Penultimate draft. Accepted for publication in *Studies in the History and Philosophy of Science Part A*

WHAT KIND OF NOVELTIES CAN MACHINE LEARNING POSSIBLY GENERATE? THE CASE OF GENOMICS

Emanuele Ratti[1], University of Notre Dame

Abstract. Machine learning (ML) has been praised as a tool that can advance science and knowledge in radical ways. However, it is not clear exactly how radical are the novelties that ML generates. In this article, I argue that this question can only be answered contextually, because outputs generated by ML have to be evaluated on the basis of the theory of the science to which ML is applied. In particular, I analyze the problem of novelty of ML outputs in the context of molecular biology. In order to do this, I first clarify the nature of the models generated by ML. Next, I distinguish three ways in which a model can be novel (from the weakest to the strongest). Third, I dissect the way ML algorithms work and generate models in molecular biology and genomics. On these bases, I argue that ML is either a tool to identify instances of knowledge already present and codified, or to generate models that are novel in a weak sense. The notable contribution of ML to scientific discovery in the context of biology is that it can aid humans in overcoming potential bias by exploring more systematically the space of possible hypotheses implied by a theory.

Keywords: machine learning; theory pursuing; theory choice; molecular biology; genomics

1. INTRODUCTION

Today's gospel in both specialized and general journals is that big data is revolutionizing the practice of science and its products. Commentators claim that the tools of data science can generate new knowledge that traditional forms of the scientific method (whatever we mean by that) could not possibly do. Machine learning (ML) and deep learning in particular, it is said, are advancing and will advance science in unprecedented ways (Maxmen 2018; Zhou et al 2018; Zhang et al 2017; Schrider and Kern 2017; Angermueller 2016; Sommer and Gerlich 2013). However, it is not clear what kind of knowledge advancement ML is stimulating. In this paper, I will answer to a few questions about the nature of knowledge generated by ML, especially when it is applied to genomics and molecular biology; what kind of knowledge does ML afford in specific contexts? Is such knowledge revolutionary, understood as knowledge that can revise the theory of the field to which ML is applied? What is the role of ML insights in the scientific process?

[1] eratti@nd.edu ; mnl.ratti@gmail.com

I will claim that ML is not a tool to generate new knowledge in a strong sense. Since ML aims at generating models, when I say 'knowledge', I refer to the models generated by ML and their relation with biological knowledge; if ML models are revolutionary, it is because they add something new to the field they are applied (in the case of the present work, both molecular biology and genomics[2]). In other words, ML is a tool to generate either models which can be interpreted as instances of biological knowledge already present and codified (i.e. non-novel knowledge), or models which yield biological interpretation in unexpected directions (i.e. new knowledge). I will distinguish three senses in which a model can be 'novel' or can lead to new biological interpretations (from the weakest to the strongest), and I will show that ML can deliver novel models only in the weakest sense. Moreover, I will show that revisionary models generated by ML have to be complemented with other lines of evidence not just for being accepted, but also for being pursued in the first place.

The choice of limiting my analysis to molecular biology and genomics is motivated by the fact that it is not possible to answer the question of the article *in general for the natural sciences*. 'New knowledge' is novel with respect to a pre-existing theory. Given that each natural science has its own theory structured in peculiar ways, then the question of novelty can be only approached on a case-by-case basis. In this paper I focus on molecular biology and the way its theory is structured.

The structure of the paper is as follows. In Section 1.1, I introduce the scope of ML and I clarify the nature of ML models. In Section 2, I introduce a distinction between novel and non-novel models (2.1), and I further distinguish three senses in which a model can be novel (2.2). In Section 3, I scrutinize the outputs of different modes of learning and I will claim that only novel models in the weakest sense will be accepted. In Section 4, I show that even in a context of theory pursuing, ML models that are revisionary need to be complemented with other lines of evidence because of how epistemically costly are. Finally, I conclude with thoughts on how this article can shed light on the relation between computer science and biology, and automated science.

1.1 Machine Learning and its scope

[2] I assume a continuity between molecular biology and genomics, namely that genomics is a data-rich instance of molecular biology

Before starting the analysis of ML, let me define the scope of this field. The term 'machine learning' was coined by Arthur Samuel in 1959, understood as the "programming of a digital computer to behave in a way which (…) would be described as involving the process of learning" (Samuel 1988, p 335).

Today, ML is a subfield of data science, which can be roughly defined as "the study of the generalizable extraction of knowledge from data" (Dhar 2013, p 64). Knowledge is not meant in any philosophical or technical sense; Dhar refers to patterns in the data. 'Generalizable' means that the patterns detected in data sets can be extracted from similar data sets as well.

ML algorithms extract patterns from a data set starting from a *problem*, which is typically defined as a given set of input variables, a set of output variables, and a data set (called the training set). The aim is to calculate the relation between inputs and outputs (or to calculate the outputs themselves) starting from the inputs and the training set. Usually, the training set contains instances of input-output pairs. This calculation in its simplest form corresponds to a model connecting data points by learning it from the training. In other applications of ML, the algorithm can simply generate a model by grouping similar data points together (i.e. clustering). The algorithm is sharply distinguished from the model – the algorithm is applied to data sets in order to learn a model representing relations between data points. Sometimes practitioners say that an algorithm 'selects' a model (from an abstract space of possible models) given a sample of data.

Consider e-commerce. ML is applied to identify quantitative relations between inputs (consisting of a set of customers' characteristics) and outputs (defined as buying specific products). The problem is composed of an input (i.e. the customers' characteristics) and a set of input-output pairs (i.e. a training data set on customers with specific characteristics buying various products). The solution to the problem is a model depicting the quantitative relation between inputs and outputs – i.e. given the characteristics *x and y* of buyers, there is a probability *p* that such buyers will buy the product *z*.

Different goals are connected to specific modes of learning. The goal of generating a predictive model is connected to one mode of learning in particular, namely *supervised learning*. In this mode of learning, the algorithm learns a function *f(x)=y* from a list of pairs *(x1, y1),(x2, y2)…,* already available (Angermueller et al 2016). In its simplest form, supervised learning works in three phases (Libbrecht and Noble 2015). First, an algorithm and a task in the form of either predicting a certain value or classifying data points into predefined classes are chosen. Next, the

algorithm is provided with a training set. This data set is labelled, in the sense that labels are assigned to data points, where 'labels' are tags specifying what the data point represents. Finally, the algorithm generates a model from the training set which is then applied to an unlabeled set to automatically label it. This mode is called 'supervised' because of the crucial role that labeled data sets play in giving 'instructions' to the algorithm on what to learn. Classification is one way in which supervised learning is used, and it is when we "design a system able to accurately predict the class membership of new objects based on the available features" (Tarca et al 2007). 'Learning' here refers to identifying the relation between data points and the labels, such that the algorithm will be able to assign 'labels' to an unlabeled data set without human supervision. In other words, the algorithm learns how to recognize the data points labelled in such and such a way in new (although similar) contexts. Therefore, classification is when we assign data points to classes, where classes are sets of objects which share some features such that they can be relevant for some purpose. Another way to express this concept, is to say that classification involves learning a conditional distribution for a finite number of classes. A second supervised learning goal is regression, which predicts the value of a variable starting from the value of another variable, where the value is learnt from training data. More precisely, regression is "a problem of predicting a real-valued label (often called a target) given an unlabeled example (Burkov 2019, p 19). As an example, Burkov reports the problem of estimating house price based on certain house features (which are the 'labelled' part of the training set).

There is also a second mode of learning, called *unsupervised learning*. This mode of learning is not necessarily connected to predictive models. An unsupervised algorithm analyzes unlabeled data sets with the goal of discovering the 'structure' of the data set itself. This is done by analyzing data points and grouping them according to some arbitrary measure of similarity or redundancy. This means that the algorithm *clusters* data points in various groups or categories. An example of clustering algorithm is K-means. With this algorithm, we specify a k number of centroids, and then we assign a data point to a cluster based on the distance between the data point and the centroid. It is important to notice that the choice of the number of centroids will make the difference to the way in which the data points will be clustered. Other widely used unsupervised

algorithms (especially in biology) are hierarchical clustering and partitioning around medoids (Tarca et al 2007)[3].

Finally, there is *reinforcement learning*, which learns from positive and negative rewards from the environment. Reinforcement learning is less relevant here because it is used more in engineering than in the natural sciences.

2. WHAT IS 'NOVEL'?

What would it be a good way of understanding ML as a methodology to produce new scientific knowledge that we cannot achieve by other traditional means? First, since we have characterized the task of ML in terms of predictions and discovery of structures in data sets, it would be by saying that ML can produce new scientific knowledge in these forms. Next, we have to associate these goals of ML with biological interpretations. Supervised learning algorithms may provide models which will count as 'novel' if they predict phenomena which are unanticipated given the theory. Similarly, unsupervised learning can generate 'novelty' by partitioning data points in specific ways when the groupings/clustering were not anticipated given the theory.

