DATA SCIENCE AND MOLECULAR BIOLOGY: PREDICTION AND MECHANISTIC EXPLANATION

Ezequiel López-Rubio[1] [2]
Departamento de Lógica, Historia y Filosofía de la Ciencia
Universidad Nacional de Educación a Distancia (UNED)
Paseo de Senda del Rey 7, 28040 Madrid, Spain
Departamento de Lenguajes y Ciencias de la Computación
Universidad de Málaga (UMA)
Bulevar Louis Pasteur 35, 29071 Málaga, Spain

Emanuele Ratti[3]
Reilly Center for Science, Technology, and Values
Department of Philosophy
University of Notre Dame

Abstract

In the last few years, biologists and computer scientists have claimed that the introduction of data science techniques in molecular biology has changed the characteristics and the aims of typical outputs (i.e. models) of such a discipline. In this paper we will critically examine this claim. First, we identify the received view on models and their aims in molecular biology. Models in molecular biology are mechanistic and explanatory. Next, we identify the scope and aims of data science (machine learning in particular). These lie mainly in the creation of predictive models which performances increase as data set increases. Next, we will identify a tradeoff between predictive and explanatory performances by comparing the features of mechanistic and predictive models. Finally, we show how this *a priori* analysis of machine learning and mechanistic research applies to actual biological practice. This will be done by analyzing the publications of a consortium – The Cancer Genome Atlas - which stands at the forefront in integrating data science and molecular biology. The result will be that biologists

[1] ezeqlr@lcc.uma.es
[2] Authors have contributed equally to this work.
[3] mnl.ratti@gmail.com

have to deal with the tradeoff between explaining and predicting that we have identified, and hence the explanatory force of the 'new' biology is substantially diminished if compared to the 'old' biology. However, this aspect also emphasizes the existence of other research goals which make predictive force independent from explanation.

1. INTRODUCTION

In the last few years, the intersection of data-intensive science and biology have attracted the interest of philosophers, historians and STS scholars (Leonelli 2011; 2016; Strasser 2011; Ratti 2015; Boem and Ratti 2016; Stevens 2013; 2015; 2017)[4], in particular where cognitive strategies and modes of knowledge production are concerned. However, how data science have shaped the desiderata and the aims of bioscience has not received enough philosophical attention. The impression is that data science changed something profound about the epistemic units of analyses, the conceptual tools and the aims of biological science itself. This concern applies in particular to molecular biology (Alberts 2012; Golub 2010; Callebaut 2012), which is the field investigated in this article. In particular, it is not clear which are the effects that employing data science techniques have for the status of typical outputs – like models and their aims - of molecular biology. In this paper, we will identify the changes that data science has stimulated in molecular biology when the characteristics and goals of models and modeling are concerned.

After having identified the received view on models in molecular biology as *mechanistic* and *explanatory* (Section 2), we will identify the aims and scope of data science, and machine learning in particular (Section 3). From this we will be able to draw a comparison between the mechanistic and the machine learning perspective. We will show the existence in machine learning of *a tradeoff between prediction and (mechanistic) explanation*, such that when predictive performances increase, the possibility of elaborating a (mechanistic) explanation necessarily decreases (Section 3.3). This may help to elucidate a common intuition according to which mechanistic models explain in terms of cause-effect relations (represented *via* diagrams), while predictions may be derived also from statistical correlations. Finally, we will see whether our philosophical analysis of machine learning,

[4] It should be emphasized that the effects of the use of computers in biology have been studied well before the studies we refer to. The use of computers and computational models has certainly boosted in the last few decades (see for instance Keller 2002, Chapter 8), but these have been mostly used as 'crutches' and instrumentally to achieve the aims that are traditionally ascribed to biology

(mechanistic) explanation and prediction applies to the practice of biology (Section 4). To achieve such an aim, we will analyze the publications of The Cancer Genome Atlas Consortium (n = 43). This consortium has been one of the first big-science projects at the forefront of the integration between machine learning and biology. The techniques it developed and the results it published have been highly influential for anyone who wants to engage in data science and biology. This means that the analysis of the publications of this consortium may have a significant degree of applicability to the entire field of molecular biology. What this analysis will show is that biologists have to deal with the tradeoff between explanation and prediction that we have identified. This has two consequences. First, it limits the explanatory force of the 'new' biology, such that models in data-intensive biology will hardly be explanatory from a mechanistic point of view. Second, it emphasizes the increasing importance of other non-explanatory aims related to this scientific field.

2. MECHANISTIC MODELS AND EXPLANATION IN MOLECULAR BIOLOGY

In order to understand whether data science has changed the status of models in molecular biology, we should first state clearly the received view on this 'status'.

Molecular biology is a field that investigates the processes and dynamics of biological phenomena at the level of complex organic molecules. Molecular biology aims at making sense of biological phenomena conceived as the result of the activities of those molecules. Complex and less complex organic molecules include nucleotides, polypeptides, lipids, etc while the activities include binding, releasing, or more complex operations as phosphorylating among others. When we say 'make sense' we mean that molecular biologists aim at explaining a specific class of biological phenomena as nothing *but* the activities, interactions and the organization of those macromolecules. By 'organization', we rely on a basic meaning spelled out by Levy and Bechtel (2013), namely *organization as 'causal connectivity'*, in the sense of "an internal division of causal labor whereby different components perform different causal roles" (p. 243). In other words, molecular biology *explains* biological phenomena by *describing* how these are produced by the way macromolecules interact. These descriptions are given in form of mechanistic models. To sum up, it is common in the literature – especially the so-called 'mechanistic philosophy' (Machamer et al 2000) – to state that molecular biology aims at (a) *explaining* phenomena (b) in terms of *mechanistic models*. Therefore, models here are mechanistic models, which

function as explanations. The explanatory dimension is usually interpreted by practitioners as the main goal of molecular biology (Alberts 2012; Weinberg 1985). However, one may say there are a plenty of mechanistic models that are not necessarily explanatory, or they are not used for their explanatory capacity, but in other ways. For instance, it is possible to use an abstract schema in the form of a mechanism sketch as a starting point to investigate a possible mechanism (Craver and Darden 2013, Ch 3 and 5). The schema itself is not explanatory, but it is a very useful resource to uncover the nature of phenomena. Therefore, a mechanistic model to be mechanistic does not need to be explanatory; however, the emphasis on the need "to transform black boxes (...) into gray boxes (...) into glass boxes" (Craver and Darden 2013, p 31) often spelled out in term of 'completeness' of mechanistic description (relatively to the purpose at hand) is an indication that biologists still see good mechanistic models as being explanatory.

