The Explanatory Role of Machine Learning in Molecular Biology

Fridolin Gross
University of Bordeaux

forthcoming in *Erkenntnis*

Abstract

The philosophical debate around the impact of machine learning in science is often framed in terms of a choice between AI and classical methods as mutually exclusive alternatives involving difficult epistemological trade-offs. A common worry regarding machine learning methods specifically is that they lead to opaque models that make predictions but do not lead to explanation or understanding. Focusing on the field of molecular biology, I argue that in practice machine learning is often used with explanatory aims. More specifically, I argue that machine learning can be tightly integrated with other, more traditional, research methods and in a clear sense can contribute to insight into the causal processes underlying phenomena of interest to biologists. One could even say that machine learning is not the end of theory in important areas of biology, as has been argued, but rather a new beginning. I support these claims with a detailed discussion of a case study involving gene regulation by microRNAs.

1 Introduction

Artificial intelligence (AI) methods based on machine learning techniques have recently led to some spectacular advances in many different fields, such as image recognition, language translation, or games like chess and Go. In the scientific context, the most recent example is the success of the AI system "AlphaFold" in predicting the folded structure of proteins based on their amino-acid sequence (Callaway, 2020; Jumper et al., 2021). Machine learning gives rise to a number of epistemological issues due to the differences from classical computational methods. A common question is whether the widespread application of machine learning methods will limit scientific activity to a merely predictive enterprise that does not provide any insight into the causal processes underlying the investigated phenomena. Already more than a decade ago, Chris Anderson made the provocative claim in an article in *WIRED* that big data will herald the "end of theory" and make the traditional scientific method obsolete (Anderson, 2008). The topic does not seem to have lost its appeal, as Laura Spinney recently discussed in *The Guardian* the question of whether we are "witnessing the dawn of post-theory science" (Spinney, 2022). Philosophers of science and scientists have taken up the central question of this debate by addressing the impact that the introduction of AI and machine learning is likely to have on the general character of scientific research (e.g. Pietsch, 2015; Canali, 2016; Coveney et al., 2016; Boon, 2020; Creel, 2020; Ourmazd, 2020; Boge and Poznic, 2021; López-Rubio and Ratti, 2021; Boge et al., 2022; Krenn et al., 2022; Duede, 2023; Andrews, 2023). The areas of genetics and molecular biology, which over the last few decades have become highly "data-centric" (Leonelli, 2016), seem particular prone to making the shift from a theory- or hypothesis-driven mode towards purely data-driven modes of research. While several philosophers have highlighted that the idea of a choice between hypothesis-driven and data-driven science is based on a false dichotomy, and that the more important question is how these modes of research can be integrated in scientific practice (e.g. O'Malley and Soyer, 2012), the discussion around machine learning seems to be commonly framed in terms of AI and classical methods as mutually exclusive alternatives.

Based on a case study from the field of microRNA research (McGeary et al., 2019), I show in this paper how in practice machine learning is often tightly integrated with mechanistic and explanatory research strategies. Far from precluding mechanistic insight, machine learning can actually contribute to the understanding of the causal factors underlying complex phenomena. My case study suggests that machine learning may in some scientific contexts not represent the end of theory, but rather a new beginning.

At this point, perhaps a comment on my use of the term "theory" is in order. Biologists often reserve this term for highly general principles that can be applied across a wide variety of contexts (e.g., Darwin's theory of natural selection, Schwann and Schleiden's cell theory, or Burnet's clonal selection theory of immunity). Consistent with other contributors to the machine learning and AI debate, I use the term "theory" here in a much broader sense to refer to an aspect of science that looks at the underlying causes of observed phenomena and uses that understanding to make predictions and propose interventions. Some biologists seem to believe that their work is not based on any kind of theorizing because their knowledge claims are somehow "directly" derived from experiments, or because they do not use formal or quantitative tools. I think this is misleading, and that theory and theorizing permeate biology whenever biologists use mechanisms or models to explain the phenomena they study.

The article is organized as follows. In Section 2, I present some basics of machine learning methods. By surveying the existing philosophical literature on the subject, I then point out that machine learning and more conventional modeling methods are generally viewed as alternative and mutually exclusive approaches. In Section 3, I describe McGeary et al.'s microRNA repression model in which machine learning plays a central role, but is integrated in a more complex research constellation involving different methods. In Section 4, I use the insights from this case study to argue for three claims of increasing strength: machine learning can be tightly coupled with conventional methods; it can indeed contribute to explanation and understanding; and it may even represent a new form of theorizing in the context of biology. Section 5 contains concluding remarks, discussing in particular the generalizability of the results of my case study.

2 Machine Learning and the End of Theory

The term "machine learning" refers to a set of computational methods that are capable of solving complex tasks not by explicit instructions, but by learning from data. Exactly what this "learning" means depends on the specific type of method being used. A basic distinction that is often made is between "supervised" and "unsupervised" learning methods. In supervised learning one trains an algorithm by providing not only a set of data, but also information about the right kind of output that the algorithm should produce for a given piece of input data. For example, if the task is to classify images, one also provides a set of labels ("cat", "dog", "ball", etc.) that corresponds to the desired classification. In unsupervised learning, by contrast, one does not provide any additional information, but lets the method find features or structure in the data set by itself. Unsupervised learning can be used to sort observations into a set of groups and thus achieve a classification of potentially relevant groups, for example to identify cell types in single-cell gene expression data. Artificial neural networks (ANNs) are arguably the best known examples of machine learning methods. They consist of several layers of artificial neurons, which can be understood as simple computational units that convert a set of inputs into an integrated output according to specific parameters. The first layer ("input layer") takes the data input directly, while the last layer ("output layer") produces a numerical output. In numerical prediction (regression) tasks, this output is directly interpreted as a prediction of the model, while in classification tasks the output can be interpreted as a vector of probabilities corresponding to the set of possible classes. The layers between input and output are called "hidden layers". In the supervised setting, "learning" consists in incrementally updating the parameters in order to minimize the deviation (also known as the "cost") between the predicted and the desired output. Depending on the task, such networks can be enormously complex with thousands or millions of parameters. Machine learning is considered "deep" to the extent that there are multiple hidden layers in the model architecture that correspond to successive transformations of the input data.

The complexity of the models and the fact that the details of the learning process are not externally controlled by human agents makes it extremely difficult, if not impossible, to understand how a task that has been learned by such a model is actually carried out. This is also known as the problem of the "opacity" of machine learning algorithms (Burrell, 2016; Creel, 2020; Boge, 2022; Zerilli, 2022). In the scientific context, machine learning models can be extremely accurate at predicting certain outcomes, but this accuracy seems to come at the cost of not knowing *how* they generate these predictions. This might be seen as problematic for several reasons. For example, it could undermine trust in or reliance on such models (Duede, 2023), in particular when it comes to generalizing beyond the kind of input data that the model has "seen" in the training phase (Räz, 2022b). On the other hand, we may be troubled by the fact that we have only a limited idea of how the structure of the models incorporates features of the external world to which they are supposed to refer (Alvarado and Humphreys, 2017). In particular, their way of making predictions seems to be altogether different from the "traditional" way of predicting based on an explicit representation of underlying processes occurring in the target system. Two different problems of understanding should be distinguished in this context: understanding *of* a model and understanding *with* a model (Räz and Beisbart, 2022). First, machine learning methods often lead to models that are very difficult or even impossible to understand themselves, i.e. the way in which they predict certain outputs given certain input data. Second, these models seem to be useless for the purpose of understanding features of the world, given that there does not seem to be any obvious representational relation between model and target system. Both of these problems add to the feeling that scientific research based on machine learning deviates considerably from the traditional theory-based approach.

