the edge B – C as B → C. This is called *orientation propagation*." (Glymour et al., 2019, p.4)

This will not always allow the learning of the full causal graph. There is, in other words, no guarantee that all edges will be covered by these two inference rules. As such, conditional independence algorithms at best converge to the Markov Equivalence class (assuming that the conditional independence estimates are correct and there are no unmeasured confounders) where some edges can go either way. Learning causal graphs in this manner is thus both highly dependent on how conditional independence is estimated (importantly Shah and Peters (2020) have shown that there exists no uniformly valid test for conditional independence) and will not always produce a full causal graph. This nicely illustrates the difficulty of learning causal graphs from data, and is in line with Pearl's quote that there are always some untested assumptions involved. In this case, those causal assumptions will be that the particular conditional independence test used on the data allows us to conclude *causal* independence of the two variables and is possibly supplemented by additional causal assumptions to complete the causal graph. For example, in the situation where there are only two variables there is no way to determine whether A causes B, or B causes A using independence testing. It is possible to observe whether the variables influence each other, but without actually changing one variable and seeing the effect of this intervention we cannot decide this situation. In practice we of course have access to many more variables than just the two, but situations will still occur where the direction of certain edges cannot be decided using a constraint-based method. Then, additional causal assumptions have to be made.

To return to the focus of this paper, conditional independence tests can be performed with the help of machine learning models. In this case, the method as reviewed by Glymour et al. (2019) is thus executed using machine learning and thus gives us an instance of a causal graph learned using machine learning. For example, Sen et al. (2017) do so by turning the question of conditional independence testing into a classification problem, which machine learning models can handle. Their basic idea is that if two variables are conditionally independent, then a classifier (such as a machine learning model, including deep neural networks) will have a hard time to predict one variable based on the other. If, on the other hand, there is some causal dependence there, then it will be possible to predict the values of the first variable using the values of the second variable. Note, however, that this requires a purpose-built machine learning model and that standard predictive models cannot be used for conditional independence tests. Specifically, Sen et al. (2017) start with $3n$ independent and

identically distributed samples from three variables X, Y and Z. They then process $2n$ samples in a nearest-neighbour bootstrap algorithm to generate n samples that are close to how the three variables would be distributed if X were independent from Y, given Z. These new samples are labelled with a 1, the original n samples that were not used in the bootstrapping are labelled with a 0. Finally, a machine learning model is trained on a mixture of the two sets of samples to predict whether it was an original sample (labelled with 1) or a generated sample (labelled with 0). If the machine learning model has a high accuracy, then the original samples are distinguishable from the conditionally independent samples and so it's likely that X and Y are not independent. If the model performs (close to) randomly, then it is likely that X and Y are conditionally independent. As is clear, this requires quite a bit of work and a specifically designed machine learning model. This is a feature that is repeated in other approaches found in the literature: e.g. Bellot and van der Schaar (2019) and Shi et al. (2020) both use specific generative adversarial networks (a type of deep neural network/machine learning model) to approximate conditional distributions and test independence hypotheses on that basis. In short, conditional independence testing with machine learning models is possible, but requires purpose-built models.

It is thus possible to use machine learning for causal discovery (i.e. learning causal graphs from data) when taking a constraint-based approach. A similar possibility for using machine learning is seen for the other method, a scoring approach. Instead of estimating conditional independence, this method uses a scoring function to search for the causal graph that best fits the observational data (where the choice of scoring function contains causal assumptions). This effectively comes down to an optimization problem (learning the graph that maximizes fit with the data) and so is well-suited for the use of machine learning. And indeed, there is a good number of approaches to do so, of which I'll highlight two here to give an impression of the possibilities. As we'll see, for all these approaches there is again an important causal assumption: that a graph more similar to the actual causal graph will get a higher score than one that is less similar. This assumption, that the scoring function tracks causality, is a rather uncertain one, as scoring methods frequently give us incorrect causal graphs. Still, it can be a useful method when arriving at a causal graph in other ways is difficult.