Before we go further in the received view on molecular biology, three things need to be specified.

First, appealing to mechanistic explanations in the form of mechanistic descriptions do not only have epistemic motivations. In other words, it is not necessarily the case that mechanistic explanations are superior epistemically to other forms of explanation[5]. Actually, there are also pragmatic reasons for these explanatory standards. Among the many, one prominent aspect is that mechanistic descriptions provide a useful way for human beings to think about control. In molecular biology material manipulations are important, and mechanistic descriptions provide easier ways to think about those manipulations. There are also historical reasons for this preference towards mechanistic models in biology. As Philip Sloan (2000) has emphasized, mechanistic programs in life sciences after the 19th century (and in molecular biology in the 20th century) have been a product "of the incursion of physical sciences into the work of medical faculties" (p 7). While entering into details of these pragmatic and historical aspects is beyond the scope of this work, it is interesting to notice that acceptance of epistemic standards do not have necessarily (or only) epistemic reasons.

Next, while molecular biologists commit to the idea of mechanism, they often have in mind a notion of mechanism that may not be as precise as the one specified by mechanistic philosophers. However, our impression is that the word 'mechanism' is not merely a term which replaces the word 'cause'. Biologists are not only interested in identifying causes and

[5] We do not commit to any particular thesis on this aspect.

causally relevant entities. Appealing to mechanisms means that we also try to find out a more fine-grained causal connectivity between causally relevant entities.

Finally, we must spend a few words on models. We will refer to models in two ways throughout this work. On one hand, the level of abstraction of a model is the degree of detail in the representation of the target (real) system. For example, one may have a model of the circulatory system where the smallest elements are blood vessels. This would be regarded as a high level model, as compared to a model of the circulatory system where the smallest elements are the red and white cells. Therefore, a model can be high or low level, also called coarse grained versus fine grained, as compared to other models, depending on the smallest (atomic) elements that it considers (Gerlee & Lundh, 2016, pages 32-33). On the other hand, the size of a model is the number of elements that it contains (Gerlee & Lundh, 2016, page 48, simple versus complex models, also page 73). For example, in a model of the nervous system where the smallest elements are the neurons, the system size could be the number of neurons. That is, a large model is one with many atomic elements, irrespective of its level of abstraction (high or low). In this work, we are interested in comparing handcrafted, mechanistic, small size models versus automatically generated, machine learned, large size models. We do not address the distinction between high level (systemic) versus low level (local) models.

2.1 Mechanistic models and desiderata of biological explanations

A mechanism is usually understood as "entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (Machamer et al 2000, p 3). More succinctly, a mechanism "is a structure performing a function in virtue of its component parts, component operations, and their organization" (Bechtel and Abrahamsen 2005, p 423). Entities figuring in a mechanistic model are those taken to be responsible for the production of the phenomenon under scrutiny. Moreover, activities or operations may be organized simply in terms of temporal sequence, even though "biological mechanisms tend to exhibit more complex forms of organization" (Bechtel and Abrahamsen 2005, p 424). In addition, mechanisms may involve more than one level of organization.

Outlining the dynamics and the organization of parts and activities of a mechanism is, in molecular biology, providing an explanation of the phenomenon one is investigating.

There is a consensus according to which "to give a description[6] of a mechanism for a phenomenon is to explain that phenomenon, i.e., to explain how it was produced'' (Machamer et al 2000, p 3).

To simplify a bit[7], a mechanistic model is explanatory when it describes the causal structure of the world, namely how a phenomenon is produced by *constitutively causally relevant entities* (Craver 2007; Kaplan and Craver 2011). Entities that are supposed to be causally relevant must be clearly identified (i.e. they must be real) and their organization – in the sense of 'causal connectivity' – must be clearly spelled out[8]. It is important to emphasize that being explanatory is not a yes/no issue, but rather it is by degree; the depth of explanatory power of mechanistic models will vary a lot, depending on the circumstances.

An important indication of the 'depth' of explanatory mechanistic models is their being predictive, in the sense of "the ability to answers questions about how the mechanism would behave if different parts of the mechanism were to work differently, or if they were to break" (Craver and Darden 2013, p 91). If the model is explanatory, then we must be able to anticipate what would happen if things were slightly different. These predictions are usually qualitative, and they can be corroborated through interventionist strategies. Importantly, interventions must be concrete, not only possible; the predictive force (in the qualitative sense) of a mechanistic model is to be measured by material and concrete interventions - interventions that we can do right now. If a mechanistic model affords predictions that hold in the real world, then clearly the model itself is on the right track to more explanatory power. However, being predictive is only a necessary condition for being explanatory - in fact, we can imagine plenty of examples of very uninformative and general predictions holding true for very abstract and general mechanistic models. But still being predictive is necessary - if predictions do not hold, then it is likely that we just got the phenomenon wrong.

2.2 Understanding and explaining

[6] 'Description' does not mean necessarily a linguistic description; actually, the preferred means for expressing mechanistic description are diagrams (see Bechtel and Abrahamsen 2005 for preliminary arguments on this matter)

[7] It is not necessary to recall in detail the complex structure of mechanistic explanations; for a full-fledged account of this issue, see (Craver 2007)

[8] There is an interesting debate on what 'relevant' might possibly mean and the difference between leaving out relevant details from the model (i.e. incomplete models) and abstracting from irrelevant details (i.e. abstraction). See in particular (Levy 2014; Bechtel and Levy 2014; Nathan and Love 2015)

While there are many accounts of mechanistic explanations and desiderata for mechanistic models, the remark we just did about organization and richness of details may not apply to all of them. Here we specifically focus on Craver's and similar accounts.