In line with this, the philosophy of science literature has tended to discuss machine learning mostly in terms of a choice between competing epistemic values. For example, Boge and Poznic (2021) raise the worry that machine learning will turn science away from the aim of explanation towards mere pattern recognition and prediction. If explanations remain an important goal, it is argued, then additional "second-order" explanatory efforts must be made to first understand how machine learning models perform their predictive tasks (cf. Zednik, 2021). This is the goal of the "explainable AI" (xAI) movement, which has received considerable attention (see e.g. Watson and Floridi, 2021; Watson, 2022a; Zednik and Boelsen, 2022; Zerilli, 2022; Beisbart and Räz, 2022; Räz, 2022a), but also generated skeptical reactions (e.g. Rudin, 2019). In any case, such explanations aim at making the decision-making of algorithms transparent without thereby achieving the goals of providing explanation and understanding about phenomena in the world.

Similarly, López-Rubio and Ratti (2021) frame the issue in terms of a choice between black box machine learning models and mechanistic strategies when scientists are confronted with complex and chaotic systems. They suggest that there is a direct trade-off between the explanatory and predictive aims of science. Boon (2020) asks whether the traditional toolkit of science, i.e. concepts, laws, models, and theories will become superfluous with the advent of machine learning. She argues that such a transition would be in line with a purely empiricist stance on the aims of science that does not worry about the unobservable causes that bring about the observed phenomena. In a very similar vein, Hooker and Hooker (2018) see the contrast between machine learning and conventional methods as an expression of the "realist/anti-realist cleavage that runs through the philosophy of science" (Hooker and Hooker, 2018, 3).

Creel points out that to the extent that explanation requires insight into underlying mechanism, opacity will be an obstacle to the explanatory goals of science (Creel, 2020). Srećković et al. (2022) imagine that machine learning is likely to lead to the emergence of two separate strands of scientific activity, one "purely predictive and detached from any explanatory efforts", and one that is "faithful to the anthropocentric research focused on the search for explanation" (Srećković et al., 2022, 179). Finally, Boge (2022) argues that deep neural nets are purely instrumental models that do not deliver explanations and stand in the way of proper conceptualizations of the target system.

While I do not want to deny the problematic aspects of machine learning, I agree with Duede (2023) when he states that "the disconnect between philosophical pessimism and scientific optimism is driven by a failure to examine how AI is actually used in science" (Duede, 2023, 1). In particular, it seems to me that most philosophical discussions are implicitly based on the idea that given a certain problem, scientists are faced with a choice between machine learning method or a "traditional" method as mutually exclusive alternatives.

Some philosophers have questioned the idea that machine learning models are incompatible with scientific understanding. Sullivan (2022), for example, argues that it is not the complexity or opaqueness of a model that limits understanding, but rather a property that she calls "link uncertainty", or "the extent to which the model fails to be empirically supported and adequately linked to the target phenomena" (Sullivan, 2022, 110). As a case study, Sullivan discusses the "deep patient model", a deep learning model that allows the prediction of health states of patients from electronic health records. Similarly, Knüsel and Baumberger (2020) aim to assess the extent to which models can serve as vehicles for scientific understanding based on their representational accuracy and coherence with background knowledge. They compare different climate models and conclude that machine learning models can serve as vehicles for understanding under the right circumstances.

However, both Sullivan and Knüsel and Baumberger consider scenarios in which the machine learning model is used in isolation for a particular scientific task. This overlooks the possibility that the most interesting contribution of machine learning to explanation or understanding may lie in combining these methods with others (Baker et al., 2018).

Underlying all views discussed so far is the notion that ML and traditional modeling methods actually differ in epistemologically relevant aspects. This "distintness claim" is challenged by Andrews (2023). She argues that it is a mistake to consider data-intensive methods such as machine learning to be theory-free, since they are typically applied in highly context-specific environments and the construction of models and preparation of data sets are theory-laden in important ways. Although I am broadly sympathetic to this position, I think there remain important methodological differences, and it seems that the theory-ladenness discussed by Andrews (2023) is largely *external* in the sense that it does not imply any specific assumptions about the causal structure of the phenomena under study (Pietsch, 2015). In this regard, it is useful to consider the distinction between “process-based models” and “data-driven models”, introduced by Knüsel and Baumberger (2020). While the former explicitly represent the causal processes that are assumed to occur in the target system, the latter are created using statistical learning methods in order to capture dependencies between observable features of the system. A standard example of a process-based model is the Hodgkin-Huxley model which describes the initiation and propagation of action potentials in neurons (see e.g. Weber, 2005; Craver, 2008). Here, the equations of the model directly represent characteristics of excitable cells, with the variables and parameters of the model standing for measurable quantities such as the capacitance of the lipid bilayer constituting the cell membrane or the voltage of ion channels. Experimental measurements serve as valuable information both to infer the basic organization of the mechanism (e.g. which components causally interact) and to determine the quantitative parameters figuring in the mathematical description of the mechanism. Process-based models can be used to make predictions, but because of the inclusion of causal processes, they are also considered capable of explaining properties of the target system. In data-driven models, by contrast, there is not necessarily a straightforward mapping between model components and features of the world. In a deep neural network, for example, the architecture, i.e., the number of layers, the number of nodes in those layers, and their connections, is usually unrelated to the structural features of the target system. Data is used to tune the parameters of the model in such a way that the model reproduces the observed input-output relationship, but the same type of model could in principle be used to fit many different kinds of such relationship. Figure 1 shows schematically the respective roles of the two types of models in scientific contexts. It should be noted that the nature of the mathematical formalism by itself does not determine whether it is used for process-based or data-driven modeling. For example, ordinary differential equations can be used in a purely data-driven manner, while neural networks can sometimes be used as process-based models, e.g., to represent the architecture of parts of the nervous system (Chirimuuta, 2021; Stinson, 2020; Cao and Yamins, 2021).

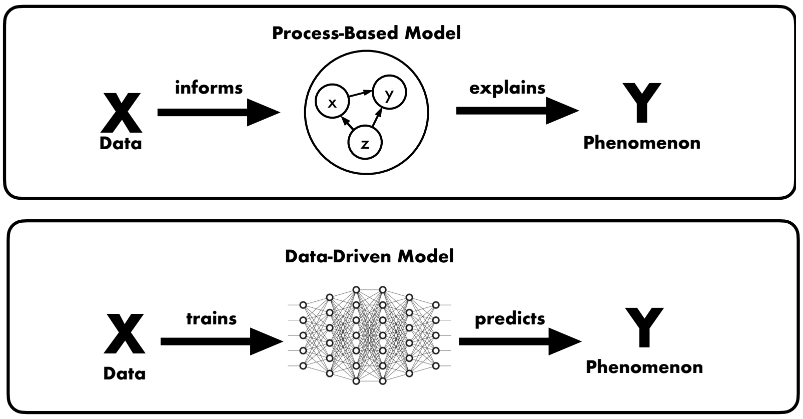


Figure 1: Two different kinds of modeling approaches according to the distinction proposed by Knüsel and Baumberger (2020). Process-based models (top panel) explicitly represent entities (indicated by the variables x, y, z) and causal processes occurring in the target system and are often used for explanation. Data-driven models (bottom panel) are constructed using techniques of statistical learning and do not aim at representing underlying causal structure. They are typically used for prediction.