I will refer to this aspect of 'prediction/clustering + new biological interpretation' as 'novel'. However, this is not enough. Let me qualify this point in a more precise epistemological sense.

2.1 Intuitive conception of novel/non-novel

In philosophy of science, novelty is sometimes referred to predictions, especially in more traditional accounts such as the one found in Popper[4] (1965) or Lakatos[5] (1970). In the case of ML, the intuitive idea of 'novelty' is that, whatever the strategy used by algorithms[6], these tools generate models that challenge what we already know about a particular phenomenon or simply expand what we already know about it. Having such voluminous data sets – combined with

[3] There are also examples of semi-supervised learning algorithms, meaning algorithms requiring on the one hand labelled data sets, but on the other hand they also make use of unlabeled examples

[4] A new theory/hypothesis "must lead to the prediction of phenomena which have not thus far been observed" (1965, pp 241-2)

[5] Novel is defined by Lakatos as "improbable or even impossible in light of previous knowledge" (1970, p 118)

[6] Interestingly, sometimes it seems is not even possible *in principle* to understand how these algorithms obtain the results they obtain (Humphreys 2011; Burrell 2016)

algorithms – allows us to see something that we could have not seen/predicted with whatever methods or conceptual tools we used before. Sometimes what we see through these tools is revisionary of the knowledge of a field.

There is something to emphasize here; when we say that ML produces revisionary knowledge, we assume that ML is applied to fields other than computer science (e.g. physics, biology, chemistry, etc) and that ML somehow modifies the knowledge (in form of a theory) of the field which is applied to. This is not a trivial point; whatever result we obtain with ML, *it should also be evaluated on the basis of the theories and the conceptual apparatus of the science to which ML is applied*. Therefore, when we anticipate a phenomenon by using a theory/model of a field *and* an algorithm of ML, and this phenomenon was not known given the particular theory/model (but not the algorithm) we were using, then the output counts as novel. It is the algorithm that is the *plus* with respect to traditional scientific methodologies (whatever we mean by that). On a similar note, we generate a non-novel output when the model does not challenge what we already know. My senses of 'novel' and 'non-novel' are related not only to the extent to which a model confirms a theory, but most importantly, to the extent a model leads us to *revise* a theory, even though (as I will show in the next section) revision may be also interpreted in the weak sense of 'enlarging the scope' of a theory. However, while non-novel is quite straightforward, there can be several ways in which a model can be 'novel', especially when we think about the different meanings that 'using a theory' may assume.

2.2 Three types of novelty

Douglas and Magnus (2013) distinguish four levels across which scientists make inference, which will be used to specify what can possibly be 'novel' in ML outputs.

Data and phenomena provide the first two levels. Data corresponds roughly to what ML practitioners understand by this term – i.e. *data points*. We may accommodate both Bogen and Woodward's view (1988) and more recent conceptualizations such as Leonelli's relational account (2015) to make sense of ML's conception of data. Phenomena are usually understood as "patterns in the world that are indicated by data" (Douglas and Magnus 2013, p 582) – even here, this is not different from Bogen and Woodward's view (1988). In the case of ML, phenomena may be understood in a similar way. Patterns in the data – regularities – are candidates for being also

indications of phenomena. 'Novelty' (in whatever way we mean it) is strictly connected to the identification of biological phenomena in data sets.

Theory and framework are the third and fourth level respectively. Douglas and Magnus (as others) understand theory as a set of models and/or laws which predicts and explains a broad class of phenomena, while the framework is a set of implicit background assumptions including auxiliary hypotheses and commitments. Since my goal is to analyze the novelty of ML outputs when applied to molecular biology and genomics, I have to discuss what theory and framework are in those fields. In molecular biology and genomics, one way to understand theory might be to think about it as a collection (more or less structured) of models. For instance, the theory of cell biology is the totality of mechanistic models explaining various phenomena in the cell at the level of macromolecules[7].

The framework is what Darden calls a *store of a field* (2006). Assuming a mechanistic perspective, "[f]or a given field at a given time, there is typically a store of established and accepted components out of which mechanisms can be constructed (…) the store also contains accepted modules: organized composites of the established entities and activities" (Darden 2006, p 51). The store will contain all those components that, as molecular biologists, we implicitly use to conceptualize biological phenomena. In this sense, the store stays in the background. There are other epistemic frameworks developed in philosophy of science to make sense of the notion of framework. For example, the store of a field can be considered as a type of *explanatory model* (Longino 1990). Longino uses the term 'explanatory model' to refer to

> "a normative and somewhat general description of the sorts of items that can figure in explanations of a given sort of phenomenon and of the relationships those items can be said to bear to the phenomena being explained (1990, p 134)

An explanatory model is a grammar and a vocabulary that we use to think about the phenomena that we are investigating. The store of field is an explanatory model because it provides exactly that vocabulary and grammar used to think about biological phenomena, interpret results, and informing the way we do experiments. By thinking in molecular biological terms, we are

[7] An example will be the totality of mechanistic models that one can consult in a textbook such as Alberts' biology of the cell (Alberts et al 2007)

necessarily assuming and committing to a specific store of a field/explanatory model. The store of the field and the explanatory model play similar roles; they give scientists a way to talk about the phenomena they are investigating. It is interesting also to point out Longino's emphasis on the normativity of explanatory models. Any contribution to a specific field must conform to the store of that field/explanatory model. This, in turn, echoed Kuhn's parallelism between 'puzzles' and the problems of normal science. Among the several grounds to compare puzzles and normal science, Kuhn argues that to solve puzzles in normal science one needs 'rules' (in a broadened use of the term) "that limit both the nature of acceptable solutions and the steps by which they are to be obtained" (1962, p 38). Explanatory models and stores of a field play a similar role.

However, in molecular biology it is difficult to distinguish sharply between theory and framework. The difficulty lies in the fact that the store functions also as theory in the sense meant by Douglas and Magnus. For instance, the store of molecular biology will contain Type II restriction enzymes, which are very specific in cutting DNA sequences. But having Type II restriction enzymes implies also having a model of restriction-modification systems which explain the abilities of such enzymes – and the model of restriction-modification systems is one model of the theory of bacteria. Therefore, components of the store will count as theoretical in the sense meant by Douglas and Magnus. For this reason, I think that (at least in molecular biology and genomics) we should not distinguish between theory and framework *as if they were two different conceptual entities*. Rather, we should distinguish between *ways of using a theory*.

A significant way to express this idea is to distinguish between theory-driven and theory-informed practices/ways of using a theory (Waters 2007). Theory-driven includes all the practices that somehow put at test a theoretical framework, or where the theoretical framework is used explicitly to predict and/or explain phenomena. Using a theory in this way means conceptualizing the theory in the way Magnus and Douglas think about the theory. On the other hand, theory-informed will cover all the practices where a theoretical framework is used to conceptualize data sets and/or to think about methodologies – this will be the framework. Theory-driven practices also imply theory-informed practices; parts of a theory will be used in the theory-driven sense, others as theory-informed. However, in cases such as exploratory experimentation there can be theory-informed practices without theory-driven practices (Waters 2007; Ratti 2015).

Patterns identified by ML in a biological data set will be conceptualized in terms of the store of the field; theory is used in the theory-informed sense. Patterns may be further scrutinized to see

in which relation they stand to theory in the theory-driven sense – i.e. to see if they might indicate (novel) phenomena. At this point, a model may be novel, where 'novel' can be understood in three main ways, from the weakest to the strongest sense of 'novel':

- The weakest sense of 'novel' (*N1*), i.e. when patterns in the data set (i.e. the model generated by the algorithm) stimulates the development of a biological model already existing in a 'family' of models; ML models in this case point to something worth investigating biologically[8];
- A stronger sense of 'novel' (*N2*), i.e. when the patterns themselves suggest that there is something wrong in the family of models about a specific class of phenomena – for instance that a mechanism or a set of entities were conceived to do *x*, while the patterns suggest that they do $\neg x$[9]. In the sense of N2, ML outputs suggest a revision of biological theory in the theory-driven sense. An example of N2 is provided by the history of the interpretation of the restriction phenomenon in bacteria, i.e. the fact that some bacteriophages grown on one strain of bacteria could not grow equally well in others. At first, the dominant model claimed that characteristics of the invading bacteriophages were responsible for the presence or absence of the restriction. However, Werner Arber provided evidence that it was the opposite, namely that host cells were responsible for blocking the growth of bacteriophages[10];
- Finally, the strongest sense of 'novel' (*N3*), i.e. when the patterns themselves suggest a modification of the theory conceived as the store of the field (i.e. theory-informed sense). This is when we realize that there is something wrong with the theory when it is used to conceptualize phenomena themselves (theory-informed practices). An example of N3 comes from the debate on the fundamental biology of cancer, and whether this can be characterized by appealing either to the tenets of somatic mutation theory or the tissue organization field theory (see Chapters 2,3 and 4 of Bertolaso 2016). These theories make

[8] In the case of molecular biology, investigating the meaning of those patterns is done with different means than the computational ones. It is the case of several cancer genomics screenings, where we discover cancer genes that we know they might be implicated in cancer because of what the proteins they encode do, but we do not know exactly how, and we have to perform more traditional wet-lab experiments to find out that (Ratti 2015)

[9] This is different from saying that something was conceived to do *x*, and then patterns suggest that it does *y*, since it is quite the rule in biology that entities or even full-fledged mechanisms are involved in more than one process

[10] For a comprehensive account of this episode, see Ratti (2018)

use of competing 'explanatory models' (in Longino's sense) to conceptualize cancer and they are in tension with each other. If we revise a theory in the theory-informed sense, then we do a much heavier revision than the others, because it implies a modification of the way we think about the pillars of biology.