An important aspect of mechanistic explanations that needs to be flagged (especially for the present work) is the fact that mechanistic models must be *intelligible* in order to be explanatory. Biologists *elaborate* such descriptions, and how entities and activities are organized should be clearly discerned by modelers, especially because modelers are *prior* to the model. For a model to exist modelers have to build it, and to elaborate it, modelers have to understand the model itself. This issue is raised here and there in the literature, but it has been rarely tackled properly. A crucial example here is Craver (2006). He mentions this problem when he talks about the notion of 'complete description of a mechanism'. In particular, he says that "such descriptions [i.e. complete descriptions of mechanisms] would include so many potential factors that they would be unwieldy for the purpose of prediction and control and *utterly unilluminating to human beings*" (p 360, emphasis added). In other words, Craver is saying that too complex mechanistic models would not explain, because such a complex model would be *non-intelligible* (though he does not use this exact word). The process of building a mechanistic model where too many details appeared would be unsuccessful, because the modeler would be unable to find out the causal structure and connectivity between components. Therefore, too many details are likely to prevent the construction of mechanistic models. This is why Craver praises the ability of modelers to understand which details should be taken into account. Hence, it seems there is a connection between the ability to explain and the ability to understand models.

2.2.1 'Build-it' test and intelligible models

Let's explore Craver's intuitions. What is the relation between (mechanistic) explanation and understanding? The debate on scientific understanding has experienced a significant growth in the last few years. Due to lack of space, it is not possible to summarize the main themes of this debate. However, we can draw some distinctions that are important for the topic at hand. First, we should distinguish between *understanding of natural phenomena*, and *understanding of the models about phenomena* (Rice 2016). One can understand a model or a theory about *x*, but if the model/theory for some reasons is not adequate, then one does not understand *x*, as in the case of phlogiston theory and the phenomenon of combustion. While most of the debate on understanding tries to flesh out the

notion in the former sense, here we are interested in the latter. In particular, we make use of de Regt's framework (2009; 2015; 2017), though modified according to the present context.

De Regt aims at deciphering the epistemic importance of what he calls *pragmatic understanding*, i.e. the *intelligibility of a theory*, defined as being able *to use* the theory[9]. Since he grounds his analysis in the framework of models as mediators between theories and the world provided by Morgan and Morrison (1999), according to his perspective a theory is being used especially for the construction of models. Here we avoid any reference to theories, and we focus on models.

When it comes to 'mechanistic models', their intelligibility is related to the use we make of them. Being able to use a model covers a plethora of things that a scientist can do with a model, and in particular what Craver and Darden (2013) call the *build-it test*. Mechanistic models have been understood, among the other things, as 'recipes for construction'. Recipes describe a set of operations performed on certain ingredients in a way that they will produce a specific thing (e.g. a cake). Mechanistic models afford something similar; they describe how specific entities, if interacting in a specific and organized way, can result in a specific phenomenon. Therefore, Craver and Darden point out that the *successful* ability to modify at will a specific experimental system on the basis of the 'instructions' provided by mechanistic models can be interpreted as a sign that the mechanistic model is somehow describing the phenomenon in a plausible way, i.e. the model explains. Therefore, by 'building' the phenomenon according to the model, we understand if we have an adequate explanation. This is especially important when it comes to models that only highlight that some entities are causally relevant, but without specifying how the entities influence each other to produce the phenomenon; a *how-possibly* model including only a catalogue of entities and activities is not enough, because we would not have instructions whatsoever as to how to make them interacting. Similarly, a *how-possibly* model depicting in the right way the organization of the mechanism, but failing to identify the relevant entities and activities will be useless as well, because we would not know which entities and activities need to be modified.

But the test is not useful only as a 'confirmation tool'. Actually, it is very useful also when we elaborate the model itself. By doing the *build-it test*, we obtain useful hints as to how we are on the right path - this is related to aspect about prediction highlighted before,

[9] de Regt's characterization of understanding is *much more rich* than this, but for the purpose at hand this definition is sufficient

and the concrete and material interventions required. But in order to elaborate such a test, the model needs to be *intelligible*. By developing Craver's concerns (2006) about models that are too detailed, if in the sketch of the mechanism there are too many entities and activities, then it becomes really complex for the scientist to find out how they are organized or, to use Levy and Bechtel's jargon (2013), it is very difficult to uncover *causal connectivity*. How do we start to stimulate or inhibit entities if we have too many of them? Which entity is to be ascribed causal relevance if we have a long list of potential ones? This means that if we do not strike the right balance between relevant entities and activities and the number of variables that the human mind can deal with, then it is unlikely that the elaboration of a mechanistic model will be successful. This means that too complex (in terms of number of variables, i.e. entities) how-possibly models are unlikely to be turned into mechanistic explanations. To use the terminology introduced above, if the size of the mechanistic model is overwhelming, it is unlikely to become explanatory - it will be remain a how-possibly model[10]. This argument has been spelled out in greater details in another work we have co-authored (Ratti & López-Rubio 2018). In that article, we translate de Regt's account of intelligibility in the mechanistic context, and by relying on studies of cognitive psychology, we make the argument that too complex models cannot be possibly intelligible because of cognitive limits of human agents. They define complexity as a function of the number of model components included in the actual model. Too complex models are not intelligible, and they cannot be turned into explanations. But one may rely on more liberal accounts of mechanistic models and explanations, thereby making the case that very complex models can be somehow explanatory. However, if we are too liberal in saying what mechanistic models are explanatory and what not - and even what counts as a mechanistic model in the first place - , then we may lose the very reasons why we needed a mechanistic account for models in the first place.

To sum up, mechanistic models have to *intelligible* to the modeler in order to be explanatory, where by 'intelligible' we mean de Regt's preliminary notion of pragmatic understanding, but remodeled for the present context, i.e. *understanding a model is being*

[10] Please note that it does not mean that large models cannot be used to draw general predictions that can be also verified experimentally. For instance, if you model protein-protein interactions with network science, you will obtain a very large model that cannot be turned into an explanation - it is impossible to draw the exact causal narrative connecting all the entities. However, network science tools identify central hub, and one can draw the very general prediction that, if I knock-down a central hub, then the network - and hence the biological phenomenon - will be disrupted. However, this prediction does not help any researcher in elaborating a mechanistic explanation

able to successfully use it. If we consider such a test as a model-developing tool, there are important consequences. In order to transform a how-possibly model into a how-actually model (or at least a how-plausibly), we need to be able to experimentally stimulate (e.g. inhibitory and excitatory strategies, see Bechtel and Richardson 2010) the phenomenon under investigation, and we do this on the basis of the strategies suggested by the model itself. Therefore, if the size of a *how-possibly* or *how-plausibly* model is too big (e.g. too many components) – and hence it is unintelligible – then there is no way to successfully turning it into an explanatory model which specifies how the process occurs.