However, it is clear that both of these scenarios abstract from important aspects of scientific activity. In practice, computational models are rarely used in isolation and are typically integrated in larger research contexts. It seems that most philosophical discussions of machine learning methods take the form of an explicit comparison with more traditional forms of theorizing in science. While such comparisons are generally informative and useful, they can be misleading if they imply that one has to choose between machine learning and conventional methods, or between process-based and data-driven models, as if they were mutually exclusive alternatives. It seems more productive to think of machine learning as an addition to the toolkit of scientists that is appropriate in certain contexts and that can be complemented with other kinds of methods. In such a bigger picture, machine learning does not necessarily threaten to turn science into a merely predictive endeavor based on opaque models, but can, on the contrary, contribute to the articulation of explanations and to the illumination of phenomena and their underlying causal processes. In the next section I will provide a detailed case study from the field of molecular biology that illustrates this productive explanatory role.

3 Case Study: A Biochemical Model of microRNA Affinity

3.1 Biological Background

For a long time, ribonucleic acid (RNA) was considered to play a subordinate role in molecular biology, compared to DNA and proteins. According to the so-called “central dogma” formulated by Francis Crick (1958), its main function was to act in the form of messenger RNA (mRNA) as an intermediary “transcript” to transmit the information contained in gene sequences in DNA required for the synthesis of proteins. The prevailing view that emerged in the following decades was that gene regulation, that is, the set of processes that determine which genes are expressed by the cell in a given context and at what levels, was due to the action of specialized proteins called “transcription factors” (Morange and Cobb, 2020). However, early on some mechanisms were described in which RNA molecules intervened more actively in this process. These were based on the principle of sequence complementarity, i.e., the idea that RNA can recognize and bind to specific sequences of DNA when the “letters” (called nucleotides) in both sequences correspond exactly. One class of such regulatory RNA molecules, later called microRNAs, was discovered in the late 1980s when scientists were unable to explain the phenotypic characteristics of a genetic mutant in the nematode worm *C. elegans* using the traditional model of gene regulation. When they investigated the underlying gene, they found that it did not give rise to a messenger RNA molecule encoding a protein, but instead produced a short non-coding RNA that contained sequences partially complementary to multiple sequences in another (conventional) gene, which suggested that complementary binding of the RNA was involved in the repression of the target gene.[1] Later discoveries showed that this regulatory mechanism is not peculiar to worms, but is ubiquitous in the animal world. It is now known that microRNAs are small ($\sim$22 nucleotides long) RNA molecules that act mainly by complementary binding to the untranslated portion of mRNAs (i.e., the part of the sequence not involved in specifying the synthesis of proteins), causing their increased degradation and thus decreased expression of the corresponding genes (see Fig. 2). The number of different microRNA species in humans has not been determined with precision, but is estimated to lie somewhere between 600 and 2000, and biologists assume that most conventional

[1] For a more extensive account of the history of microRNA discovery, see Burian (2007); O’Malley et al. (2010).

genes are targets of at least one microRNA. They are involved in important biological functions and have been implicated in a number of human diseases (Bartel, 2018).

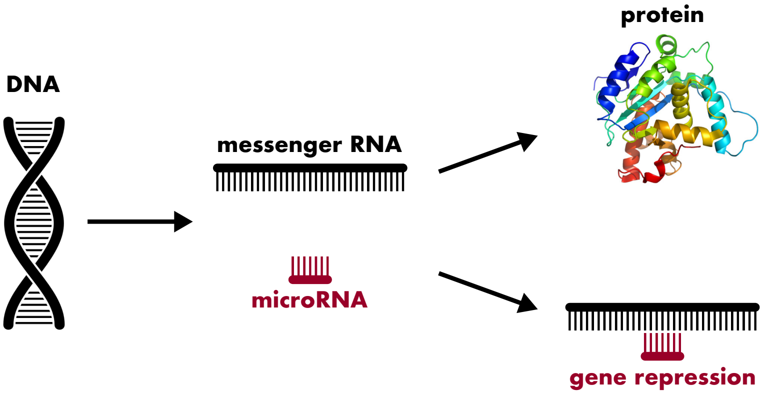


Figure 2: Schematic illustration of gene regulation by microRNAs.

Importantly, an individual microRNA can target a set of different mRNA molecules, and a particular mRNA can have binding sites for several different microRNAs. Therefore, a critical task in understanding the specific ways in which microRNAs affect gene regulation is to determine the targets for a given microRNA in a given cellular context and, conversely, to determine the set of microRNAs that affect a given mRNA. The suppressive potential of a given microRNA appears to be very specific and directly related to the number and affinities of the binding sites on the target molecule. In particular, complementary binding of nucleotides 2-7 of the microRNA, the so-called seed region, is of central importance. However, not all of the binding sites with direct seed pairing are equally effective, and non-canonical sites (i.e. with non-perfect pairing) have been shown to be relevant in some cases as well. Importantly, microRNAs do not act alone, but are incorporated into a protein complex called RISC (for “RNA-induced silencing complex”), which affects the binding characteristics of a microRNA molecule by modifying its three-dimensional configuration. As a result, the affinities of microRNA binding sites depend in complex ways on the wider molecular context and are very poorly predicted based on chemical principles of RNA site pairing alone. Up to recently, the most successful methods in predicting target affinity relied on indirect, correlative approaches, fitting multiple linear regression models based on the main features thought to be informative, including, for example, the conservation of binding sites across different organisms. However, even the best of these models have been shown to explain only a small fraction of the effects on gene expression that are attributable to microRNAs.

3.2 Predicting and Understanding Gene Regulation by microRNAs

In 2019, the group of David P. Bartel at MIT, one of the key figures in microRNA research, published a study called “The biochemical basis of microRNA targeting efficacy” that aimed at providing the foundation for a more detailed mechanistic understanding of microRNA action, while at the same time generating more accurate predictions of target repression (McGeary et al., 2019). Interestingly, machine learning played a key role in this endeavor together with an innovative experimental technology, called RNA Bind-n-Seq (RBNS). In addition, the approach included a second model, which the authors call a “biochemical model,” that describes the interaction of microRNA and mRNA molecules. And finally, it relied on so-called “transfection experiments” in human cell lines. The cell lines were genetically engineered to express high levels of a specific microRNA, and the effect of this overexpression on the cellular levels of mRNAs was measured by genome-wide RNA sequencing. To appreciate the role of machine learning in the context of this project, it is important to clearly understand how these different components were integrated into the overall approach.