Lachapelle et al. (2019) propose a method that uses the inner workings of small neural networks to learn causal graphs. The basic idea is that they train fully-connected neural networks to predict each variable of interest, based on all the other variables that are considered. If variable X is not used to predict the value of Y, then that is evidence that there is no arrow from X to Y. If X is used to predict Y, then that is evidence for an arrow from X to Y. In a little more detail: this method starts with the fitted neural networks and takes the sum of all path products from input (variable) i to output (variable) k. That's feasible, since the neural networks are small: e.g. when there are two

hidden layers than there are three weights on any path from i to k, one from the input to hidden layer 1, one from hidden layer 1 to hidden layer 2 and a final one from hidden layer 2 to the output. If every path from i to k has a product of the weights along that path that equals zero, then the two are independent (i.e. there is no arrow between them in the causal graph). In practice, a threshold of $\epsilon = 10^{-4}$ is applied during optimization, where any weights are permanently set to zero if they fall below the threshold. The non-zero sums result in a matrix that then needs to be adjusted such that the corresponding graph (with an arrow from i to j iff matrix entry $A_{ij} > 0$) is acyclic. That is where the scoring function comes in, and a maximum likelihood optimzation problem is formulated using the acyclicality constraint, which is solved to provide the best fitting causal graph.[2] As such, the method produces a unique causal graph, but without strong guarantees that it is the correct graph. For example, the method has to employ various strategies to avoid overfitting the neural networks to the data to obtain more plausible graphs. Lachapelle et al. (2019) note that adding edges can never reduce the maximal likelihood, and it is therefore necessary to stop training as soon as the performance no longer increases, as well as to prune the graph after training and to do a preliminary selection for any neural networks with more than 50 nodes. You get a unique, completely filled in, causal graph at the end but there are no guarantees that any of the edges are correct.

The other approach to highlight has similar benefits and drawbacks. Kalainathan et al. (2018) present a method to learn causal graphs using generative adversarial networks (a type of neural network), and produce unique causal graphs – but with a good number of false positive and false negative arrows in the resulting graphs. Still, the approach is interestingly different: instead of training shallow neural networks for all the variables at the same time it uses a single, second, machine learning model to score a range of neural networks representing alternative causal graphs. First, different machine learning models are trained to generate the data distribution (predict the values of one variable) based on constraints in line with alternative causal graphs. A model only uses those variables to predict Y that have a path leading to Y in the candidate causal graph. So, a causal graph with an arrow from X to Z but no arrow from Y to Z will be represented by a neural network that aims to predict/generate values for Z based solely on the values of X. Had there been arrows from both X and Y in the causal graph, then the values for Z would be generated based on the values of both X and Y. By creating a large number of neural networks, each corresponding to different causal graphs, it then becomes possible to evaluate which causal graph best fits the actual situation. To do so, a second machine learning model is trained to distinguish between the true data and the data generated by the various

[2]To be precise, the method looks at neural networks j for each variable X (indexed 1 to d). The parameters (i.e. weights) of these neural networks are represented in vector $\phi_{(j)}$. The maximum likelihood problem solved over all these neural networks is then the equation $max_{\phi}\mathbb{E}_{X\ P_X}\sum_{j=1}^{d}\log p_j(X_j|X_{\pi_j^{\phi}};\phi_{(j)})$, where $X_{\pi_j^{\phi}}$ is the set of parents of node j in graph $\mathcal{G}_{\phi}$. Essentially, the idea is that one optimized the predictive accuracy of all these neural networks together, where each neural network aims to predict the value of variable X_j in terms of the values of all the other variables.

neural networks. Whatever generated data is discriminated worst from the true data (i.e. produces the most realistic results) is then considered to be based on the true causal graph. Machine learning does all the work here. At the same time, the results are imperfect. There is no guarantee that a good score in mimicking the data distribution is due to a stronger similarity to the causal graph, and in experimental findings a wide range of causal graphs is found for the same dataset. For example, for the true arrow $x_1 \rightarrow x_2$ the method only correctly identified this edge in 54 out of 100 runs, and included the arrow in the opposite direction in 35 runs. Whether edges are correctly identified varied greatly; some were found in 92 of the 100 runs, others in only 4, and finally one incorrect edge was included in as many as 95 runs. The method is sensitive to how the weights are initialized, but even then it performs better than many other causal discovery methods that do not use neural networks. Still, the variation in causal graphs identified as well as the sometimes very confident false positives show the difficulty of learning causal graphs from data. One misses the essential element of interventions in the world that is the only sure way to identify causes, and so where the constraint-based method would fall short in identifying some edges the scoring-based method will make mistakes in its push to always return a fully filled-in, unique, causal graph.