3. DATA SCIENCE, MACHINE LEARNING AND THE VALUE OF PREDICTIONS

In the previous section we have identified the characteristics and aims of models and modeling in molecular biology. We have emphasized that molecular biologists look for explanations of biological phenomena, where explanations have to be understood as mechanistic models. Such descriptions stress the causal structure/connectivity holding between components producing the phenomenon we want to explain. We have also emphasized that in order to build explanatory models and to 'test' the adequacy of models, these have to be *intelligible* to the user, where this is measured by the ability of the user to use the model in various ways. This means that explaining and understanding are strictly related, at least in mechanistic models. Therefore, the question about the relation between data science and biology is exactly the question of whether data science modifies the picture just sketched. In order to understand this, in this section we introduce data science, its aims and what motivates them.

3.1 Data Science, algorithmic models and machine learning

Dhar defines data science as "the study of the generalizable extraction of knowledge from data" (2013, p. 64). In this definition, *data* is a set of samples, such that each sample is a realization of the variables whose joint probability distribution underlies the observations. Also, *generalizable* means that the patterns automatically extracted from the available samples are expected to occur in other samples of the same process under study. *Knowledge* is not meant in any philosophical sense. Since the idea is to extract patterns between variables, knowledge here simply refers to such patterns.

In data science, patterns are understood in terms of predictions, and they are extracted from samples starting from a *problem*. For the purpose of this work, *a problem* is a set of *input variables* (available at the time that the prediction must be delivered), a set of *output variables* (to be predicted, and hence not available at prediction time), a set of samples (previously observed input-output pairs), and a set of *real-world situations* where the dependency among inputs and outputs can be assumed to be the same[11]. For example, evaluating a mortgage application from a potential borrower is a problem. The inputs could be the address, year built and price of the real estate property, whether the applicant has previously succeeded in repaying a previous loan, her annual income, etc. The outputs could be whether money should be lent to the applicant, and if so, the maximum amount, for how many years and the interest rate. *Data science aims to discover the quantitative relations between inputs and outputs that can possibly hold also in the set of real-world situations assumed to be similar to the one depicted in the samples*. Such relations between inputs and outputs – in the form of a *predictive model* - are obtained by elaborating and applying algorithms. For our purposes, *an algorithm* is a sequence of actions such that, given a certain kind of input, calculates an output in a finite time, after the execution of a finite number of steps using a finite amount of memory[12]. *Prediction* here is the computation of the values of the output variables for an input whose associated output is unknown (though, as we will see, model construction by algorithms is based on cases where similar pairs of input-output are available). In what follows, 'prediction' will have this meaning unless specified otherwise.

Data science may include several disciplines. Here we are interested in *machine learning* (ML). This is the discipline that is central to the so-called 'data revolution' or 'data deluge' (Leonelli, 2012, p. 3). ML enables the automated construction of predictive models, which is the reason of its central role. Very broadly, ML is a branch of data science devoted to the development of computer systems that can improve their performance for a given task progressively as more samples are provided, where performance is quantitatively measured. Usually this performance improvement comes from the execution of an algorithm (a *learning*

[11] There are similarities with the notion of problem representation in a problem space as famously indicated by Newell and Simon (even though here we refer to the formulation made by Bechtel and Richardson). Bechtel and Richardson's concept of problem is an instantiation of the four elements of the problem representation (initial state, goal, defining moves, path constraints). For machine learning, a problem is an instantiation of an underlying input-output relation (a function) which has been sampled in order to obtain the dataset to be supplied to the learning algorithm.

[12] While there are algorithms that do not halt for some inputs or may require an unlimited amount of memory, they are not used in data science, so we will ignore them in what follows.

algorithm[13]) that builds and refines a model. Since such models are built by applying algorithms, we will call them *algorithmic models*. An algorithmic model is a model quantitatively depicting the relation between variables, and it is built starting from a set of samples and an algorithm[14]. The idea is that we infer a quantitative relation between already available pairs of inputs and outputs that will be used in similar cases to generate a predicted output where only the input is available.

ML uses a learning algorithm to build an algorithmic model from a set of *training samples*, called the *training set*. The time sequence is as follows: first a set of data is collected, then an algorithm is chosen to be run on those data, and finally the algorithm is run on the input data, so that the algorithm generates an algorithmic model. Several learning algorithms can be run on the same training set to yield alternative algorithmic models. Then, a set of *validation samples* which were not available at training time is used to measure the performance of each model, so that the algorithmic model which performs best on the validation set is selected. After that, a third test set, which is disjoint with both the validation and test sets, can be employed to carry out a *final assessment* of the selected model. These training, validation and test procedures are usually repeated many times (*replications*) with different random splits of the available data into training, validation and test sets. Through this process, the statistical confidence on the performance of the selected model is improved, as measured by suitable statistical tests.

3.2. What are algorithmic models of machine learning about?

Technically speaking, ML algorithmic models are *predictive models*[15]. When we apply a selected algorithmic model[16] to an input that was not used to train the algorithm in order to produce a predicted output, then we are generating a new prediction. Predictions refer to the

[13] Algorithms in machine learning can be supervised, unsupervised and reinforcement learning. We will refer to supervised and unsupervised algorithms because they are relevant to 'discover' quantitative relations between inputs and outputs. On the other hand, reinforcement learning algorithm are the less relevant here, which are used especially in engineering rather than natural or social sciences

[14] Here we must remark that each learning algorithm builds a different kind of algorithmic model. Some learning algorithms build complex models composed by many submodels which are combined by a consensus subalgorithm. These are called ensemble algorithms and models

[15] "The generalization performance of a learning method relates to its prediction capability on independent test data. Assessment of this performance is extremely important in practice, since it guides the choice of learning method or model, and gives us a measure of the quality of the ultimately chosen model" (Hastie et al 2013, p. 219)

[16] The algorithm produces the model, which in turn is applied to test new data

local context of a given problem. That is, a model which has been learned by a learning algorithm to predict whether a mortgage borrower will repay the loan is not expected to perform well if used to try to predict whether a student will repay a student loan.