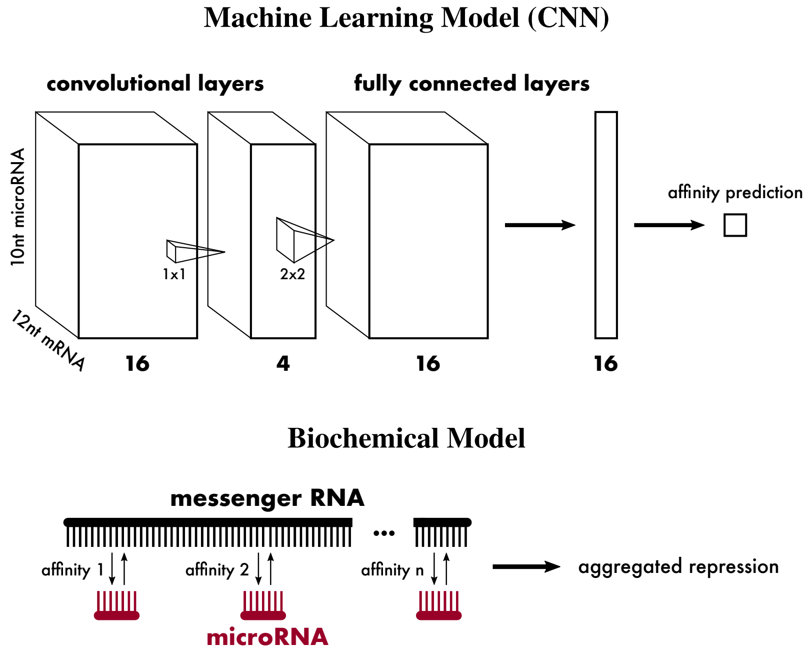
3.2.1 Experimental Approach

In the RBNS experiments the binding affinities of a small set of six microRNAs to all 262,144 possible 12 nucleotide long RNA sequences with at least four contiguous matches to the microRNA seed were determined *in vitro*, that is, in a controlled chemical test-tube environment containing only the RNA molecules and AGO2, the relevant component of the RISC protein complex. These experiments were interesting in their own right, as they provided an exhaustive view on the possible transcript interactions and led to the discovery of several new types of binding sites. In addition, the measurements showed that the affinity for the same type of binding site can differ substantially between microRNAs, thus further highlighting the importance of taking into account the wider sequence context. As expected, the measured affinities also differed substantially from direct predictions based on the free energy of site pairing. In the transfection experiments five of the six microRNAs already used in the RBNS experiments, as well as 10 others, were overexpressed in a human cell line (HeLa) in order to measure resulting changes in gene expression. Both types of experiments were used to generate data to train and test the combined model of microRNA regulation.

3.2.2 Two Types of Models

Since it is not realistic to measure for each microRNA the affinities for all possible sequences, the authors decided to use a convolutional neural network (CNN) model to “learn” the affinity between the sequence of any microRNA and any target sequence based on the experimental data generated in the first step. CNNs are often used to analyze images, and their architecture is inspired by the organization of the visual cortex in animals. They allow for the detection and subsequent combination of complex features in the data by using filter functions that combine information coming from contiguous patches of an image and that are optimized along with the other parameters of the model. The input data used by McGeary et al., which consisted in partial sequences of the microRNA and the mRNA, were presented to the CNN as $10 \times 12 \times 16$ matrices, corresponding to 10 nucleotide positions of the microRNA, 12 nucleotide positions of the mRNA, and 16 possible pairings of the four RNA nucleotides (adenine (A), cytosine (C), guanine (G) and uracil (U)), which is in line with the three-dimensional input format of image data (image height × image width × number of color channels). While convolutional layers are typically used to capture spatial dependencies between pixels in an image, McGeary et al. used them to capture dependencies between adjacent nucleotide positions in the molecular interactions between mRNA and microRNA. In addition to two convolutional layers, they added two fully connected layers in order to allow for more general types of interaction between the two molecules (see Fig. 3, top).

Based on the idea that binding affinity is directly related to the target repression observed in cells, they additionally built a model that infers the degree to which a given microRNA represses a given mRNA. This model considers full length mRNA transcripts and predicts the total occupancy of this transcript by the microRNA (i.e. how many copies of the microRNA are on average bound



Machine Learning Model (CNN)

Biochemical Model

Figure 3: The two models used in the study. The pyramidal shapes connecting the layers of the CNN indicate the size of the convolutional filters (1x1 or 2x2). Both are simplified and modified versions of diagrams found in McGeary et al. (2019).

to the mRNA at equilibrium) by aggregating the affinities of all potential binding sites on the mRNA. This "biochemical" model was derived from an explicit description, in terms of ordinary differential equations, of the dynamic processes by which microRNAs bind to and detach from the mRNA and of the enhanced degradation caused by this occupancy. Apart from the site affinities, the model contained parameters describing the relative concentration of unbound microRNA complexes and the repression caused by a single bound microRNA, as well as a parameter that accounts for the fact that sites within translated regions are less effective for repression than sites in untranslated regions.

Figure 4 schematically describes the training process, which consisted in linking both models to both sets of experiments. The CNN was used to predict binding affinities of microRNAs to 12 nucleotide length sequences, and the biochemical model to in turn link those affinities to experimentally measured intracellular repression levels. The overall cost function to be minimized during training was composed of two parts: the cost due to the differences between affinities measured in the RBNS experiments and those predicted by the CNN, and the cost due to the differences between the gene expression changes measured in the transfection experiments and those predicted by the biochemical model. The parameters of both models were simultaneously optimized in the process.

After training, McGeary et al. tested the combined model using microRNAs not present in the training set and additional transfections experiment performed in a different human cell line (HEK). They found that their new method showed an impressive 50% improvement over the best existing prediction tools.

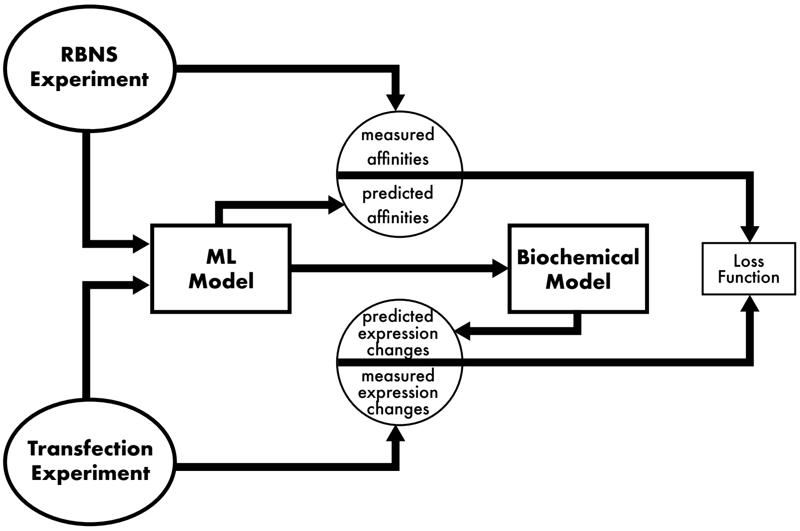


Figure 4: Schematic illustration of the training setup. This is a simplified and modified version of the diagram shown in Fig. 6A of McGeary et al. (2019).

4 Analysis of the Case Study

In this section I derive several insights from the microRNA case study. These insights go against the picture, widespread in the literature on AI, according to which there is a trade-off between the predictive and explanatory aims of science, and that machine learning stands in the way of mechanistic insight or theoretical understanding due to the opacity of its models. Clearly, a single example is not sufficient to understand the evolution of an entire field of science, but as I will suggest, it is plausible to view this example as representative of a broader trend at least in the domain of molecular biology.