It is interesting to note how the three sense of novel can be interpreted. While N2 and N3 are somehow revisionary of a theory, N1 is progressive. In fact, here 'novel' is in the sense that it may lead to the progress of a theory by enlarging its scope, without challenging it. Strictly speaking, N1 can fall under the remits of Kuhn's normal science, since it stimulates the solution of 'puzzles' by starting a process that will lead to the reinforcement of – for instance – the explanatory force of a theory.

Therefore, the claim put forward by ML enthusiasts may be interpreted as follows; *ML generates models which can be novel in the three senses of 'novel' just delineated*. First, we assume that the model represents a phenomenon. Next, this phenomenon has to be understood biologically (i.e. framed in biological theory). Models can foster the development of what we already know in an unexpected direction (N1). In stronger cases, ML leads to revise theory itself in the theory-driven sense (N2), and sometimes even theory understood as the store of the field. In the latter sense (N3), ML is even more revisionary. Given these three meanings, in which sense (if any) does ML generate 'novelty' when applied to other fields (in this case, biology)? I will answer to this question by proceeding as follows. First (Section 3), I will analyze how in the practice of biology ML novelty claims are received from the point of view of theory choice – whether they provide reasons to reject or accept hypotheses in a biological context. Next (Section 4), I will do the same, but from the point of view of theory pursuing – whether ML models provide reasons to biologists to pursue a hypothesis or not. Moreover, since ML can be either supervised or unsupervised, we need to scrutinize the claim about 'novelty' in both cases.

3. NOVELTY IN MACHINE LEARNING AND BIOLOGY

In this section I will analyze supervised and unsupervised approaches to establish in which sense in biology they can generate novelty. I will also mention a few other algorithms used in ML (e.g. algorithms for causal discovery).

3.1 Novel prediction in *supervised learning*

Supervised ML do not generate novelty, neither in the sense specified by N2 nor N3. In supervised learning the entire training of algorithms relies heavily on the theory of the phenomena investigated (thereby undermining N3), and its success is established only by comparing the model generated with the theory itself (thereby undermining N2). Strictly speaking, we do not stimulate biological interpretation in ways that are unexpected given a theory and/or model. Quite the contrary, we look for the same things over and over again – we explicitly do so.

In order to understand exactly what I mean, we should emphasize how the algorithm is trained. In supervised learning, we use labeled data sets. These are labeled in the sense that labels/tags are assigned to data points, and these labels/tags specify what biological entity the data points indicate (a transcription start-site, a gene, etc). Supervised learning algorithms are trained on labeled data sets in order to 'learn' a model of those data points that are tagged as particular biological entities (e.g. a gene, a TSS, a protein, etc). But these labels/tags – and the way they are assigned - show the importance of theoretical commitments in the form of the store of the field. The work of Sabina Leonelli (2016) provides evidence for this claim. Labeling systems (called bio-ontologies[11]) are computationally implemented, but they are based on biological theory. Famously, Leonelli analyzed the labelling-system of Gene Ontology (Ashburner et al 2000), and she emphasizes two aspects; first, bio-ontologies are controlled vocabularies, such that labels/tags are connected to formal definitions of how a biological entity must be understood. Second, datasets must be annotated, meaning that we have to assign specific labels/tags to data points. But in order to assign tags, we must understand what the tag represents biologically, and why the data point should be assigned that tag; in other words, we need a biological understanding of the data set. Therefore, the practice of labelling data sets depends heavily on theory, to the extent that theory – in the form of bio-ontologies – is actually formalized in a robust way[12]. This means that anytime we use a labelled data set to train an algorithm to recognize data points labelled/tagged in specific ways, we are also committing to and assuming a certain theoretical way of thinking about the entity we want to identify in the new data set:

[11] 'Bio-ontologies' are networks of terms, each denoting biological entities or processes.
[12] As Boem (2016) reports, there are Gene Ontology meetings where practitioners discuss how to formalize the definitions of specific biological entities/activities

"Researchers who use bio-ontologies for data retrieval implicitly accept, even if they might not be aware of it, the definition of biological entities and processes contained within bio-ontologies" (Leonelli 2016, p 125)

This shows that practitioners training an algorithm on labeled data sets are assuming a specific store of a field[13], and hence N3 cannot be achieved, *not even in principle*. In other words, we cannot generate something new in the sense of N3, because a formalized store of the field is the *conditio sine qua non* for the functioning of supervised learning itself.

One may point out that the algorithm itself is 'unaware' of the theory-informed practices. However, it is problematic to isolate the algorithm from the whole system of practice (Chang 2014) that is part of, such as the choice of variables, the procedure to acquire data, the labelling system, etc. In fact, the coherence of training an algorithm on specific data sets respond to the need of being in accordance with a specific explanatory model – and hence the store of the field.

Supervised learning does not provide anything new in the sense specified by N2 either. Consider the case of the algorithm trained by Down and Hubbard (2002) to identify transcription start-sites[14] (TSS)[15]. First, the algorithm was trained on data sets containing well-labelled TSSs. Next, it was further trained with a data set of verified mRNA transcripts. It generated a model where TSS are generally located downstream of CpG island, upstream TATA box motifs and other constraints[16]. Interestingly, the model was accepted (and hence the training was considered a

[13] I cannot cover much more than the 'theory-laden' aspects of this process, but there are other aspects (emphasized by Leonelli) that play a fundamental role, such as "the strong interdependence between technical decisions and political, economic and cultural interpretations of what counts as evidence, research, and even knowledge" (Leonelli 2016, p 195).

[14] To simplify a bit, a TSS is where transcription of a gene starts, precisely at the 5'-end of a gene sequence

[15]The example of TSS is representative of genomics, and this is why I have decided to focus on it. Identifying TSS from well-curated databases is something pretty common today. Most studies of genomics are focused on the identification of the genetic basis of phenomena, and they look for entities and activities coalescing around genes. Therefore, looking for TSS is something pretty common and representative of routine activities in genomics. Moreover, TSSs have also been characterized in detail, so that labeling a data set for TSS is not particularly challenging (though it might be time consuming).

[16] TSS are usually associated with TATA boxes. These are sequences of Thymine-Adenine-Thymine-Adenine where a general transcription factor binds, and it is supposed to 'indicate' where transcription of DNA into messenger RNA should approximately begin, since it is located 25 nucleotides upstream TSS. The idea is that anytime there is a TATA box (but other signaling sequences as well), it is likely to have a TSS in the proximity. But we know much more about TSS, for instance that they are located downstream of CpG islands.

success) because it was in accordance with what we know about TSS. This means that N2 does not apply, because we would not accept the final model if it were in tension with what we know about the phenomenon. The authors admit it when they list the features of the final model, and they recognize that both the TATAAA motif and the CpG island enrichment are aspects widely recognized in TSS – and they imply that this 'external' consistency is another reason to believe that the model is good. TSSs have been characterized as being related to all these aspects, and data points are labelled as TSSs because a biologist recognizes them as such. Moreover, the further training with verified mRNA transcripts revealed the importance of the assumption of a connection between TSS and mRNA transcripts which is part of the theory. It looks like they are suggesting that the model made sense also because of these widely shared aspects. Therefore, the whole practice of training an algorithm includes and assumes the knowledge that we have about TSS – we would not have labelled data sets with TSS if we did not have expectations on what to find. In addition to the more technical validation in training, there is also this more 'theoretical' and informal validation, which is basically the comparison of the model with what we know about the phenomenon modeled (see Figure 1 for examples about TSS and genes).