An algorithmic model is designed to predict the behavior of a single target system, which is defined *as an aspect of reality which is associated to a probability distribution of possible cases from which data samples can be drawn at random, in order to obtain the training and validation datasets, and from which the test cases also come*. In addition, the role of machine learning can be further extended by automatically *discovering classes*[17] which would be later predicted. In biomedicine this is pretty common, and classes of (for instance) mutated genes and associated diseases might be later used in order to predict prognoses for new patients (Akbani et al., 2015, pp. 1687-1688).

Algorithms used in ML build algorithmic models of problems whose predictive accuracy typically enhances as more data are supplied[18]. We have reported in another work (Ratti and López-Rubio, 2018) the example of the algorithm PARADIGM (Vaske et al 2010). This algorithm is used in biology to predict the state of molecular biological entities that for some reasons cannot be measured directly. In particular, PARADIGM is aimed at inferring how genetic changes in a patient influence genetic pathways. Pathways are collections of molecular entities that are causally relevant to higher cellular processes. Because of biological complexity, pathways are large and complex. Therefore, it is impossible to keep in mind all the possible entities and their relations into mind - there are databases that keep track of all the details. PARADIGM integrate several genomic datasets and will predict whether a patient will have specific pathways disrupted given the status of some measured entities (e.g. specific genes mutated). The model generated by PARADIGM will specify the expected state of several biological entities which are not experimentally measured, by drawing the

[17] Given a universe of objects, a class is a subset of the universe whose elements share some features which makes the class relevant for some scientific or technological purpose

[18] Abundance of data facilitates a better model selection and assessment, because the estimations of the performance of a model are more accurate (Hastie, 2009, p. 222). Infrequent but relevant cases can only be observed in sufficient numbers if the number of samples is large enough (Junqué de Fortuny et al, 2013, p. 216). For an important kind of algorithmic models such as maximum likelihood estimators, there are formal proofs that under mild conditions they are both asymptotically unbiased, i.e. the bias reduces to zero as the number of samples tends to infinity; and asymptotically efficient, i.e. the variance reduces to the lowest possible for any model as the number of samples tends to infinity (Sugiyama, 2015, pp. 140-143). In some fields such as natural language processing, as the number of samples increases, a point is reached where all relevant cases are sufficiently covered in the dataset, so that memorization of examples outperforms models based on general patterns (Halevy et al, 2009, p. 9).

information from databases. We have here the characterization of ML through the idea of a problem; there is the input (i.e. the measurements on patients), the samples (i.e. the database), and the output (i.e. the status of entities given inputs and samples). Intuitively, the larger the inputs and samples, the more precise the algorithm performances will be, because by adding more information the algorithm will be able to perform more precise calculations on the expected status of unmeasured entities.

Models generated by ML are *black boxes* (Frické, 2015, p. 655). 'Black-box' is a central concept to understand ML's focus on predictions. By 'black-box' model we mean models that *while being precise in associating inputs with outputs, they are nonetheless unable to clearly depict the (causal) relations between them* (Hastie, 2009, pp. 351-352)[19]. Consider an example. Let's say that we have a model elaborated on the basis that anytime there is a specific set of customer characteristics, it is very likely that the customer population with such characteristics will buy a specific type of product. Let's say that we can calculate quite precisely the 'very likely'. This model is said to be 'black-box' because even if it can clearly point out that there is a quantitative relation between specific inputs (i.e. the set of customer characteristics) and outputs (i.e. the probability of buying a specific product), still it is not at all able to uncover the causal connectivity between the characteristics of the customers and the mental processes which lead them to buy the product. In other words, we know *how to associate* inputs and outputs, but we do not know *what is in between*. Machine learning algorithms generate models that 'black-box' the relation between inputs and outputs, though they are very specific in saying that *there is* a relation. This association between inputs and outputs without saying what's in between may be interpreted in a mechanistic framework as the fact that such black-boxed models are not explanatory[20].

However, one may say that the algorithmic black-boxed models generated by ML may be turned, in principle, into 'white-box' models, namely models where the causal connectivity between components is uncovered. This is usually what we do with how-possibly mechanistic models. Let's imagine that we have a straightforward correlation between the mutations of certain genes and specific phenotypes. Let's hypothesize that we are working with a very well known human gene, such as *PTEN* or *KRAS*. Since we already know the processes that one of these genes can possibly trigger, we can elaborate a causal

[19] Please note that this concept of 'black box' differs from Latour's, who refers to ready-made computer systems which are assumed to perform their function correctly (Latour, 1987, pp. 3-4).

[20] Any how-possibly model is in a sense a black-box model, because it establishes that there are components that are clearly involved in a phenomenon, but we do not know exactly how.

narrative (see Ratti and López-Rubio 2018) - though idealized and/or abstract, of course - connecting such genes with the phenotypes we are observing. We can refine the causal history by instantiating well-known experimental strategies for discovering mechanisms, such as the ones mentioned in (Bechtel and Richardson 2010; Craver and Darden 2013). By depicting the causal connectivity between components, we try to make sense of why there is a relation between inputs and outputs and hence we say what's in between inputs and outputs. In other words, by connecting inputs and outputs we usually obtain an *explanation*. However, an algorithm used in the context of ML does not do anything like that; it does not yield an explanation, it just points out a connection between two entities that must be explained by experimental strategies for discovering mechanisms.

3.3 Predictivism in machine learning

The appeal to predictive performances is motivated by specific epistemological reasons. Even if ML is oriented towards predictions (a quantitative relation between an input and an output), one may think that, at least in principle, this does not prevent this discipline from being oriented towards mechanistic-like explanations as well. However, this is not the case; ML is *only* about predictions. This is not a sort of *predictivism* (in the sense of Kaplan and Craver, 2011); predictivism in ML is *unavoidable.* In other words, a predictivist would say that a scientific model which conveys accurate predictions counts as an explanation (Kaplan and Craver, 2011, p. 605), while ML focuses on producing models with a good predictive performance without addressing the question of whether the obtained models have an explanatory value[21]. In any case, ML models are not required to be simulations of the real world processes that they are associated to. In order to explain why ML focuses on prediction, first we will say when and how we need machine learning. Next, we will say why the problems that ML tackles can be approached just by appealing to predictions.

3.3.1 When and how we need machine learning

[21] There are some machine learning methods like Bayesian networks (Spirtes et al., 1993) which can learn causal connections among variables from data. They can ascertain that some variables are causes of other variables, but they cannot say anything about the specific mechanism that is behind such causal connection. In other words, Bayesian networks by themselves cannot produce any explanations, since the specific mechanisms must be found by the scientist.