For the purposes of this article, it will not be necessary to go into the intricacies of the debates over mechanistic explanation and scientific understanding. My arguments rest on assumptions that I think are fairly uncontroversial: I assume that an account of the causal processes, or the mechanism, underlying a phenomenon of interest can, under the right circumstances, serve as a means of explaining that phenomenon. Furthermore, I assume that such causal or mechanistic explanations, again under the right circumstances, can serve as vehicles for scientific understanding. The "right circumstances" may be formulated, for example, in terms of a criterion of intelligibility (De Regt and Dieks, 2005) and the resulting understanding may be considered as a skill rather than as a special type of knowledge (De Regt, 2015). Note that in saying this, I am not assuming that all explanations are causal or mechanistic, nor that I hold any particular view of causal or mechanistic explanation. Nor do I assume that all scientific understanding is based on explanations (cf. Lipton, 2009). Although there has been some debate about whether explanations involving mathematics or computational models should be distinguished from mechanistic explanations (e.g. Bechtel and Abrahamsen, 2010; Lange, 2013; Issad and Malaterre, 2015; Craver and Povich, 2017), I believe that most philosophers agree that mathematical or computational modeling techniques are sometimes successfully used by scientists to describe causal processes or mechanisms and thereby contribute to the goals of explanation and understanding. This is the way in which I understand the contribution to explanation and understanding of process-based models as characterized in Section

2.

4.1 Machine learning is tightly integrated with traditional "mechanistic" methods

The case study shows that scientific projects involving machine learning are not necessarily limited to generating data and training a model for purposes of prediction. McGeary et al. combined machine learning with detailed experimental analysis of the causal factors underlying the regulatory role of microRNAs and with process-based modeling.

First, they used the RBNS experiments not only as a training set but also to gain very specific insights into the causal determinants of microRNA binding affinity. Thus, they discovered, for example, that some of the experimentally investigated microRNAs have binding sites in which complementary binding is extended beyond the seed region. Furthermore, they found pronounced differences in the affinities of the same type of binding site across different microRNAs, pointing to the causal importance of the sequence and conformation of the whole microRNA molecule. Finally, they gained insight into the importance of the nucleotides directly flanking the seed regions on the target sequences and explained their effect in terms of "the propensity of these nucleotides to stabilize RNA secondary structure" (McGeary et al., 2019, 8).

Second, they developed their "biochemical model" by connecting the affinity values determined in the RBNS experiments (or predicted by the CNN) to the mRNA repression observed in the cellular context of the transfection experiments. In the terminology proposed in Section 2 this is clearly a process-based model, in the same way as the Hodgkin-Huxley model, because it explicitly represents the process of microRNAs occupying their target transcripts in terms of causally interpretable variables and parameters (e.g. the dissociation constant of target binding, the concentration of free microRNA molecules, the transcript degradation rate). Such a model can be interpreted as a mathematical representation of a causal mechanism and therefore as potentially conducive of explanation and understanding.

The CNN model had a particular role in the overall project: to transfer the causal-mechanistic insights from the experiments and the biochemical model to other microRNAs that were not among the ones studied experimentally. There is thus a direct continuity between the causal-mechanistic analysis and the machine learning part of the study. As shown in Figure 4, the CNN was directly integrated with the biochemical model to turn sequence-based affinity predictions into predictions of transcript repression. Thus, the integration of data-driven and process-based models occurred in a very direct sense as the integration of different software modules in the overall computational approach.[2]

In addition to this very explicit type of integration, it is also worth noting that the CNN model itself can be understood to some extent as integrating causal-mechanistic knowledge. For example, McGeary et al. did not use a generic neural network, but the architecture of the model actually reflects features of the target system, with nodes corresponding to nucleotide positions and the choice of convolution filters corresponding to molecular interactions considered relevant. Moreover, the choice of training data amounts to selecting mechanistic features of the target system that seem important to the scientists.[3]

[2] As can be seen in the accompanying software repository (`https://github.com/kslin/miRNA_models`).

[3] I thank one of the anonymous reviewers for highlighting this point. A much more detailed account of the "theory-ladenness" of machine learning applications in science can be found in Andrews (2023).

4.2 Machine learning does not impede but contributes to mechanistic insight

As we have seen, it is a common view that machine learning models will enable scientists to make predictions about the behavior of very complex systems without providing any insight into how this behavior is brought about by the interplay of the systems' components. Consequently, machine learning is seen as impeding mechanistic insight and understanding. The example of the microRNA study shows this does not have to be the case. As discussed in the previous section, machine learning can go hand in hand with mechanistic methods, but based on the case study one can even make the stronger claim that it actually *contributes* to mechanistic understanding. This is because the predictions generated by machine learning in the microRNA study are not "high-level phenomena" like, for example, the disease status of a patient. Instead, the predictions serve to quantify the affinity of microRNA molecules to their target binding sites, which is a relatively "low-level" mechanistic feature that is used by biologists in detailed explanatory accounts of specific gene regulation mechanisms, and the fact that these affinity values are generated by a machine learning method does not necessarily make these accounts any less explanatory. However, in line with Machamer et al. (2000), one might object that a model containing machine learning based predictions is merely a "mechanism sketch", i.e., an incomplete explanation containing "black boxes" that have yet to be filled in. Accordingly, the explanation would not be truly complete until the machine learning prediction is replaced by an explicit description of the causal factors that give rise to the affinity between molecules. It is clear, however, that the obligation to open black boxes cannot apply unconditionally, since biologists are not usually required to go to the level of elementary particle physics in their explanations. Indeed, Machamer et al. themselves emphasize that the accounts of nested mechanisms "bottom out" in a discipline-specific lowest level of description. For the scientists of this discipline this is the level where "[t]he explanation comes to an end, and description of lower-level mechanisms would be irrelevant to their interests" (Machamer et al., 2000, 13). Coming back to the case study, it seems that a detailed account of the interactions of the individual nucleotides participating in the bonding between the two RNA molecules lies below the bottoming out level of most biologists who are interested in mechanisms of gene regulation. Thus, even if it does not provide an explanatory account of molecular affinity, the CNN provides important information about *who interacts with whom*, and therefore directly contributes to the construction of causal-mechanistic explanations. In further support of this idea, it seems highly plausible that biologists would accept an account of microRNA repression as mechanistic if the affinity values were not determined by machine learning, but by experimental measurement, even though this would not illuminate the causal factors underlying affinity either.[4]

One reason that this type of explanatory role has not been considered may be that typical examples of machine learning involve high-level predictions based on a broad and undifferentiated database, e.g., disease diagnoses based on gene expression (Watson, 2022b). In such circumstances, it appears that the machine learning approach is diametrically opposed to a representation of the actual causal processes mediating between the different levels and replaces it with a purely correlative model. In my case study, however, the machine learning approach remains confined to the molecular level, and the data were processed to ensure causal relevance. Due to these features, it can be integrated with a mechanistic account rather naturally.