But the impossibility of achieving N2 is more in practice than in principle. This is because *in principle* the algorithm can elaborate a model which is in tension with theory in a theory-driven sense. However, N2 is not achieved *in practice*, because N2 is not even relevant in the context of supervised learning. In fact, supervised learning does not really respond to a need of novelty in the first place. The practice of automated annotation by training a supervised algorithm is not a practice devoted to discovery – in the typical way of discovering new or revolutionary theoretical things. Actually, it is a practice which is in itself aimed at identifying things we already know. We train the algorithm on labelled data sets because we think that we can automate the process of identifying things that we have already well characterized – if we have a labeling system in place, it means not only that a biological entity is well characterized, but also that its definition is formalized. The clear advantage of using a ML supervised algorithm is that a biologist will not waste time in manually recognizing TSSs, and she can *trust* the algorithm because it is trained on data sets which are characterized and structured in light of the right theory. Moreover, as seen in the example of TSS, the adequacy of the model is also tested against what we know about the particular phenomenon (that, again, has been characterized clearly and in detail). The success of a training is

measured by how close the algorithmic model is to the corresponding model in biological theory understood in a theory-driven sense.

To sum up, supervised learning is pervaded by *non-novelty*. Supervised algorithms identify (in this case) TSSs that we, as humans, would expect to find by ourselves if we were as fast and precise in calculation as computers are. We cannot generate something novel in the sense of N3, not even *in principle*, because the framework is assumed in the labelled data set. Moreover, we do not even look for something novel in the sense of N2; algorithms are used to identify something we already know. Supervised learning, among the other things, responds to a need of automation of scientific practice, in the sense that we need is the computational power of ML to identify what we already know in a large data set which we cannot scrutinize by ourselves.

There is a sense of novelty though; if we train an algorithm to recognize TSS, then we identify in new data sets *new instances* of TSS. Imagine a team of biologists working in a lab devoted to whole-genome sequencing of a particular species, say mice. Once genomes have been sequenced, the team starts various computational analyses, and they will apply different algorithms to detect different things. Among these, TSS may be important. We identify new TSS, but we do not discover anything new about TSS that any textbook of cell biology would not mention; identifying instances of what we already know is not a sense of 'novel' that is disruptive and groundbreaking in the way ML practitioners seem to imply when they show the abilities of their tools.

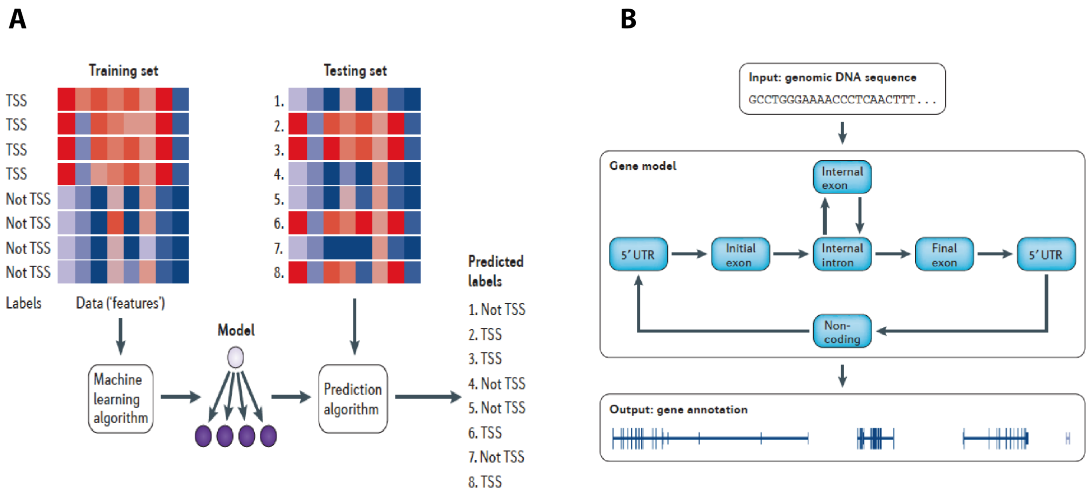

Figure 1. Supervised algorithm in action. (A) supervised algorithm to identify transcription start sites (TSS). (B) supervised learning algorithm for gene finding requiring as inputs a training set of labelled DNA sequences specifying the main features of the structure of genes. Both images taken from (Libbrecht and Noble 2015)

3.2 Novel prediction in *unsupervised learning*

In the case of unsupervised learning, establishing the nature of outputs is trickier.

In unsupervised learning, an algorithm is provided with an unlabeled data set *and a number of unspecified labels* which will be evaluated on the basis of how well they fit the data. The idea is that we want to identify classes of data points, and we do this without having previous examples of this grouping. As emphasized above, one popular unsupervised algorithm is K-means, which defines a *number k* of centroids, and then data points are assigned to the closest centroid. This is the general idea of *clustering*; which data points are similar to each other? How many categories of data points are computationally meaningful in the data set *x*? On this aspect, the main difference between supervised and unsupervised is that, for the modeler, the supervised approach is embedded in a theory from the very beginning (at least from the moment in which data sets are labelled), while in the case of unsupervised learning the modeler approaches the problem merely from a data science perspective without (directly or indirectly) imposing (at least not at the beginning) any theoretical (in the sense of biological) interpretation, and she just uses the algorithm to cluster together similar data points. Therefore, one may say that unsupervised learning may generate novelties because of the absence of labeled data sets and the absence of biological interpretation.

But even though in unsupervised learning there are uninterpreted labels, interpretation comes sooner than you think. Unsupervised algorithms will generate a model of the data by clustering data points according to a number of labels. However, when the output is generated, the labels have to be interpreted *biologically* and not merely computationally, i.e. we have to assign a biological concept to the label. Let's make another example to explain this. Let's say, for instance, that we want to interpret a heterogeneous collection of data from ENCODE[17] (Germain et al 2014; Libbrecht and Nobel 2015) in order to understand how the genome is transcribed. This is a common theme in contemporary genomics. Let's say that we focus on chromatin data (Hoffman et al 2012), as many do. Chromatin is a complex structure of DNA and proteins that forms chromosome. Its function is to efficiently and properly 'package' DNA into a small volume.

[17] I consider the ENCODE consortium example as important and representative; even though ML techniques have been used in biology since the 1990s (see Gillies 1996), ENCODE was instrumental in promoting a more radical and widespread use of them in the biological sciences, especially because of the overwhelming data sets that it generated and that needed to be handled computationally.

Because of the fact that DNA is 'packaged' into this structure, chromatin must open for gene expression to happen; by studying chromatin, we study gene expression as well. Hoffman and colleagues (2012) use unsupervised learning to analyze chromatin data from ENCODE project; they clustered data points around a fixed number of labels (i.e. 25) and *next* they interpreted such labels in a biological way. They say explicitly that the number of labels was established arbitrarily in order "to remain interpretable by biologists" (p 474) and that functional categories (i.e. biological well-known categories) were later assigned. For instance, the algorithm 'rediscovered' protein-coding genes, their chromatin patterns, and other well-characterized biological processes. Therefore, models will not be 'novel' in the sense of N3 because the labels will be interpreted in light of background knowledge or, to use a terminology introduced above, *labels are interpreted by making use of biological theory in a theory-informed sense*. Since we use the theory in a theory-informed sense to interpret the empty labels themselves, we cannot even *in principle* revise the store of the field, because *the store of the field is the condition of possibility of our interpretation*.

Similar considerations apply to N2, even though it is not an 'in-principle' argument. Because the categories we use for interpreting data sets are the ones forming the backbone of the theory both in a theory-driven and a theory-informed sense, it seems unreasonable to think that they might find something like 'the patterns suggest that y does $\neg x$ while before we thought it did x'. With unsupervised learning, we fit the data into theory. To stress further this point, consider that usually numerical solutions are evaluated on the basis of how well they fit the theory, especially in a theory-driven sense. An example comes from cancer genomics[18]. The aim of cancer genomics is to characterize the genetic basis of cancer by focusing on genetic mutations and/or the way chromosomes are disrupted. The approach is brute-force and it consists in sequencing either the exomes (i.e. the totally of exons in a genome) or the whole-genomes of cancer samples to identify non-synonymous mutations[19] in cancer genes (i.e. genes involved in the development of cancer). Algorithms used to identify cancer genes worked under the assumptions that a gene's mutation rate being higher than a certain threshold is a proxy for the gene being a cancer gene. This is based on an evolutionary argument saying that since mutations fostering cancer confer a

[18]This field is another paradigmatic case of the way ML methodologies are applied to biology – because of the amount of data cancer genomics has been generating for more than a decade, there has been a push and investments in computational technologies to make sense of those data sets.