We need ML when black-boxes are the best options to discover something about target systems[22]. This is more likely to happen in disciplines where the system under study is complex or chaotic. For example, epidemiology and social sciences are amenable to black boxes because the underlying processes involve a huge number of relevant variables with highly nonlinear dependencies among them[23]. Such complexity makes it very difficult to organize variables in a coherent mechanistic-like model. Either we simplify the model or we tackle complexity and settle for black-boxes. This is because even if we could elaborate intelligible models of chaotic and complex systems, these would likely be oversimplifications with limited predictive performance in practical settings. For example, if you want to predict the probability that a person will develop osteoporosis, you may consider a simple model which considers that such probability is a linear function of the age of the person. This is certainly intelligible, and it could provide a first approximation to the problem. But if you need better predictions and a richer (in terms of details) model, you will have to opt for a more complex and (necessarily) less intelligible model where all the risk factors such as age, gender, family history, bone structure, body weight, broken bones, ethnicity, smoking, alcohol, medications, and previous related diseases interact in nonlinear and non obvious ways. Moreover, if the mechanism under investigation involves a complex relation among variables, then an input-output relation cannot be written down in intelligible terms. This can happen even if the number of variables is small, i.e. for small size models, depending on the intrinsic complexity of the relation among the variables in the system. Under these circumstances, only unintelligible algorithmic models can yield good predictions.

3.3.2 Machine learning performances and the bias-variance tradeoff

We need ML when complex systems are our targets, and ML is oriented towards predictive models. What is the best predictive model in this context[24]? This is a pretty technical matter; there are theoretical difficulties in obtaining a model with high predictive performances. This

[22] The real-world process and the model are categorically different. For trained machine learning models, they might not even be structurally similar, depending on the kind of algorithm that is used to learn the model. There are algorithms that aim to learn the structure of the biological interactions, which can be regarded as white box algorithms, while other algorithms do not try to yield a model which resembles the target biological system..

[23] Well known cases include stock market prediction, modeling the spread of communicable diseases on a population, and recommendation systems for online marketing.

[24] Please note once again that predictions here do not necessarily overlap with predictions in the mechanistic context

difficulty can be explained by appealing to the so-called *bias-variance tradeoff*. Let us see what this is about.

Following Shmueli's notation (Shmueli, 2010), for a given problem there is an abstract level of reasoning where the existence of an abstract function F is postulated, which relates some abstract concepts X to other abstract concepts Y, where $Y=F(X)$. Then, an operationalization must be carried out to translate these abstract concepts into measurable variables X, Y linked by the function f which captures the exact relation among them, $Y=f(X)$. The last step is the construction of a model f', which is an approximation of f because it cannot capture all the details of the phenomenon under study. Here the phenomenon must be associated to a single target system from which samples (X, Y) can be drawn at random. In other words, a phenomenon is associated to a specific probability distribution $p(X, Y)$ which can be sampled by observing the target system. Data science comes into play whenever the abstract function F is too complex to be ascertained by a human[25] (Pietsch, 2015, p. 915; James et al., 2013, p. 19), or it cannot be used to derive a suitable predictive model f'. Data science proceeds by learning f' from examples of pairs X, Y (James et al., 2013, p. 17) and it considers f' as an acceptable model of reality until a better one is found, even if f' does not provide an explanation about how f works[26].

How does data science compare models f_i' to choose the best one? In this context, most model performance criteria try to strike a balance between the complexity of the model and how well it fits the data (Bishop, 2006, pp. 32-33). At the heart of this issue is the *bias-variance tradeoff*. A well-known performance criterion for predictive models is *the expected error* which is committed when using a model f' for prediction. This error is the sum of three non-negative quantities (James et al., 2013, pp. 33-35; Hastie et al., 2009, pp. 223-228); *the irreducible error*, which stays the same no matter how accurate the model is; *the bias*, which is caused by the difference between the true function f and its approximation f'; and *the variance of the learning method*, which is caused by the oscillations of f' as the training dataset is changed. *The best model is the one for which the sum of bias and variance is very low*. However, this is very difficult to obtain.

[25] Some data sets are so large and complex that it would be almost impossible for a human to find significant patterns without the help of algorithms which automatically elaborate models to fit the data (Dhar, 2013, p. 68).

[26] A difference between the desiderata of models f' for explanation-based and data science approaches is that for the former we want f' to have a causal structure that is similar (not understood in the philosophy of science technical sense) as much as possible to the phenomenon we are investigating, while for data science the predictive performances for new cases is the main goal.

Imagine that we wish to predict the ambient temperature tomorrow at 0:00 GMT for all locations on the Earth surface, given the temperatures measured today at 0:00 GMT in all weather stations in the world, so that the problem has one input variable for each weather station. We might use a model f_1' which considers that the average temperature is the same on all points of the Earth surface, and it is computed as the average of the recorded temperatures on all stations. This model would have a high bias, since it is far from the real function *f*. For example, it would grossly overestimate the temperature at the polar regions, and it would underestimate at the deserts. We may also use another model f_2' which divides the Earth surface into 510 patches of about 1 million square kilometers (the surface of the Earth is about 510.1 million square kilometers), and then predicts the temperature at each patch to be the average of the recorded temperatures on all stations within the patch. The model f_2' would have a smaller bias than f_1', since the overestimations and the underestimations would be less severe. However, f_2' would have a higher variance than f_1', because patches for f_2' contain fewer stations, so that any random changes in the temperature recorded at one station (training dataset change) would cause a larger oscillation in the temperature prediction for the patch that the station belongs to, since those random changes happen independently from one station to the others. There could be a third model f_3' which divides the Earth surface into 510,000 patches of about 1000 square kilometers. Now f_3' would have an even smaller bias than f_2', since it would represent even smaller details of the temperature distribution. But f_3' would have a very high variance, since there would be very few weather stations in each of these small patches, so that the computed average on each patch would be even more sensitive to random changes in the temperature recorded at a station, i.e. dataset changes. As we go from f_1' to f_2' and then to f_3' both the model complexity and the variance increase, while the bias diminishes. This is the tradeoff. However, data science techniques *can reduce the variance of highly complex predictive models (while keeping bias low as well) by harnessing large numbers of samples*[27].