How do all these machine learning models fit in with the overarching question of whether machine learning can be used to arrive at new scientific explanations? In all of these cases, the machine learning models are crucial to make the step from a situation/data set where we do not know what the causal relations might be to a causal graph that maps out (possible) causal relations between the real-world variables in the data. This means that the result of using the machine learning models is a causal graph, which we didn't have before, that describes which causal relations there are in relation to the target phenomenon. As such, the machine learning models really take centre stage in the causal discovery process (in virtue of them we get the causal graphs) and thus in the mapping of causal relations, which on the causal accounts of explanations entails that they also provide scientific explanations. Moreover, these are how-actually explanations (as opposed to how-possibly explanations) as they are, if all goes well, the causal relations that are present in the actual world captured by the data. The only caveat is that in all of the methods described in this section we get causal graphs with as of yet unknown causal strengths. We may learn that X causes Y, but we do not know how big the influence of X is on Y. To determine that, we need to calculate treatment effects, to which I turn next.

3 Calculating treatment effects

Given a causal graph (or an equivalent structural causal model) from the causal discovery phase it is possible to estimate the strength of causal relations. Typically this is referred to as calculat-

ing/estimating the (average) treatment effect, as the change of the value of the cause is called treatment – and we want to know the effect of that change. And so put very simply, we can take a linear scenario where this treatment effect is simply the coefficient θ regulating the effect of treatment T on output y in the equation:

$$y = \theta T + g(X) + U$$

Where X is the set of variables that also influence y, and which thus need to be controlled for when estimating the strength of the causal relation between T and y. This formulation also conforms well to what standard predictive machine learning models do (in a more general setting where we take $y = g(T, X) + U$ to not assume linearity): they give the output as a function of the different variables at play. However, simply estimating the treatment effect based on this single equation will not work: it neglects the effect that variables X can have on treatment T. As Chernozhukov et al. (2017, 2018) discuss in detail, direct estimations of treatment effects (i.e. based on standard predictive machine learning models) are systematically biased precisely because they miss this extra effect. So, it isn't possible to correctly estimate the strength of causal relations using predictive machine learning models.

To illustrate that problem, it helps to look at an example of one such machine learning model. Caruana et al. (2015) report on a machine learning model that predicted the risk of death for patients entering the hospital with pneumonia. Curiously, this model assigns a lower risk score to patients with asthma than to patients without asthma. It does so for the simple reason that in the data that the model learned from, patients with both asthma and pneumonia are immediately sent to the intensive care, and therefore receive more intensive medical treatment than patients without asthma. That extra confounder (the level of medical care) biases the estimation of the causal effect (from asthma to probability of dying from pneumonia). As such, the predictive model cannot be used to estimate the strength of the underlying causal relations, because it isn't possible to control for the effects of other variables on the causal relation we're interested in. What is needed instead is *double* machine learning, as developed by Chernozhukov et al. (2017).

Their basic idea is that there are two equations that need to be taken into account (keeping it linear for the sake of simplicity, but note that in general the equations are non-linear to take full advantage of the machine learning techniques used):

$$y = \theta T + g(X) + E_1$$

$$T = m(X) + E_2$$

Where the second equation gives us the effect of the other variables on the treatment. If we look back

at figure 1, that means that θ covers the arrow from T to y, $g(X)$ the arrow from X to y and $m(X)$ the arrow from X to T that was previously missing. By solving two prediction problems it becomes possible, then, to arrive at an unbiased estimation of the treatment effect. The first step here is to fit a machine learning model to estimate $m(X)$ in the bottom equation, and thus get an estimate of residual $\hat{W} = T - m(X)$. Second, a different machine learning model is fitted to $g(X)$ in the top equation, giving the residual $\hat{V} = y - \hat{g}(X)$. Finally, the residuals $\hat{V}$ are regressed on $\hat{W}$ to remove the bias of the effect from X to T, and yielding the 'debiased' treatment effect θ. A last bit of bias removal is done by ensuring that the machine learning models are fit on different data than the residual regression, to prevent overfitting should the error terms E_1 and E_2 be correlated (Chernozhukov et al., 2017).