[19] Non-synonymous mutations are mutations that alter the amino acids sequence of a protein. Sometimes mutations are synonymous because the so-called 'genetic code' is redundant

growth advantage to cancer cells (i.e. cancer cells proliferate without control), then they should be positively selected, and hence they will be detected more often than passenger mutations (Ratti 2016). Algorithms have worked under the assumptions of identifying genes mutated at a high rate[20]. Biologists were expecting that, by using more voluminous data sets, both sensitivity (i.e. identifying true cancer genes) and specificity (i.e. distinguishing passenger from driver mutations) of genomics algorithms would both increase. However, "the list of apparently significant cancer-associated genes grows rapidly and implausibly" (Lawrence et al 2013, p 214), *where 'implausibly' has a biological meaning and not a computational one*. The problem was that "many of the genes seem highly suspicious on the basis of their biological function or genomic properties" (p 214). Certain genes that, because of their characteristics (i.e. because of biological theory), are not likely to be involved in cancer, were actually *involved* according to the analysis. At this point, ML could have really revised theory in the N2 sense, but *its results were rejected*. Lawrence and colleagues had to recognize that results of algorithms could make sense numerically, but then they had to be discarded *because theoretically they were spurious and suspicious*. Some genes may not simply 'code' for proteins that play a significant role in a specific tumor; the algorithm may target very long genes, which you will expect to be more mutated than others, because when the genome is replicated there is a chance of 'replication' mistakes which are usually categorized as 'mutations'. These are all considerations based on biology, not on ML. This is not an isolated example, since this article has stimulated revisions of algorithms used in cancer genomics to include biological criteria when numerical solutions are in tensions with current biological knowledge. Complementing computational analysis with other computational analyses that try to connect outputs with concepts belonging to biological theories (e.g. by doing Gene-ontology enrichments[21]) is *evidence that ML results are always interpreted (and their validity evaluated) on the basis of biological criteria*. The fact that field-centered theoretical interpretation (in this case, biological interpretation) dictates the acceptance of the outputs of an algorithm also tells much about the collaborative nature of any ML analysis. However, I cannot honestly say if this is going

[20] The very concept of 'mutation rate' is strictly biological

[21] Gene-ontology term enrichment is when we assign to a set of genes to terms belonging to Gene Ontology system of classification, thereby providing an idea of the processes in which the genes have been found to be involved in. This is a way to complement statistical analysis with biological interpretation

to be the rule also in the future[22] . The examples I have described are based on the specificities of genomics and molecular biology, in the sense that this field has a particular way of interpreting results which differs considerably from the way ML practitioners interpret their own results. Since ML is instrumental to biology, it seems reasonable to think that the biological interpretation will come first, but I cannot say if in the future ML will have a more substantial role[23].

At this point, one may think that ML – both in the supervised and unsupervised version – provides merely non-novel outputs. However, this is not exact; outputs in both supervised unsupervised learning are not completely non-novel either. There is a sense in which supervised and unsupervised ML may provide something new, but it is more in a 'progressive' sense rather than 'revisionary'.

3.3 ML as a way to amend specific human biases

The argument so far is that ML does not generate outputs which are new in the sense of N3 and N2. Let me recapitulate this.

There is no new knowledge in the sense of N3 and N2 because what ML identifies are instances of an already codified knowledge. In the supervised learning case, algorithms literally learn how to identify instances of something that is contained in an explicit form in the training data sets, which are labeled by relying heavily on biological theory in a theory-informed sense. Moreover, a mechanistic model as specified by biological theory (in the theory-driven sense) is used to evaluate the adequacy of the model itself. Finally, supervised learning is linked to a need of automating the identification of well-established entities, rather than the discovery of new entities, and hence N2 and N3 are not even relevant. In the unsupervised learning case, once a number of labels correctly clusters data points, the results will be interpreted in light of the background theory of biology– in other words, the formerly uninterpreted labels have to be interpreted by means of already known biological concepts. Moreover, as in the supervised case, patterns are interpreted by comparing them to biological theory in a theory-driven sense, where theory dictates the validity of patterns.

However, ML does not generate just obvious outputs either – it can generate outputs that are *novel in the sense of N1*. There is a class of cases which exemplifies N1 perfectly. Garraway

[22] This is why N2 is not impossible in principle, but only in practice
[23] I thank a reviewer for encouraging me to emphasize this aspect

and Lander (2013) in a seminal review wonder whether a cancer genomics approach – that is, a data-intensive approach which can be both supervised and unsupervised – can discover new cancer genes, where 'new' would be defined as novel with respect to the list of cancer genes we already have. The answer, they say, is *yes*.

In a famous paper, Parsons and colleagues (2008) sequenced more than 20,000 genes in several samples of glioblastoma multiforme and, by means of computational analysis, found that a gene involved in cell metabolism (*IDH1*) was recurrently mutated. This led researchers to recognize a possible causal connection between genes governing cell metabolism and cancer. This work has been important to stimulate new research in cancer genomics investigating cell metabolism and its connections to cancer. But if you think about the very nature of cell metabolism – as it is likely that Parsons and colleagues did -, then the connection between this and the hallmarks of cancer does not seem surprising. In fact, it is very likely that metabolism of tumor cells may be altered in order to modify and increase the way cells acquire necessary nutrients, or to assign nutrients to pathways contributing to the hallmarks of cancer. Given the nature of tumors, and the nature of metabolism, this is all very reasonable. This is also likely to be the reason why the correlation was not discarded as spurious. While Garraway and Lander say that evidence pointing towards the involvement in cancer of genes of cell metabolism was a surprising result, still the connection between metabolism and cancer has been noted well before the advent of cancer genomics (Pavlova and Thompson 2016).

The discovery of *IDH1* is a case of N1 novelty. What computational tools in this case have provided is a way to *explore a theory* in a direction that has not been done in the past. Scientists have always suspected that cell metabolism was involved in cancer because of compelling evidence and the nature of cell metabolism itself. In this case, *computational tools were able to amend a specific cognitive limitation of human agents, namely the limit in the capacity of exploring the space of possible hypotheses*. The limitation is that humans cannot possibly generate and test hypotheses on the cancer-driven nature about each and every gene in the human genome, even though based on what we know about each and every gene there might be thousands that could be relevant to cancer. History of molecular biology has led researchers to start investigation about cancer genes selected with *ad hoc* criteria, such that there was only a handful of genes which have been investigated (Patel et al 2012; Butcher 2003; Weinberg 2014; Ratti 2016) and genes related to cell metabolism were not among the most preferred. Being able to sequence the human genome

and have simultaneously data about all genes is a method to generate predictions that our cognitive limitations could not do right away (Dulbecco 1986; Ratti 2015). When we sequence *all* genes, we are performing an exploratory experiment to overcome our own limitations. Outputs generated are not novel in the sense of N2 and N3, but they are in the sense of N1 – ML methodology in cancer genomics points to a reinterpretation of previous observations (e.g. connections between metabolism and cancer) that for contingent reasons were not explored before. Please note that this sense of novelty does not apply only to unsupervised, but it can be also valid for supervised learning. In the case of the identification of TSS, once the algorithm generates a model of TSS starting from a labelled data set, we may notice an emphasis on aspects of TSSs that were only barely noticed in the literature and which intuitively is in accordance with what we already know. This is not something new in the sense of N2 and N3, but it can be interpreted as a case of N1. These remarks can be interpreted, as already mentioned, in Kuhnian terms – ML provides a way to focus on 'puzzles' that for some contingent reasons were ignored in the past.

	N3 (revision of theory in the theory-informed sense)	**N2 (revision of theory in the theory-driven sense)**	**N1 (model development)**
Supervised Learning	Theory in the theory-informed sense is used to build the algorithm	Theory in the theory-driven sense is absent	The algorithm may point to theoretical developments to be pursued by other means
Unsupervised Learning	Theory in the theory-informed sense is used to interpret uninterpreted labels	Theory in the theory-driven sense can be used to evaluate the quantitative results of algorithms	The algorithm may point to theoretical developments to be pursued by other means

Table 1. Novel predictions in machine learning when applied to biology

3.4 Algorithms for general causal discovery

One may point out that I have not considered the whole spectrum of algorithms usually applied to natural sciences such as biology. While there are many algorithms, and it is difficult to cover them all, still there is a class of algorithms that should be mentioned. I mean those algorithms that are somehow able of causal discovery. There are ML algorithms that infer causal connections among variables from data (Spirtes et al. 2000; Spirtes and Zhang 2016) – by, for instance, establishing that some variables are causes to other variables. However, as shown in (Lopez-Rubio and Ratti 2019), these methods cannot by themselves generate full-fledged biological explanations, and the variables themselves have to be interpreted biologically, as in the other cases of ML just reviewed. We interpret the causal relation between variables by using theory in a theory-informed sense, and hence N3 cannot be achieved. As in the case of unsupervised learning, N2 is possible in principle, though in practice models are evaluated on the basis of their consistency with what we know about a biological phenomenon.