The computational techniques that are used to manage large numbers of samples are also suitable for large numbers of input variables. Therefore, a typical way to enhance the

[27] Let's consider a variation of the example of temperatures. We may say that the temperature tomorrow at 0:00 GMT in a particular weather station will be the average of the temperatures recorded on the same day of the year at 0:00 GMT in the same weather station, computed over the available weather data. We can diminish variance as follows. As the number of years with available data (the number of samples) increases, the variance diminishes because the output that *f*' produces for unseen test data will be less sensitive to oscillations in the training dataset, i.e. extremely cold or extremely hot years in the historic record which is used for training.

predictive accuracy of a model is to increase the number of input variables that are supplied to the algorithms. The complexity of the learned models usually increase as the number of input variables grows, but this is not a pressing concern for machine learning because both the computer hardware and software are capable of managing large numbers of input variables and samples at the same time.

3.3.3 Why we cannot obtain explanations when using machine learning

Let us summarize what has just been said:

1. ML generates algorithmic models which are predictive models in nature.
2. Algorithmic models are necessary to study very complex systems[28].
3. The best (predictive) algorithmic model is (by definition) the model which yields the best predictions, and this is that which attains the minimal sum of bias and variance. Therefore, we want to keep both bias and variance low.
4. Keeping both bias and variance low is hard to achieve because of the bias-variance tradeoff. However, through ML, the variance of complex models could be reduced while keeping the bias low by *supplying large amounts of training samples*.
5. The same computational techniques that are used to manage large numbers of samples are also employed to increase the number of input variables, with the aim of further enhancing the predictive performance.

And here come the troubles; *supplying larger and larger numbers of input variables means also making the model even more complex* (i.e. the size of the model increases). The argument can be summarized as follows. If we want to generate reliable models about complex systems we have to use algorithms (as defined above). In this context, a model will be a predictive model, because algorithms of data science obtain exactly that. We do not want just models; we want *good models* in terms of their predictive performance (which, again, is not the same as predictive performances in the mechanistic context). If we want a better predictive model (i.e. one for which the sum of variance and bias is low), we have no choice but to make the model more complex by supplying larger and larger amounts of training data. Since this usually means more variables, *ML works better when the number of variables is*

[28] This makes sense only if we assume that complex systems must be approached by taking into account the contribution of each of their components. There might be other approaches to complexity (e.g. systemic or holistic approaches) which may not require the attitude we are describing here

overwhelming, but this also means that in order to connect inputs and outputs we have to identify the causal connectivity or causal structure between an increasing number of components.

This worry can be framed in terms of the relation between explanation and intelligibility that we have delineated in section 2.2 and argued in (Ratti and López-Rubio 2018). An algorithmic model is very unlikely to be 'turned' (or 'upgraded') into an explanatory (mechanistic) model. This is because machine learning algorithms *do not generate intelligible models* (in the sense delineated in Section 2.2); there are too many variables, and we do not know how to use the model in order to instantiate a series of 'built-it' tests (in case we are in a mechanistic context) or other procedures such that we can elaborate clearly the causal connectivity (i.e. the organization) between the components. Therefore, it seems that when we use machine learning because of the magnitude of the data sets, at the same time *we cannot even in principle* obtain explanatory (mechanistic) models, but only predictive models. Algorithmic models obtain predictions, and predictions are all we can obtain from them[29].

A consequence of this is that, under the circumstances where ML is employed, prediction and explanation stand in a *relation of tradeoff*. The tradeoff[30] consists in the fact that the more we use machine learning because we have larger volumes of data, the less our chances of elaborating (mechanistic) explanations will be. That is, the underlying biological processes are so complex that intelligible mechanistic models are oversimplifications of reality. Only more complex, less understandable data-based models can capture the complexity of the underlying biology. Mechanistic models in molecular biology are not very good at prediction (in the machine learning sense) because their bias is very high, i.e. they are

[29] It can be argued that there are algorithms which learn the structure of the biological system under investigation, so that we can relate somehow the work of algorithms to mechanistic descriptions. For example, the algorithm PARADIGM mentioned above does learn the structure of the interactions among the entities. But it cannot ascertain the specific mechanisms which underlie behind the interactions. So the lack of intelligibility comes from the inability of algorithms to learn those specific mechanisms rather than a dissimilarity between the learned interaction structure and the real one. There are other algorithms such as Prediction Analysis of Microarray (PAM, Tibshirani et al., 2002), which do not intend to learn the biological structure in any way, since they are aimed to prediction only. When algorithms like PAM are used, it means that scientists are not particularly interested in the structures, but in the predictions. In other words, when the biological network under investigation is too complex to obtain an explanatory mechanistic model, then the only option is to use a black box prediction algorithm which does not intend to learn the structure of the analyzed network.

[30] Let us clarify again: this tradeoff between explaining and predicting is a consequence of the way machine learning deals with the bias-variance tradeoff. In other words, this tradeoff and the bias-variance tradeoff are different tradeoffs.

far from representing the real target system due to the unavoidable oversimplification that they entail. As explained before, any predictive model must keep both bias and variance low to yield good predictions. Mechanistic models are expected to yield the best predictions when the complexity of the modeled system is small, so that there are no significant oversimplifications in them.

Before proceeding, it is important to stress something important about the tradeoffs (bias-variance, and explanatory abilities-predictive performances) we talked about. In philosophy of science, there has been an interesting discussion on tradeoffs in modeling for a few decades. The discussion originated from Levins' influential paper on model-building in population biology (1966), and it has been recently developed by Weisberg in several papers (for instance Weisberg 2006; Matthewson & Weisberg 2009). Even though Levins does not use the word 'tradeoff', it seems he identified a sort of three-way tradeoff between three different features of models. These features are precision, realism, and generality. By following Weisberg's careful analysis (2006), generality can be defined as the number of target systems to which a model can apply to. Next, realism is how well a model represents its target system. Finally, precision is the "fineness of specification of parameters" (Weisberg 2006, p 636). The type of tradeoffs identified by Levins are

(1) you cannot increase both realism and precision without sacrificing generality,
(2) you cannot increase both generality and precision without sacrificing realism
(3) you cannot increase both realism and generality without sacrificing precision.