While this involves some extra work, double machine learning has shown to be of value when conducting causal inference, thus providing a clear case where machine learning models can help us arrive at causal relations and with that causal explanations. For again it is really the machine learning that does the work of inferring the treatment effect. Yes, we need a causal graph that is specified beforehand to know which equations the machine learning models should be fitted to (supplying the causal assumptions), but this is no different than standard practice in causal inference (which is also split up into these two phases). We still go, thanks to machine learning, from a situation where we have a causal graph with unknown treatment effects, to one where we have a causal graph with estimated treatment effects. Moreover, machine learning (and especially deep neural networks) can have added value to the process, as Baiardi and Naghi (2020) nicely discuss in the context of econometric research. They highlight that the flexibility of complex machine learning models (in terms of what functions they can approximate) is helpful when estimating treatment effects in situations with complex interactions between variables. ML models also allow for the inclusion of a large number of covariates, where other methods are often more limited. And, machine learning models are helpful for estimating heterogeneous treatment effects, where the strength of the causal relation varies depending on other variables. In short, the nonlinearity of machine learning models is a helpful feature for these estimations of treatment effects and makes them more robust.

As a final example to show that we do acquire causal scientific explanations at the end of this process of inferring a causal graph and estimating treatment effects, consider the work done in Cao et al. (2022). They conducted a study on the use of chemical dispersants to tackle oil spills, and the effect of the salinity of the water on their effectiveness. They start out with a causal graph (drawn in figure 2) and used double machine learning to estimate the strenghts of the various causal relations. As can be seen, this calculation directly explains why oil degrades in the presence of dispersants: the dispersants increase cell abundance and productivity, after which it is these extra, and more productive, cells that speed up oil degradation. This causal graph doesn't show effects of salinity, but that is merely

because it shows the *average* treatment effect. A calculation of heterogeneous treatment effects shows that "dispersant addition had negative effects on oil biodegradation ratio under low salinities, but positive effects under high salinities. On the other hand, the effects of dispersant addition on oil biodegradation increased along with the salinity rise. The results indicated that dispersant addition could alleviate the negative impact on oil biodegradation caused by salinity increasement." (Cao et al., 2022, p.7) A clear (how-actually) explanation of the degradation of oil, and one arrived at thanks to a machine learning model.

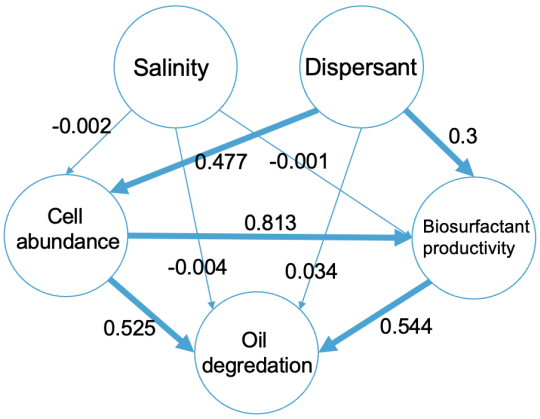

Figure 2: Reproduction of the results in Cao et al. (2022), showing the estimated strengths of causal relations obtained through double machine learning

To sum up, we see that machine learning models can be used arrive at a (causal, how-actually) scientific explanation. Both causal inference steps can be conducted using machine learning, meaning that we can go from a situation where we have merely observations of a target phenomenon to one where we have identified possible causes, confounders and colliders (causal discovery) and where we have quantified what effect these have on the phenomenon in question (estimated treatment effects) using nothing but machine learning and the right data. Of course, in practice we will have to evaluate the causal graph and we will have to be careful about the data we use, but that is no different for other methods used to find the causes for a phenomenon. The important point is that it is, for a specific set of machine learning models, clearly possible to use them to arrive at causal relations and thus at scientific explanations (if something like the causal account is correct).