To make this point, I have not even considered whether it might be possible to learn something about the causal factors underlying binding affinity from the machine learning model itself. As

[4] This is similar to the idea that AI sometimes functions like a "computational microscope" (Krenn et al., 2022). Nevertheless, there are important epistemological differences between machine learning and the use of scientific instruments (Duede, 2022).

detailed above, McGeary et al. used a specifically designed convolutional neural network for the task of affinity prediction. This allowed them to adapt the network architecture to specific features in the data that they expected to be particularly relevant:

> The first layer of the CNN was designed to learn important single-nucleotide interactions, the second layer was designed to learn dinucleotide interactions, and the third layer was designed to learn position-specific information. (McGeary et al., 2019, 10)

Thus, even though the study itself does not elaborate on this, it is plausible that such a network can be interrogated using techniques of xAI in order to gain more insight into the relevant causal factors underlying microRNA binding. Indeed, as shown in Soutschek et al. (2022), the model can be used, for example, to extract information about the specific importance of individual nucleotides within a microRNA for binding to its target sites.

4.3 Machine learning is not the end but a new beginning of theory

Anderson (2008) argued in his provocative article that scientists in general, and biologists in particular, should accept that an approach based on models and theory is becoming obsolete. He proposed, instead, that they would be more effective at advancing science by mining massive data sets for correlations. My case study suggests that this is not necessarily the way in which biology is heading. On the contrary, it points to the possibility that data-driven methods can contribute in sophisticated ways to new ways of theory building.

First of all, McGeary et al. did not use machine learning from the start as a method of simply “crunching the data”, but they introduced it only after more direct ways of gaining insight into the causal factors determining microRNA binding affinity were exhausted. It is important to note that if there had been a way to determine affinity based on a few simple rules (e.g. the free energy of the pairing nucleotides), then a machine learning model would not have been necessary and a process-based model would have been preferred. But the experiments on just a few microRNAs revealed that there is a non-trivial dependency of affinity on features of the wider sequence context and, even though data-driven, a CNN model appeared to be the adequate tool to capture this kind of complexity and context-sensitivity. This, however, suggests that machine learning was not simply used as a “black box” to generate predictions based on data, it rather promised to give rise to a model that, although not process-based, “embodies” in some sense the principles of the complex biochemistry underlying microRNA activity. Although this interpretation is admittedly vague, it is supported by the account of the scientists themselves. For example, they describe their aim as gaining “deeper understanding”, and their (combined) model as a way to “understand and predict” and of going beyond “correlative models” towards a “principled, biochemical model” (McGeary et al., 2019, 1). Even more tellingly, toward the end of their study, the authors discuss the limitations of their current model, which is restricted to a relatively small number of nucleotides and does not take into account features of the target sequence that are further away from the binding site:

> Perhaps the most promising strategy for accounting for these more distal features will be an unbiased machine-learning approach that uses entire mRNA sequences to predict repression, leveraging substantially expanded repression datasets as well as site-affinity values. In this way, the complete regulatory landscape, as specified by AGO within this essential biological pathway, might ultimately be computationally reconstructed. (McGeary et al., 2019, 12)

This strongly suggests that the authors view machine learning as more than just a predictive tool. The goal of "reconstructing the regulatory landscape" indicates that they view it as an adequate way to provide a theoretical underpinning for the phenomena of microRNA-driven gene regulation. Thus, rather than heralding the end of theory in biology, machine learning may actually lead to a new way of conceiving the meaning of theory in biology.

An obvious objection to this view is that machine learning is inherently unable to capture underlying principles and remains on the surface of "behavioristic" input-output relationships. Comparison with a case in climate science discussed by Kawamleh (2021) may be instructive in this regard. Because it is computationally intensive to explicitly represent small-scale processes such as cloud formation in climate models, some scientists have attempted to replace them with neural network models that are trained using the output of small-scale process-based models. These neural network models are then integrated into large-scale process-based models, giving rise to a similar combination of process-based and data-driven models as in the case study. As Kawamleh argues, the neural network architecture is in principle able to approximate *any* input-output relationship, and it fails "to learn any meaningful underlying principle or physical relations among relata" (Kawamleh, 2021, 1018). She argues that as a consequence, the hybrid model fails to generalize and to make accurate predictions in new contexts, such as temperature regimes beyond the range that has been encountered during training. At first glance, this problem should also affect the microRNA model.

The first thing to note is that, unlike in the climate example, there is no alternative process-based "gold standard" model to which McGeary et al.'s machine learning model could be compared. For this reason, it is unlikely that the data-driven component of the computational treatment of microRNA affinity can be replaced by a more transparent model. More importantly, however, there is a difference between the kind of task that the machine learning model has to learn in the two cases. The scenarios to be considered relevant in the case of the climate system are potentially infinite as we can always imagine the causal factors involved in climate phenomena (such as temperature) to be extended beyond their previously encountered range. For this reason, it is difficult to train a machine learning model on a truly representative sample of possible scenarios. In the context of microRNA regulation, by contrast, the problem is structured such that there is a finite (albeit very large) number of possible microRNA-mRNA interactions. Therefore, a model that predicts all corresponding affinities with acceptable accuracy is conceivable, and it does not seem absurd to consider this model as embodying the "theory" of microRNA regulation.[5]

The central difference, therefore, lies in the fact that the biological case, despite the involved level of complexity, offers a well-circumscribed and finite "problem space". More specifically, the controlled cellular environment limits the range of possible scenarios that are physiologically relevant in a given organism.[6] In addition, the discrete nature and finite number of elements of the genetic code contribute to the numerical manageability of the problem. Considering that many problems in molecular biology exhibit these, or similar, features, it seems plausible to consider this idea of "theory" as more generally applicable in the biological realm.

[5] That is, provided the model is not effectively just a "list" of all measurements. Some degree of compression seems necessary to speak of "theory".

[6] It is not obvious whether McGeary et al.'s model will work across different organisms because relevant features of the molecular context, e.g. the conformation of the RISC complex (or its homologue), may be different. But I do not see this as threatening the idea of theory involved. It just means that the theory may look different for different organisms. For applications of the model to mouse and rat data, see Soutschek et al. (2022).

5 Conclusion

My goal in this paper was to counterbalance a widely held view in the philosophical literature on machine learning according to which scientists face a choice between the competing goals of prediction and explanation. My case study shows that at least in the context of molecular biology, machine learning models do not necessarily work as isolated tools that generate predictions without providing theoretical insight. Instead, such models are often tightly integrated with other methods commonly considered as giving rise to explanation and understanding. Moreover, the predictions generated by the machine learning model in the case study are not related to "high-level" system behavior but quantify important parameters that have a clear causal interpretation and can be used to build more complex accounts of the causal processes underlying gene regulation. Finally, I have suggested that the biologists themselves consider the machine learning model as the adequate vehicle to capture the phenomenon of microRNA binding at a general level and, therefore, as a form of theory.

The context of microRNA biology may be considered too specific to warrant a new outlook on theory in biology. However, as I have argued above, it is plausible that this perspective may capture important parts of the field of genomics more generally. Many phenomena and processes in molecular biology exhibit a similar kind of complexity and context sensitivity that defies understanding in terms of simple rules, but at the same time is translatable into well-circumscribed problems. In line with this, methods of machine learning have been successfully applied in other cases that are related to predicting complex aspects of sequence related biology (e.g. Alipanahi et al., 2015; Whalen et al., 2016; Cuperus et al., 2017).