4. NOVELTY, THEORY PURSUING, AND AUTOMATED SCIENCE

In the previous sections, I have claimed that, while in principle ML can generate all three types of novelty, in practice N2 and N3 will be rejected. In the case of supervised learning, N3 is not possible because the data sets used to train the algorithm itself are labeled by relying heavily on biological theory in a theory-informed sense, while the possibility of N2 is made difficult by the fact that theory in a theory-driven sense is used as a benchmark to evaluate the model generated by the algorithm. Similar arguments apply to unsupervised learning. In this section I want to develop these claims by responding to two possible challenges. The first is that my analysis does not shed light on the role of ML in biological discovery. One may say that the claim has been exaggerated, and that ML practitioners may just mean that ML plays an essential role in a long and complicated process. I will describe what the role is. The second challenge comes from the fact my claims seem to imply the impossibility of automated science. While no automation is now on offer, it does not mean that in principle it cannot be offered[24].

4.1 Weakening the demand of ML-based biology

[24] I thank a reviewer to point out this issue

One may say that the ML practitioners have been exaggerating in their claims. We should not think about ML as generating knowledge and theoretical revisions *on its own*, but rather evaluate its potential in an integrated process within the practice of genomics/molecular biology. One can make the claim that ML plays an essential role in *hypothesis/theory pursuing*, rather than hypothesis/theory choice. Let me start with a digression on the relation between computer science and biology, and then I will get to my main claim on the challenge at hand.

In concluding his book on bioinformatics, Hallam Stevens (2013) depicts a suggestive mini-history of biology under the lens of information technologies. Biology 0.0 is the 'pre-informatic' biology before the World War II, while Biology 1.0 is the introduction of the information paradigm (and metaphor) in molecular biology. From the mid-1980s there is the rise of Biology 2.0, where biological entities and processes are represented in and manipulated by computers (which is much of contemporary bioinformatics). Stevens also claims that Biology 2.0 will become Biology 3.0 when it will be coupled more systematically with Web 3.0. In particular, he seems to imply (2013, p 219) that biology and computer science will become the same thing. This can mean several things:

1. The work at the bench will not be required anymore
2. Computer scientists and biologists will have the same training
3. Biological knowledge and computational models will be the same thing.

While 1 is unlikely, and 2 is a real challenge, this article so far has shown that the third possibility is really problematic. Given the relation between computers, humans, and biological interpretation, it will be difficult for computer scientists to be autonomous in analyzing data sets from a disciplinary point of view, in the sense that algorithms and algorithmic models alone could not provide biological meaning. In order to label data sets in a reliable way, a person needs extensive knowledge of the discipline of the data set itself (Leonelli 2016). Even if we use unsupervised learning, still we need to interpret clusters of data points according to the theory of the field. Therefore, the real issue of Biology 2.0 (and 3.0 if ever) is not just a problem of computing power and reliable data sets, but it is also the challenge of how to foster and make more effective collaborative research and knowledge integration. Even if we use tools such as bio-ontologies, no enrichment of data sets can be done without a biologist scrutinizing the significance of associating certain biological concepts to data points.

Molecular biology has been undergoing 'mathematization' for a while because of the use of computational models and tools, but still the theoretical core remains qualitative[25]. This means that there is a precedence of (qualitative) biological interpretation over the results obtained with more formal methods. This takes two forms. First, the mathematization of biology is understood instrumentally, in the sense that biologists use computers and algorithms as convenient instruments to deal with complexity, but "they describe the organism solely in terms of the genetic and biochemical properties that biologists have thus far elucidated" (Keller 2003 p 250). But even if we use algorithms that, *per se*, are not strictly connected to biological theory (as in the case of most algorithms discussed here), still their results have to be discussed and interpreted biologically – and hence biology takes the precedence over mathematics, in the sense that the mechanistic interpretation of the results generated cannot be obtained by using computational tools alone. These considerations suggest that ML – as other computational methods - is only a part of a very intricate multidisciplinary process. Therefore, the demands of ML can be rephrased as follows; rather than generating novelties by itself, ML will provide (by itself) justificatory reasons which are robust enough to pursue and explore N1, N2, and N3 hypotheses.

In the case of N1, this is exactly what happens - *by definition*. Because of its ability to master and analyze significant amount of data, ML will provide to biologists models that can enrich the complicated structure of biological theory. Models provided by ML can help developing already existing models of well-known processes, as the fictional (though realistic) example of TSS shows. If, for instance, a supervised training will lead to the construction of a model that suggests some new features of TSSs, then the model may be likely *pursued* by other (experimental) means to develop it. In this case, ML model by itself provides reasons that are sufficient. Another important instance in which ML has been shown to be important is in (Ratti 2015). Methodologies and tools designed to handle big data sets are used to extend and strengthen eliminative inductive strategies in fields such as cancer genomics. In this context, whole-exome and whole-genome data are collected by means of sequencing technologies. ML, in conjunction with background assumptions, generates a vast set of initial hypotheses starting from the data gathered. These hypotheses are in the simple form of 'the gene x plays a crucial role in cancer'. Because the list is usually awfully long, biologists have to find a way to check in an efficient way which hypothesis

[25] In addition to this, theory in molecular biology is tightly connected to its practical (*viz* experimental) dimension (Keller 2003, pp 261-264)

is worth pursuing, where 'pursuing' means, in the first instance, Gene Ontology enrichment, which eventually will lead to experimental validation. Once we realize the role of ML in the actual pipeline of contemporary biology, we identify what is its role in generating novelties. ML plays a significant role in hypothesis pursuing, namely that it provides reasons why we pursue a hypothesis or a set of hypotheses rather than another. In the case of cancer genomics, ML tools allow us to see more clearly what's in the immense data sets that we gather. In the fictional (though realistic) case of TSS, it allows us to see aspects of our knowledge of TSS that for some reasons were not considered before and that are worth to be pursued.

It is possible, as I have recognized, that ML can also substantially contribute to N2 and N3. However, in the case of N2 and N3 merely deciding to pursue revisionary hypotheses because of ML can be prohibitively costly. Let me explain this idea by conceptualizing the theory of cell biology in an unusual way. This conceptualization is more of an analogy, and it has several limitations (as I will point out), but it captures an important idea. Let's conceptualize cell biology theory by means of network science (Barabasi and Otvai 2004), which offers a way to quantify and model domains represented as networks composed of nodes connected by edges. The topology of a network is important to determine its behavior and performance, and the topology is given, as the name suggests, by a few topological features. For instance, each node has a basic characteristic called *degree*, which measures how many edges a node has with other nodes. In addition to this, there is also the incoming degree, which measures how many links point to the node (this is because edges may have a direction – in this case they are called 'arcs'). Nodes with high degree tend to play a more important role in maintaining network structure (Zhu et al 2007). Another relevant aspect is the clustering coefficient, which measures the degree to which nodes tend to cluster together not randomly. There are also other factors such as degree distribution, shortest path, etc which are less relevant for the analogy. The idea is that, given these characteristics, different types of networks are possible. We can think about a biological theory as a network, where nodes stand for models and/or components of the store of the field, while directed arcs stand for the contribution of a node to another node (i.e. when the content of a node is assumed by the content of another node). For instance, the model of protein synthesis will have many edges directing to all sort of models, because this model is assumed in many other research contexts in molecular biology. We can zoom-in and zoom-out as much as we want: for instance, we can think

about protein synthesis as a cluster of different nodes, including models of how TSS works, splicing, etc.

A biological theory such as cell biology can be represented as a scale-free network rather than other types of networks (e.g. random networks). This is because there are parts (nodes) of the theory (network) that are more important than others. Some components of the store of the field (e.g. genes) or models (e.g. protein synthesis) are necessary to connect other parts of the theory together. By using network science terminology, parts of the theory will be *central hubs* (i.e. highly connected nodes), and this means that if you knock down central hubs, several other parts of the theory will go down too[26]. There is a dose of holism, and we can revise a theory by poking and prodding with different parts of the theory (i.e. nodes) without disrupting everything. ML can be very precise in telling us which parts of the model are in tension with our knowledge. However, when we disrupt highly connected nodes or central hubs, the theory inevitably collapse. Therefore, the moment we start revising important nodes in terms of N2 – by, for instance, knocking down central models of cell biology - and even N3, there is a serious risk that the edifice of a particular theory will collapse. This means that the more the models generated by ML are revisionary, the more they will knock down highly connected hubs, and the more epistemically costly they will be. It is very unlikely that a biologist will pursue a highly risky hypothesis just because a trained algorithm suggested so, especially because pursuing a hypothesis means putting in place a complicated multidisciplinary research project to gather more compelling evidence. Much more than just ML is needed to revise central hubs of a network.