4 Prospects for explanations from other ML models

The above result is very specific to a class of especially designed ML models. So, what can we expect from the more standard predictive ML models? Can we get scientific explanations from them? As

mentioned in the introduction, a range of authors has answered this with a cautious yes, often giving low link uncertainty as the criterion for acquiring such explanations. If we re-evaluate these claims in the light of whether they produce clearly new causal scientific explanations, then it becomes clear that low link uncertainty on its own does not automatically give us new causal explanations. As mentioned, Sullivan (2019) has been thoroughly criticized by Räz and Beisbart (2022) and so I won't discuss those cases further. Instead, I will look at Meskhidze (2021) as a more detailed example arguing for the ability to get scientific explanations from ML models.

Her example involves a machine learning model used in cosmology. She discusses PkANN, a machine learning model that predicts the large scale distribution of matter, using as its training data the outputs of a large number of runs of (physics-based) N-body simulations. The machine learning model emulates these N-body simulations, and does so with good accuracy (less than 1% deviation from the N-body simulation outputs) at a much lower computational cost. The setup is thus one where a machine learning model has been trained on a physics-based model, and is used in practice primarily for performance-based reasons. Meskhidze (2021) then proceeds to discuss what explanations can be obtained from this machine learning model in detail. She relates this in particular to the distinction Batterman (1992) makes between type (i) and type (ii) why questions: type (i) why questions ask why a phenomenon occurred in some particular circumstance while type (ii) why questions ask why phenomena of this general type occur across a variety of circumstances.

To start with, she sees limitations on what one can expect: "cosmological N-body simulations can answer Batterman's type (ii) why questions: why phenomena of this general type occur across a variety of circumstances. As minimal models of structure formation, they allow one to abstract away from any details of particular cosmological models, and, in doing so, reveal patterns evident across various instantiations of cosmological parameters. Machine learning algorithms exploit these patterns. This means, however, that machine learning algorithms cannot answer Batterman's type (ii) why questions." (Meskhidze, 2021, p.12) In other words, the lack of physical representations and coherence with physical laws is considered problematic to the acquisition of certain explanations. This is typical of ML models trained in this way. Kawamleh (2021) gives a nice discussion of an ML model predicting climate change which violates conservation of energy laws (especially outside its training domain) and argues this is going to be the case for all similar ML models because they all lack physical representations. And, indeed, Meskhidze (2021) observes the same thing for PkANN. However, she argues that another type of why question can be answered by the machine learning model: "machine learning algorithms can answer type (i) why questions: why, for example, our universe has the particular distribution of matter it does. By filling out the parameter space of interest, such methods can point cosmologists to the relevant values of the cosmological parameters that led to a particular distribution of matter."

(Meskhidze, 2021, p.12) As a machine learning model can supply particular values with high accuracy, the idea is that it can therefore contribute to these 'how-actually' explanatory questions.

Yet, are these particular values enough for an answer to the why question? If we look at causal accounts of explanation, then something is missing. For Woodward (2005) a scientific explanation consists of both the actual values *and* a generalization covering counterfactual cases. Similarly, for mechanistic accounts one needs both a causal mechanism (sketch) and particular values to answer these why questions. In the cosmology case we do, of course, have access to a covering rule (causal relation) and mechanism sketch, in the form of the physical laws used in N-body simulations. So, we can get the type (i) explanations by combining the correct values with these physics-based models. But isn't the physics-based model doing the explaining in that case? It seems that PkANN can make finding values more efficient, but that it is not the machine learning model itself that supplies the causal explanation. After all, the physics-based model is needed to meet the requirements of causal accounts of scientific explanations. To make the same point in a different way: had we just had PkANN then we would not have been able to supply a causal explanation, because PkANN does not give us a causal mechanism sketch or the kind of generalization that Woodward (2005) is asking for. PkANN just like other predictive ML models, does not account for confounding variables. Nor do we know what function PkANN is approximating (Das and Rad, 2020), which further complicates the use of just PkANN to provide a (causal) scientific explanation. As a result, the machine learning model plays a very minor role here, as essentially a more convenient way to calculate the values figuring in an explanation[3] that still employs the physics-based model. Without a clearer path to getting a causal explanation out of the ML model on its own I'm skeptical that this kind of case gives us an example of how ML can be used to acquire (causal) scientific explanations.