Even the aforementioned success of AlphaFold in predicting the structure of folded proteins may be seen as fitting into this scheme, although it has been called "a black box" or "an oracle" (Krenn et al., 2022), which does not suggest that it adds much to scientific understanding or explanation. It is important to consider, however, that protein structure prediction is rarely an aim in itself and that scientific problems are usually organized in a modular fashion. Thus, AlphaFold's achievement is generally seen as a contribution to the investigation of the mechanisms underlying important biological phenomena driven by proteins. One observer of the field considers that the most exciting prospect that it can lay the foundation for a new "structural systems biology":

> [S]tructure is the common currency through which everything in biology gets integrated, both in terms of macromolecular chemistries, i.e., proteins, nucleic acids, lipids, etc, but also in terms of the cell's functional domains, i.e., its information processing circuitry, its morphology, and its motility. A structural systems biology would take this seriously, deriving the rate constants of enzymatic and metabolic reactions, protein-protein binding affinities, and protein-DNA interactions all from structural models. (AlQuraishi, 2020)

Although a detailed account of the AlphaFold model is beyond the scope of this article, clear parallels to the prediction of microRNA affinity can be recognized.[7] Both models are used in areas where conventional models have failed and direct experimental measurement is costly or impossible. In addition, both models are used to generate key ingredients for building more comprehensive process-based models. The fact that parts of these models are based on data-driven and opaque sub-models is not seen as threatening their explanatory potential.

[7] See Andrews (2023) for an excellent philosophical analysis of the case of AlphaFold.

In emphasizing that the opacity of such models is not per se a threat to their ability to contribute to explanation and understanding, I do not mean to deny that there are often good reasons to seek transparency. In particular, methods of explainable AI might be used in the process of model validation and in using such models for discovery (Watson and Floridi, 2021; Watson, 2022b).

To conclude, it seems likely that the construction of knowledge in much of biology will remain a piecemeal endeavor, integrating various experimental and theoretical methods to uncover the mechanisms underlying the phenomena of life. Machine learning appears to be an important new tool that can help in this process of knowledge construction.

References

Alipanahi, B., A. Delong, M.T. Weirauch, and B.J. Frey. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology 33*(8): 831–838. https://doi.org/10.1038/nbt.3300 .

AlQuraishi, M. 2020. AlphaFold2 @ CASP14: "It feels like one's child has left home.". Blog entry at https://moalquraishi.wordpress.com/2020/12/08/alphafold2-casp14-it-feels-like-ones-child-has-left-home/, accessed on 2022-05-02.

Alvarado, R. and P. Humphreys. 2017. *New Literary History 48*(4): 729–749. https://doi.org/10.1353/nlh.2017.0037, .

Anderson, C. 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. WIRED magazine, https://www.wired.com/2008/06/pb-theory/, accessed on 2021-05-02.

Andrews, M. 2023. The Immortal Science of ML: Machine Learning & the Theory-Free Ideal. Preprint at https://rgdoi.net/10.13140/RG.2.2.28311.75685.

Baker, R.E., J.M. Peña, J. Jayamohan, and A. Jérusalem. 2018. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology Letters 14*(5): 20170660. https://doi.org/10.1098/rsbl.2017.0660 .

Bartel, D.P. 2018. Metazoan MicroRNAs. *Cell 173*(1): 20–51. https://doi.org/10.1016/j.cell.2018.03.006 .

Bechtel, W. and A. Abrahamsen. 2010. Dynamic mechanistic explanation: Computational modeling of circadian rhythms as an exemplar for cognitive science. *Studies in History and Philosophy of Science Part A 41*(3): 321–333. https://doi.org/10.1016/j.shpsa.2010.07.003 .

Beisbart, C. and T. Räz. 2022. Philosophy of science at sea: Clarifying the interpretability of machine learning. *Philosophy Compass 17*(6): e12830. https://doi.org/10.1111/phc3.12830 .

Boge, F.J. 2022. Two Dimensions of Opacity and the Deep Learning Predicament. *Minds and Machines 32*(1): 43–75. https://doi.org/10.1007/s11023-021-09569-4 .

Boge, F.J., P. Grünke, and R. Hillerbrand. 2022. Minds and Machines Special Issue: Machine Learning: Prediction Without Explanation? *Minds and Machines 32*(1): 1–9. https://doi.org/10.1007/s11023-022-09597-8 .

Boge, F.J. and M. Poznic. 2021. Machine Learning and the Future of Scientific Explanation. *Journal for General Philosophy of Science 52*(1): 171–176. https://doi.org/10.1007/s10838-020-09537-z .

Boon, M. 2020. How Scientists Are Brought Back into Science—The Error of Empiricism, In *A Critical Reflection on Automated Science: Will Science Remain Human?*, eds. Bertolaso, M. and F. Sterpetti, 43–65. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-25001-0_4.

Burian, R.M. 2007. On MicroRNA and the Need for Exploratory Experimentation in Post-Genomic Molecular Biology. *History and Philosophy of the Life Sciences 29*(3): 285–311 .

Burrell, J. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society 3*(1): 2053951715622512. https://doi.org/10.1177/2053951715622512 .

Callaway, E. 2020. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature 588*(7837): 203–204. https://doi.org/10.1038/d41586-020-03348-4 .

Canali, S. 2016. Big Data, epistemology and causality: Knowledge in and knowledge out in EXPOsOMICS. *Big Data & Society 3*(2): 2053951716669530. https://doi.org/10.1177/2053951716669530 .

Cao, R. and D. Yamins. 2021. Explanatory models in neuroscience: Part 1 – taking mechanistic abstraction seriously. Preprint at https://arxiv.org/abs/2104.01490.

Chirimuuta, M. 2021. Prediction versus understanding in computationally enhanced neuroscience. *Synthese 199*(1-2): 767–790. https://doi.org/10.1007/s11229-020-02713-0 .

Coveney, P.V., E.R. Dougherty, and R.R. Highfield. 2016. Big data need big theory too. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374*(2080): 20160153. https://doi.org/10.1098/rsta.2016.0153 .

Craver, C.F. 2008. Physical Law and Mechanistic Explanation in the Hodgkin and Huxley Model of the Action Potential. *Philosophy of Science 75*(5): 1022–1033. https://doi.org/10.1086/594543 .

Craver, C.F. and M. Povich. 2017. The Directionality of Distinctively Mathematical Explanations. *Studies in History and Philosophy of Science Part A* 63: 31–38. https://doi.org/10.1016/j.shpsa.2017.04.005 .

Creel, K.A. 2020. Transparency in Complex Computational Systems. *Philosophy of Science 87*(4): 568–589. https://doi.org/10.1086/709729 .

Crick, F. 1958. On protein synthesis. *Symposia of the Society for Experimental Biology* 12: 138–163 .

Cuperus, J.T., B. Groves, A. Kuchina, A.B. Rosenberg, N. Jojic, S. Fields, and G. Seelig. 2017. Deep learning of the regulatory grammar of yeast 5' untranslated regions from 500,000 random sequences. *Genome Research 27*(12): 2015–2024. https://doi.org/10.1101/gr.224964.117 .

De Regt, H.W. 2015. Scientific understanding: Truth or dare? *Synthese 192*(12): 3781–3797. https://doi.org/10.1007/s11229-014-0538-7 .