I am not suggesting that it is never rational to pursue risky hypotheses. This claim would need to be supported by an in-depth (formal) epistemic investigation, which it would require one or more articles. What I am saying is that it is good practice in contemporary biology to exclude risky hypotheses (and hence not pursue them) if they are not supported by other lines of evidence as well. This emerges from the biological literature and from more empirically-oriented

[26] The analogy, as I mentioned, has several disanalogies. For instance, here I am measuring the degree of nodes only as 'out-degree' (Zhu et al 2007). If I were to count also the 'in-degree' then the number of hubs will increase, because a model specifying the functioning of a protein will have a lot of inbound arcs coming from the entire network. Here I want only to emphasize that there are basic models and/or concepts that sustain and support big chunks of the network. Another disanalogy is that in network science central hubs guarantee the robustness of the network, in the sense that you have to take down several central hubs to collapse the structure of the network. In my analogy, certain models are just essential to the structure of the network (the theory)

investigations (e.g. Leonelli 2016). For instance, in (Bailey et al 2018), cancer genomicists re-analyze the entire exome data set of TCGA by means of 26 software tools in order to identify new cancer genes and validate old ones. However, despite the apparent robustness of the computational tools (which include ML tools), results were subjected to manual curation, and 45 genes were excluded for extra-computational reasons. Moreover, final results were also subjected to functional (viz. experimental) validation. This means that, from an epistemic point of view, knowledge-based interpretation and methods, as well as experimental validation, in the molecular field take precedence over purely quantitative and statistical outputs.

To sum up, ML by itself is not sufficient to pursue models that are potentially N2, let alone N3[27], and it needs to be complemented by other lines of evidence. These considerations have far reaching consequences, and they are in agreement with the digression at the beginning of this section, namely that the possibilities of a pure *in-silico* biology are rather slim.

4.2 Automating Scientific Discovery

I want to close this article by considering another topic that has been connected to the issues discussed here. Because ML stays at the forefront of AI developments, it came to be seen as a methodology that can somehow automate science (Ratti 2019). In this way, we could bypass biologists' scrutiny, and have tools that generate a variety of models, both novel and non-novel. However, here I will explicitly claim that automated science is not possible, in the sense that even in principle it is problematic.

Ideas on automated science and the logic of discovery have been extensively explored in the past in philosophy (Hempel 1966; Nickles 1980; Laudan 1980). The development of computational sciences has suggested new alliances between AI and philosophical ideas on the logic of discovery in order to create algorithms that can make non-trivial, if not novel, discoveries (see Langley et al 1987 for an overview). However, this line of work explicitly recognizes that novelties such as N3 are not its target. Langley et al (1987) distinguish between scientific theories as *descriptive generalization* (i.e. detection of regularities and invariants) and *laws of qualitative structure*. The former is the object of inquiry of programs they describe, while the latter are scientific theories based on qualitative concepts such as germ, natural selection, or cell. They

[27] In fact, ML is not even necessary

survey some of the programs developed to automate scientific discovery such as DENDRAL, AM, and EURISKO. These programs, while they can make non-trivial discoveries in the sense of descriptive generalizations, all incorporate some domain-specific knowledge which, if applied to the natural sciences, would constitute a store of a field/explanatory model. Even the well-known BACON (in all its versions) only detects "lawfulness in data and extract that lawfulness in the form of equations that fit the data" (Simon 1984, p 251)[28].

However, one may claim that these considerations do not speak against the possibility of automating scientific discovery *tout court*, and hence N2 can still be possible. One may just grant that there is indeed a lot of background knowledge and interpretative (qualitative) work that goes into biological discovery. But why can't we just represent this knowledge, define rules of interpretation, and devise an algorithmic system that is completely automated, of which ML will be a (significant) part? One arguing for this scenario (let's call the 'automated scenario') will have to renounce to N3 right away[29]. But one can still save the possibility of N2 in principle, by saying

[28] These observations also provide an interesting angle to read the novelty claim put forth by ML practitioners. Laudan (1980) claims that the viability of a logic of discovery depends on what the object of science is. When we think about scientific theories we have in mind grand ontological frameworks such as quantum theory or cell biology. These ontological frameworks are composed of sophisticated theoretical entities (which sometimes are arranged in 'laws of qualitative structure', to use Langley's expression). It is very difficult to imagine a way to derive mechanically the idea of the genetic code and genes from pure biochemical data about DNA. If science's objects are sophisticated ontological frameworks, then a logic of discovery is implausible – there are no rules to derive them, as shown by the impossibility of designing programs to construct them. However, Laudan continues, "if what we want to discover are general statements concerning observable regularities, then mechanical rules for generating a universal from one of more of its singular instances are not out of question" (p 178). Laudan further notices that proliferations of logics of discovery happened during times when grand theoretical characterizations were seen as impossible, and only generalizations were sought. Laudan's analysis is useful to understand claims and misunderstandings about novelties of ML. Since what algorithms do is to generate models by identifying patterns in the data (and patterns are after all regularities), then ML practitioners think that it is possible to generate novelties. However, they do not consider that *novelty should be referred not to the patterns*, but rather to the theoretical frameworks to which ML is applied to. These explanatory theories (1) are used to interpret the patterns themselves, and (2) are only elaborated and revised by humans, and no mechanical procedure exists that can be incorporated in the ML pipeline. Therefore, it seems that computer scientists and (in this particular case) biologists have different ideas in the first place on what the adjective 'novel' should be applied in the first place.

[29] This is because the hypothetical scenario assumes that there is already background knowledge – a theory in a theory-informed sense – that is used to inform the design of the algorithmic procedure. In fact, I have noted above that biological knowledge is already formalized for computational purposes through Gene Ontology. This means that whatever procedure we use, Gene Ontology and its labels/categories/concepts will be assumed for the procedure itself, and any revision will necessarily come from an outside perspective. This is exactly what happens. In particular, the way in which new concepts are created and old are modified and/or discarded requires extensive negotiation. There are meetings between different players of Gene

that even if *right now* we do not have an algorithmic procedure of this kind, it does not rule out its possibility in the future. We lack an agreed standard for theory choice in biology that specifies when enough evidence is enough to discard a hypothesis and support another one (as N2 is supposed to do), and yet we have theories of theory choice. We can think about different standards, and it is not entirely crazy to hypothesize that a group of biologist will agree on one, and implement it in the training of an algorithm. But this reasonable conjecture is informed by misleading ideas. It implies that ML is a sort of mechanism that can be initiated just by pushing a button and will inevitably lead to one result. This in turn assumes that, for some reasons, each step of the ML pipeline can be unambiguously determined by what happens at the previous step. However, this is false, because in ML procedures are underdetermined by their context, in the sense that there is no univocal choice that one can do in training an algorithm. In (Ratti 2020; Libbrecht and Noble 2015), it is shown that there is no single answer, among the many steps of ML pipeline, on to how to encode prior knowledge in a given context; how to select relevant features; how to evaluate performances; etc. At each step you can make different choices that are 'good enough' for the peculiar aspects of a context and the background experience and tacit knowledge put forth by the researchers involved. Given that there is no univocal way of going through the steps of the ML pipeline, this means the entire procedure of biological discovery itself cannot be automated[30]. Moreover, different combination of factors composing the ML pipeline can generate similar results, and this will create the challenge of choosing between ways of automating that generate equivalent results in the first place – in this and similar situations, one has to make a judgement call: choosing, on a case-by-case basis, the characteristics that make a standardized automated procedure the best one given a set of contextual factors and goals of the research itself – a process that can hardly be automated.

Ontology (biologists, computational biologists, etc) that are called *content meetings* (Boem 2016; Diehl et al 2007). In these meetings several methods are employed; in order to construct a new ontology for a domain, authoritative textbooks and up-to-date reviews are used to provide higher-level terms. Also, it should be noted that these 'terms' denoting biological complexes and entities are all based on much more basic biological concepts that, in a network-type representation of biological theory, will be highly connected nodes (such as protein, RNA and DNA molecule, etc). In addition to this, there is also a sort of 'bottom-up' approach, which is the negotiation of the definitions of terms between computer scientists (interested in computational consistency and feasibility) and expert domains (most notably biologists).