Rather than looking at ML models in terms of accuracy compared to training data and physics-based models (as Knüsel and Baumberger (2020) and Jebeile et al. (2021) do in a similar way to Meskhidze (2021)), I find it more promising to look at the prospects of explainable AI techniques. If we better understand why an ML system provides us with a certain prediction then we might use that insight into the ML models to acquire explanations. Explainable AI has theoretically been linked to causal scientific explanations (Beckers, 2022; Buijsman, 2022) and there have recently been some attempts in explainable AI to extract structural equation models from standard (deep) neural networks that naturally link to the question central to this paper. Biswas et al. (2022) is one example, where a structural equation model is created using variables representing familiar properties in an attempt to mirror the behaviour of the ML system. More interesting here are the techniques aiming to abstract

[3]Note that these are not, as in section 3, values of treatment effects but rather are values of variables figuring in the explanation. The causal graph thus remains the same, it is only instantiated in a particular way based on the outcomes of PkANN.

a set of structural equation models from (deep) neural networks (Geiger et al., 2021, 2023; Wu et al., 2023) based on theoretical work on abstracting causal models by Beckers and Halpern (2019). These explainable AI techniques aim to abstract a set of structural equations (often displayed in graph form, to mirror the graph representation of the neural network they start with) from the much larger set of equations that make up the neural network. The idea is that we can start with the neural network as a directed acyclic graph, and in fact one that meets all the requirements of a structural equation model. However, it lacks representations and has so many nodes that some abstraction is needed. So, the method aims to find a smaller, interpreted, structural equation model that is a more abstract version (in the way defined by Beckers and Halpern (2019) in terms of there being a translation function τ that maps the effects of interventions from the less onto the more abstract causal model) of the neural network. Geiger et al. (2021) do this by first formulating a hypothesis structural equation model, whose more abstract representations are then linked to parts of the neural network. The abstract model is then verified through the performance of interventions on the abstract model, the effects of which are checked against their effect on the neural network itself. Wu et al. (2023) applied this technique to Alpaca, a large language model with 7 billion parameters. For a specific task consisting of an instruction ('Please say yes only if it costs between [X.XX] and [X.XX] dollars, otherwise no.') and then a price (e.g. '3.50 dollars') they managed to formulate simple structural equation models consisting of at most two additional nodes. Despite this simplicity these models correctly captured the behaviour of Alpaca on the task in 85% of all test cases, additionally generalizing well to variations of the instruction (different price boundaries). So, while there is undoubtedly a lot of work still to do in developing this method, there are some promising first steps towards techniques that can extract manageable structural equation models from (very large) deep neural networks.

As a cautionary note before moving on, the authors of these papers typically present the resulting structural equation models as causal models. I am hesitant to directly speak of causation in the setting where the output of an ML model is 'caused' by the input of the ML model. ML models are best seen as software and these are standardly classified as abstract objects (Turner, 2011; Duncan, 2017), so applying talk of causation directly to the ML model (as opposed to a particular implementation on a physical machine) is somewhat problematic. There is counterfactual dependence though, and structured in such a way that it's possible to use the techniques of causal inference. Thanks to that the resulting models are structural equation models using physical variables (e.g. a temperature variable as the outcome and variables for CO_2 and amount of sea ice as initial variables) and can therefore be interpreted as candidate causal models for the physical world, separately from their link to an ML model.