De Regt, H.W. and D. Dieks. 2005. A Contextual Approach to Scientific Understanding. *Synthese 144*(1): 137–170. https://doi.org/10.1007/s11229-005-5000-4 .

Duede, E. 2022. Instruments, agents, and artificial intelligence: Novel epistemic categories of reliability. *Synthese 200*(6): 491. https://doi.org/10.1007/s11229-022-03975-6 .

Duede, E. 2023. Deep Learning Opacity in Scientific Discovery. *Philosophy of Science*: 1–11. https://doi.org/10.1017/psa.2023.8 .

Hooker, G. and C. Hooker. 2018. Machine Learning and the Future of Realism. *Spontaneous Generations: A Journal for the History and Philosophy of Science 9*(1): 174. https://doi.org/10.4245/sponge.v9i1.27047 .

Issad, T. and C. Malaterre. 2015. Are Dynamic Mechanistic Explanations Still Mechanistic?, In *Explanation in Biology: An Enquiry into the Diversity of Explanatory Patterns in the Life Sciences*, eds. Braillard, P.A. and C. Malaterre, 265–292. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-017-9822-8_12.

Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S.A.A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature 596*(7873): 583–589. https://doi.org/10.1038/s41586-021-03819-2 .

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. https://doi.org/10.1086/714877 .

Knüsel, B. and C. Baumberger. 2020. Understanding climate phenomena with data-driven models. *Studies in History and Philosophy of Science Part A* 84: 46–56. https://doi.org/10.1016/j.shpsa.2020.08.003 .

Krenn, M., R. Pollice, S.Y. Guo, M. Aldeghi, A. Cervera-Lierta, P. Friederich, G. dos Passos Gomes, F. Häse, A. Jinich, A. Nigam, Z. Yao, and A. Aspuru-Guzik. 2022. On scientific understanding with artificial intelligence. *Nature Reviews Physics 4*(12): 761–769. https://doi.org/10.1038/s42254-022-00518-3 .

Lange, M. 2013. What Makes a Scientific Explanation Distinctively Mathematical? *British Journal for the Philosophy of Science 64*(3): 485–511. https://doi.org/10.1093/bjps/axs012 .

Leonelli, S. 2016. *Data-Centric Biology : A Philosophical Study*. Chicago: University of Chicago Press.

Lipton, P. 2009. Understanding Without Explanation, In *Scientific Understanding: Philosophical Perspectives*, eds. de Regt, H.W., S. Leonelli, and K. Eigner, 43–63. Pittsburgh: University of Pittsburgh Press.

López-Rubio, E. and E. Ratti. 2021. Data science and molecular biology: Prediction and mechanistic explanation. *Synthese 198*(4): 3131–3156. https://doi.org/10.1007/s11229-019-02271-0 .

Machamer, P., L. Darden, and C.F. Craver. 2000. Thinking about Mechanisms. *Philosophy of Science 67*(1): 1–25 .

McGeary, S.E., K.S. Lin, C.Y. Shi, T.M. Pham, N. Bisaria, G.M. Kelley, and D.P. Bartel. 2019. The biochemical basis of microRNA targeting efficacy. *Science 366*(6472): eaav1741. https://doi.org/10.1126/science.aav1741 .

Morange, M. and M. Cobb. 2020. *The Black Box of Biology: A History of the Molecular Revolution*. Cambridge, MA: Harvard University Press.

O'Malley, M.A. and O.S. Soyer. 2012. The roles of integration in molecular systems biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 43*(1): 58–68. https://doi.org/10.1016/j.shpsc.2011.10.006 .

Ourmazd, A. 2020. Science in the age of machine learning. *Nature Reviews Physics 2*(7): 342–343. https://doi.org/10.1038/s42254-020-0191-7 .

O'Malley, M.A., K.C. Elliott, and R.M. Burian. 2010. From genetic to genomic regulation: Iterativity in microRNA research. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 41*(4): 407–417. https://doi.org/10.1016/j.shpsc.2010.10.011 .

Pietsch, W. 2015. Aspects of Theory-Ladenness in Data-Intensive Science. *Philosophy of Science 82*(5): 905–916. https://doi.org/10.1086/683328 .

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence 1*(5): 206–215. https://doi.org/10.1038/s42256-019-0048-x .

Räz, T. 2022a. ML Interpretability: Simple Isn't Easy. Preprint at https://arxiv.org/abs/2211.13617.

Räz, T. 2022b. Understanding Deep Learning with Statistical Relevance. *Philosophy of Science 89*(1): 20–41. https://doi.org/10.1017/psa.2021.12 .

Räz, T. and C. Beisbart. 2022. The Importance of Understanding Deep Learning. *Erkenntnis*. https://doi.org/10.1007/s10670-022-00605-y.

Soutschek, M., F. Gross, G. Schratt, and P.L. Germain. 2022. scanMiR: A biochemically based toolkit for versatile and efficient microRNA target prediction. *Bioinformatics 38*(9): 2466–2473. https://doi.org/10.1093/bioinformatics/btac110.

Spinney, L. 2022. Are we witnessing the dawn of post-theory science? The Guardian, https://www.theguardian.com/technology/2022/jan/09/are-we-witnessing-the-dawn-of-post-theory-science, accessed on 2023-09-21.

Srećković, S., A. Berber, and N. Filipović. 2022. The Automated Laplacean Demon: How ML Challenges Our Views on Prediction and Explanation. *Minds and Machines 32*(1): 159–183. https://doi.org/10.1007/s11023-021-09575-6.

Stinson, C. 2020. From Implausible Artificial Neurons to Idealized Cognitive Models: Rebooting Philosophy of Artificial Intelligence. *Philosophy of Science 87*(4): 590–611. https://doi.org/10.1086/709730.

Sullivan, E. 2022. Understanding from Machine Learning Models. *The British Journal for the Philosophy of Science 73*(1): 109–133. https://doi.org/10.1093/bjps/axz035.

Watson, D.S. 2022a. Conceptual challenges for interpretable machine learning. *Synthese 200*(2): 65. https://doi.org/10.1007/s11229-022-03485-5.

Watson, D.S. 2022b. Interpretable machine learning for genomics. *Human Genetics 141*(9): 1499–1513. https://doi.org/10.1007/s00439-021-02387-9.

Watson, D.S. and L. Floridi. 2021. The Explanation Game: A Formal Framework for Interpretable Machine Learning, In *Ethics, Governance, and Policies in Artificial Intelligence*, ed. Floridi, L., 185–219. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-81907-1_11.

Weber, M. 2005. *Philosophy of Experimental Biology*. Cambridge: Cambridge University Press.

Whalen, S., R.M. Truty, and K.S. Pollard. 2016. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature Genetics 48*(5): 488–496. https://doi.org/10.1038/ng.3539.

Zednik, C. 2021. Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology 34*(2): 265–288. https://doi.org/10.1007/s13347-019-00382-7.

Zednik, C. and H. Boelsen. 2022. Scientific Exploration and Explainable Artificial Intelligence. *Minds and Machines 32*(1): 219–239. https://doi.org/10.1007/s11023-021-09583-6.

Zerilli, J. 2022. Explaining Machine Learning Decisions. *Philosophy of Science 89*(1): 1–19. https://doi.org/10.1017/psa.2021.13.