[30] Here I am just focusing on ML, but it is doubtful that the other phases of 'discovery' in the biological context can be automated and do not suffer from the same problems highlighted for ML

4. CONCLUSION

In this article, I have discussed the role of ML in the process of scientific discovery, and the extent to which it can lead to serious revisions of scientific knowledge. I would like to close this article with a suggestion. At the beginning, I have pointed out how difficult it is to extend any of the claim that I make about the novelty of ML when applied to biology to other natural sciences. This is because the claim of novelty should be evaluated against the theoretical background to which ML is applied to. Therefore, depending on the specificities of the scientific theory, the claim of novelty may change – and hence, novelty should be assessed on a case-by-case basis. But the claims about novelty in ML when applied to molecular biology are facilitated by two features of molecular biological background. First, there is the background of the mechanistic tradition. Next, this background is understood and interpreted in a qualitative way, as emphasized above. We envision biological situations by representing them with diagrams, by drawing analogies with machines, and we think about the mechanistic dynamics in concrete and material ways (pulling, pushing, etc), while the more abstract mathematical dimension is less prominent. Of course, I am not saying that where mathematics is prominent there is no need for interpretation - mathematical models require interpretation too. But it is easier to see why human supervision is needed the most in 'biological' interpretation. I do not see why the claims I have made about novelty of ML when applied to biology, cannot work in a scientific discipline where the two conditions just discussed are realized.

Acknowledgements: I am indebted to James Nguyen, David Teira, Ezequiel Lopez-Rubio, and two anonymous reviewers from SHPS. A few useful suggestions came also from reviewers of another journal

REFERENCES

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2007). *Molecular Biology of the Cell* (5th Edition). Garland Science.

Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*, *12*(7), 878. http://doi.org/10.15252/msb.20156651

Bailey, Matthew H., Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, et al. 2018. "Comprehensive Characterization of

Cancer Driver Genes and Mutations." *Cell* 173 (2): 371–85.e18. doi:10.1016/j.cell.2018.02.060.

Bertolaso, M. (2016). *Philosophy of Cancer*. Dordrecht: Springer.

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 205395171562251. http://doi.org/10.1177/2053951715622512

Butcher, S. P. (2003). Target Discovery and Validation in the Post-Genomic Era. *Neurochem Res*, *28*(2), 367–371.

Darden, L. (2006). *Reasoning in Biological discoveries*. Cambridge, Uk: Cambridge University Press.

Dhar, V. (2013). Data Science and Prediction. *Communications of the ACM*, *56*(12), 64–73. http://doi.org/10.2139/ssrn.2086734

Douglas, H. E. (2009). Reintroducing Prediction to Explanation. *Philosophy of Science*, *76*(4), 444–463. http://doi.org/10.1086/648111

Douglas, H., & Magnus, P. D. (2013). State of the Field: Why novel prediction matters. *Studies in History and Philosophy of Science Part A*, *44*(4), 580–589. http://doi.org/10.1016/j.shpsa.2013.04.001

Down, T. A., Hubbard, T. J. P., Down, T. A., & Hubbard, T. J. P. (2002). Computational Detection and Location of Transcription Start Sites in Mammalian Genomic DNA Computational Detection and Location of Transcription Start Sites in Mammalian Genomic DNA, 458–461. http://doi.org/10.1101/gr.216102

Dulbecco, R. (1986). A Turning Point in cancer research: Sequencing the Human Genome. *Science*.

Garraway, L. a, & Lander, E. S. (2013). Lessons from the cancer genome. *Cell*, *153*(1), 17–37. http://doi.org/10.1016/j.cell.2013.03.002

Germain, P., Ratti, E., & Boem, F. (2014). Junk or Functional DNA? ENCODE and the Function Controversy. *Biology & Philosophy*, *29*(3), 807–831.

Gillies, Donald. (1996). *Artificial Intelligence and Scientific Method*, Oxford University Press

Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., & Noble, W. S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, *9*(5), 473–476. http://doi.org/10.1038/nmeth.1937

Humphreys, P. (2011). Computational science and its effects. In M. Carrier & A. Nordmann (Eds.), *Science in the Context of Application*. Dordrecht: Springer.

Keller, E. F. (2003). *Making Sense of Life: Explaining Biological Development with Models, Metaphors and Machines*. Cambridge, Massachusetts, and London, England: Harvard University Press.

King, R., Whelan, K., Jones, F., Reiser, P., Bryant, C., Muggleton, S., … Oliver, S. (2004). Functional genomic hypothesis generation and experimentation. *Nature*, *427*(January), 247–252. https://doi.org/10.1038/nature02169.1.

Lakatos, I. (1970). Falsification and the Methodology of Scientific research Programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the Growth of Knowledge*. Cambridge, UK: Cambridge University Press.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V, Cibulskis, K., Sivachenko, A., … Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, *499*(7457), 214–8. http://doi.org/10.1038/nature12213

Leonelli, S. (2015). What Counts as Scientific Data? A Relational Framework. *Philosophy of Science*, *82*(5), 810–821.

Leonelli, S. (2016). *Data-centric Biology*. Chicago: University of Chicago Press.

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, *16*(6), 321–332. http://doi.org/10.1038/nrg3920

Lipton, P. (2005). Testing Hypotheses: Prediction and Prejudice. *Science*, *307*(5707), 219–221. http://doi.org/10.1126/science.1103024

Longino, H. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press.

Love, A. C. (2009). Explaining Evolutionary Innovations and Novelties: Criteria of Explanatory Adequacy and Epistemological Prerequisites. *Philosophy of Science*, *75*(5), 874–886. https://doi.org/10.1086/594531

Maxmen, A. (2018). Deep learning sharpens views of cells and genes. *Nature*, *553*, 9–10.

Parsons, D. W., Parsons, D. W., Jones, S., Zhang, X., Lin, J. C., Leary, R. J., … Kinzler, K. W. (2008). An Integrated Genomic Analysis of. *Science*, *321*(September), 1807–1812. http://doi.org/10.1126/science.1164382

Patel, M. N., Halling-Brown, M. D., Tym, J. E., Workman, P., & Al-Lazikani, B. (2012). Objective assessment of cancer genes for drug discovery. *Nature Reviews Drug Discovery*, *12*(1), 35–50. http://doi.org/10.1038/nrd3913

Pavlova, N. N., & Thompson, C. B. (2016). The Emerging Hallmarks of Cancer Metabolism. *Cell Metabolism*, *23*(1), 27–47. http://doi.org/10.1016/j.cmet.2015.12.006

Popper, K. (1965). *Conjectures and Refutations*. London: Routledge and Kegan Paul.

Ratti, E. (2015). Big Data Biology: Between Eliminative Inferences and Exploratory Experiments. *Philosophy of Science*, *82*(2), 198–218.

Ratti, E. (2016). The end of "small biology"? Some thoughts about biomedicine and big science. *Big Data & Society*.

Ratti, E. (2018). "Models of" and "models for": On the relation between mechanistic models and experimental strategies in molecular biology. *British Journal for the Philosophy of Science*.

Ratti, Emanuele. 2020. "Phronesis and Automated Science: The Case of Machine Learning and Biology." In *A Critical Reflection on Automated Science - Will Science Remain Human?*, edited by Marta Bertolaso and Fabio Sterpetti. Springer.

Samuel, A. (1988). Some studies in Machine Learning using the game of checkers. In D. Levy (Ed.), *Computer Games I*. New York: Springer.

Schrider, D. R., & Kern, A. D. (2017). Machine Learning for Population Genetics: A New Paradigm. *Trends in Genetics*. http://doi.org/10.1101/206482

Sommer, C., & Gerlich, D. W. (2013). Machine learning in cell biology – teaching computers to recognize phenotypes. *Journal of Cell Science*, *126*(24), 5529–5539. http://doi.org/10.1242/jcs.123604

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, Prediction, and Search*. Springer-Verlag.

Tarca, A. L., Carey, V. J., Chen, X. wen, Romero, R., & Drăghici, S. (2007). Machine learning and its applications to biology. *PLoS Computational Biology*, *3*(6). https://doi.org/10.1371/journal.pcbi.0030116

Waters, C. K. (2007). The Nature and Context of Exploratory Experimentation. *History and Philosophy of the Life Sciences*, *29*, 1–9.

Weinberg, R. a. (2014). Coming full circle-from endless complexity to simplicity and back again. *Cell*, *157*(1), 267–71. http://doi.org/10.1016/j.cell.2014.03.004

Zhang, L., Tan, J., Han, D., & Zhu, H. (2017). From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discovery Today*, *22*(11), 1680–1685. http://doi.org/10.1016/j.drudis.2017.08.010

Zhou, Z., Tu, J., & Zhu, Z. J. (2018). Advancing the large-scale CCS database for metabolomics and lipidomics at the machine-learning era. *Current Opinion in Chemical Biology*, *42*, 34–41. http://doi.org/10.1016/j.cbpa.2017.10.033