This suggests an easy route towards (causal) scientific explanations from predictive ML models:

we need to extract/abstract a structural equation model from the predictive model and then we have a candidate explanation for the scientific phenomenon that the ML model predicts. However, I do not expect things to be quite as easy as that. The problem of confounding variables remains, and if we were to abstract a structural equation model from the ML model predicting risk of death from pneumonia based on one's medical information (see top of section 3) then we will still get a causal model on which having asthma causes a lower risk of death from pneumonia. Likewise, if we were to abstract a structural equation model from the climate ML model discussed by Kawamleh (2021) then the resulting causal model would fail to respect conservation of energy laws. We may get a candidate for a causal scientific explanation, but it is certainly not going to give us a correct explanation *for one's actual risk of dying*. In other words, the structural equation model may accurately capture the dependencies inside the ML model, it does not capture the causal dependencies in the actual world. Perhaps if the confounding variables, in this example that of receiving additional treatment, are included among the input variables of the ML model such biases can be mitigated. Pietsch (2016) for example, argues for a representative data set that includes all (or at least a substantial part) of the relevant variables. However, the general issue that predictive ML models are biased will remain, and adding variables will not be enough to get out of that problem, at the very least because a predictive ML model has nothing that accounts for the causal relations between the input variables. Rather, it might be a way to end up with more useful structural equation models that can then be refined through testing procedures that involve interventions rather than mere predictions. If we obtain causal scientific explanations at the end of such a longer procedure we can still say that the explanation was obtained in part because of the initially extracted structural equation model from the ML model. Predictive machine learning might not give us immediate explanations in such a scenario, but it still plays an important role in finding new causal scientific explanations. For the more general case there is therefore also some reason for optimism, though it is tied (on my exposition of it here) to identifying what dependencies the ML model exploits to determine the output. That is precisely the kind of requirement of explainability that Sullivan (2019) aim to avoid by focusing on link uncertainty.

5 Conclusion

Can we acquire scientific (causal) explanations from machine learning models? I argue that we can answer positively, though with some reservations. I've shown that machine learning models specifically designed for the two steps of causal inference (learning causal graphs and estimating treatment effects) fit the bill. We can identify possible causes using purpose-built machine learning models *and* we can estimate the strength of the causal relations using double machine learning methods. How-

ever, standard machine learning models are not suited for this type of causal inference, as they are systematically biased due to the lack of control for confounding variables. Likewise, standard machine learning models seem difficult to use for learning causal graphs/causal discovery as they do not give us any information on the causal relations between the input variables.

That being said, there are promising techniques that aim to extract structural equation models from predictive ML models. These can give us new candidate explanations to test, based on ML models that manage to predict physical phenomena particularly well. As these structural equation models will inherit the problems of predictive ML models (i.e. fail to account for confounding variables as well as include violations of physical laws) I consider it unlikely that they will immediately yield good explanations without further work on our part. Still, they may be valuable starting points in the formulation of new scientific explanations and as such this gives us a route to causal scientific explanations from ML models that is more concrete than that offered by e.g. Meskhidze (2021). In addition, it introduces some caution to the arguments of Pietsch (2016), who does not discuss confounders or the tendency of ML models or the law violations that complicate the extraction of causal explanations from ML models.

References

Agarwal, S., Abdalla, F. B., Feldman, H. A., Lahav, O., and Thomas, S. A. (2012). Pkann—i. non-linear matter power spectrum interpolation through artificial neural networks. *Monthly Notices of the Royal Astronomical Society*, 424(2):1409–1418.

Baiardi, A. and Naghi, A. (2020). The value added of machine learning to causal inference: Evidence from revisited studies. *SSRN Electronic Journal*.

Batterman, R. W. (1992). Explanatory instability. *Nous*, 26(3):325–348.

Beckers, S. (2022). Causal explanations and xai. In *Conference on Causal Learning and Reasoning*, pages 90–109. PMLR.

Beckers, S. and Halpern, J. Y. (2019). Abstracting causal models. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pages 2678–2685.

Bellot, A. and van der Schaar, M. (2019). Conditional independence testing using generative adversarial networks. *Advances in Neural Information Processing Systems*, 32.

Biswas, S., Corti, L., Buijsman, S., and Yang, J. (2022). Chime: Causal human-in-the-loop model explanations. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 27–39.

Buijsman, S. (2022). Defining explanation and explanatory depth in xai. *Minds and Machines*, 32(3):563–584.

Cao, Y., Kang, Q., Zhang, B., Zhu, Z., Dong, G., Cai, Q., Lee, K., and Chen, B. (2022). Machine learning-aided causal inference for unraveling chemical dispersant and salinity effects on crude oil biodegradation. *Bioresource Technology*, 345:126468.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Das, A. and Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*.

Duncan, W. D. (2017). Ontological distinctions between hardware and software. *Applied Ontology*, 12(1):5–32.

Geiger, A., Lu, H., Icard, T., and Potts, C. (2021). Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586.

Geiger, A., Potts, C., and Icard, T. (2023). Causal abstraction for faithful model interpretation. *arXiv preprint arXiv:2301.04709*.

Glymour, C., Zhang, K., and Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524.

Halpern, J. Y. and Pearl, J. (2005). Causes and explanations: A structural-model approach. part ii: Explanations. *The British journal for the philosophy of science*.

Jebeile, J., Lam, V., and Räz, T. (2021). Understanding climate change with statistical downscaling and machine learning. *Synthese*, 199(1):1877–1897.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

Kalainathan, D., Goudet, O., Guyon, I., Lopez-Paz, D., and Sebag, M. (2018). Structural agnostic modeling: Adversarial learning of causal graphs. *arXiv preprint arXiv:1803.04929*.

Kawamleh, S. (2021). Can machines learn how clouds work? the epistemic implications of machine learning methods in climate science. *Philosophy of Science*, 88(5):1008–1020.

Knüsel, B. and Baumberger, C. (2020). Understanding climate phenomena with data-driven models. *Studies in History and Philosophy of Science Part A*, 84:46–56.

Lachapelle, S., Brouillard, P., Deleu, T., and Lacoste-Julien, S. (2019). Gradient-based neural dag learning. *arXiv preprint arXiv:1906.02226*.

López-Rubio, E. and Ratti, E. (2021). Data science and molecular biology: prediction and mechanistic explanation. *Synthese*, 198(4):3131–3156.

Meskhidze, H. (2021). Can machine learning provide understanding? how cosmologists use machine learning to understand observations of the universe. *Erkenntnis*, pages 1–15.

Milkowski, M. (2013). *Explaining the computational mind*. Mit Press.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.

Piccinini, G. (2010). The mind as neural software? understanding functionalism, computationalism, and computational functionalism. *Philosophy and Phenomenological Research*, 81(2):269–311.

Pietsch, W. (2016). The causal nature of modeling with big data. *Philosophy & Technology*, 29:137–171.

Rasp, S., Pritchard, M. S., and Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, 115(39):9684–9689.

Räz, T. and Beisbart, C. (2022). The importance of understanding deep learning. *Erkenntnis*, pages 1–18.

Schmidt, J., Marques, M. R., Botti, S., and Marques, M. A. (2019). Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, 5(1):1–36.

Sen, R., Suresh, A. T., Shanmugam, K., Dimakis, A. G., and Shakkottai, S. (2017). Model-powered conditional independence test. *Advances in neural information processing systems*, 30.

Shah, R. D. and Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538.

Shi, C., Xu, T., Bergsma, W., and Li, L. (2020). Double generative adversarial networks for conditional independence testing. *arXiv preprint arXiv:2006.02615*.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. (2000). *Causation, prediction, and search*. MIT press.

Srećković, S., Berber, A., and Filipović, N. (2021). The automated laplacean demon: How ml challenges our views on prediction and explanation. *Minds and Machines*, pages 1–25.

Stinson, C. (2018). Explanation and connectionist models. *The Routledge handbook of the computational mind. New York, NY: Routledge*.

Sullivan, E. (2019). Understanding from machine learning models. *The British Journal for the Philosophy of Science*.

Turner, R. (2011). Specification. *Minds and Machines*, 21:135–152.

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford university press.

Wu, Z., D'Oosterlinck, K., Geiger, A., Zur, A., and Potts, C. (2023). Causal proxy models for concept-based model explanations. In *International Conference on Machine Learning*, pages 37313–37334. PMLR.