Are these tradeoffs similar to the tradeoffs developed here? We do not see any resemblance between the tradeoff we have identified between explanatory abilities and predictive performances, and Levins' tradeoffs. The tradeoff between bias and variance might deserve a little bit more of discussion. High precision can be interpreted as low variance, since variance measures the variability of parameters as they are learned from training data. High realism might mean low bias, since bias can be pretty much defined in the same way Weisberg characterizes Levins' notion of realism. However, generality is much more difficult to interpret in the context of machine learning. In fact, as it has been specified above, the scope of ML models is not that broad, since the models are learned with specific datasets to solve specific problems. Hence, generality will be usually low. Given the fact that Levins' theory of tradeoff is a three-way tradeoff, and the tradeoff between bias and variance is a tradeoff

between two features, we find it difficult to find a straightforward correspondence between them.

4. BIOLOGY AND MACHINE LEARNING

Because ML masters complexity efficiently, it is likely to be used more and more in the next decades in several scientific fields, including biology. How does molecular biology deal with such a tradeoff? It is time to make explicit the comparison between mechanistic models and machine learning predictive models, especially in light of the tradeoff between explanation and prediction.

In the case of mechanistic models, explanation and understanding (in the sense of 'intelligibility' as specified above) are strictly related; one desideratum of *explanatory* mechanistic models is that they spell out the organization (understood as 'causal connectivity') between the entities and processes producing the phenomenon of interest. Pinpointing the organization implies that we can 'use' (as specified above) the model in many ways, in particular *via* 'built-it' tests. Therefore, in order to elaborate an explanation in this context, the model must be intelligible to us, both in the phase of construction and confirmation. Hence, explanation and understanding are strictly related to the task of grasping how entities and activities are organized.

On the other hand, predictive models of machine learning are more efficient in their performances when the number of variables increases, i.e. larger size models are preferred[31]. If we use the language of mechanistic models, this means that the number of 'entities' and 'activities' (variables) included in the model has to be high for machine learning algorithms to generate efficient models. But since the number of variables increases, the difficulty of finding out the organization between a high number of variables increases as well; the more machine learning is efficient, the less these models are intelligible. In other words, the more predictive models of machine learning are 'good' models according to the standards of the discipline, the less intelligible they are, and hence the less they can be turned into explanations. *As soon as predictive performance increases, explanatory abilities (and hence power as well) decrease*.

This general analysis suggests that data science (in particular ML) has to impact the aims and goals of any science where explanatory mechanistic models are central. But is this

[31] For network modeling in molecular biology, the number of variables is fixed by the number of detected compounds, so that there is no flexibility to choose the size of the model to be learned.

really the case with molecular biology? In order to understand this, we fully scrutinize an important 'big science' project of molecular biology that has been at the forefront in applying machine learning to biological questions. This project is *The Cancer Genome Atlas*[32] (TCGA).

TCGA has started in 2005 and ended in 2016. The main goal of the project was to study the biology of cancer through a massive use of sequencing technologies and bioinformatics. The project background is rooted in the evolutionary view of cancer as genomic instability, which in turn is grounded in the view that there are certain genes (called cancer genes - be they oncogene or tumor suppressor (Morange 1998; Weinberg 2014)) that play a prominent causal role in the development of tumors. The framework of TCGA is the typical reductionist and mechanistic framework that has been guiding molecular biology for at least 60 years (Weinberg 2014; Tabery et al 2015). The structure of TCGA studies is centered around the use of sequencing technologies to accumulate big data sets (usually about genes, mutations or structural variations both at the genetic and epigenetic levels) which are then analyzed by computer scientists, who rely heavily on the use of data science (and machine learning in particular). In most cases, such studies identify a certain number of genes, mutations and/or structural variations that are then associated in various ways to specific phenotypes[33] (i.e. specific types of tumors). In what follows, we will provide an epistemic analysis of *all* studies published by TCGA, in order to evaluate whether relying on machine learning has, in at least this highly representative case, changed the goals and priorities of molecular biologists.

The motivation to use TCGA is straightforward. TCGA is a pioneering project in the development of sequencing technologies and bioinformatics tools to adapt biology to the new big data trend, and its studies and its data portal are extensively used as 'models' (both in technical and non-technical senses!) for anyone in contemporary biology who wants to deal with big data sets, especially in cancer biology research.

4.1 The screenings

The information about TCGA studies can be found in Supplementary Table 1. In total, we scrutinize 43 publications. 31 out of 43 are official publications (1-31) of TCGA[34]. The rest

[32] https://cancergenome.nih.gov
[33] For a more thorough exposition of the structure of TCGA studies, see (Ratti 2015)
[34] https://cancergenome.nih.gov/publications

of publications (32-43) are part of an initiative called *TCGA Pan-cancer Analysis*[35], which is a meta-analysis of data sets accumulated by TCGA across different types of cancer. We report different characteristics of the screenings analyzed.

First, we indicate the magnitude of the input analysis, which is usually the number of cancer samples subjected to next-generation sequencing and computational analysis. Numbers may vary considerably, especially in light of the availability of samples for a given cancer. For instance, study 16 has a low number of samples (n = 66) because of the nature of tumor analyzed (i.e. chromophobe renal cell carcinoma) for which the availability is substantially low, especially if compared to more common cancers such as breast cancer (see for instance study 21, where sample is n = 817). When we report the number of samples, we usually specify the number of samples subjected to specific analyses (e.g. whole-exome sequencing, whole-genome, etc). Studies 32-43 have a high number of samples because they are meta-analyses of different data sets of TCGA consortium.

Next, we indicate the magnitude of the output, namely the amount of mutations, genes or structural variations that are strongly associated with the input. In most screenings data sets are not limited to the genetic or epigenetic levels, and a focus on the proteomic level is provided. This further level of biological analysis is evidence for the data-intensive turn in biology, but our analysis focuses especially on whole-exome, whole-genome and somatic copy-number alterations data[36]. These are, respectively, the mutational analysis of exons across genomes, the analysis of whole genomes, and the analysis of structural variations. Therefore, each study from 1 to 31 (with the exception of study 15) identifies a quantitative relation between a particular phenomenon (i.e. a type of cancer) and specific molecular events/entities (e.g. mutations, structural variations, epigenetic alterations, etc).

Thirdly, we create a taxonomy of aims and goals for such screenings. In studies from 1 to 32, some have the only aim of understanding in quantitative terms how tumors are mutated, and which mutations are better associated to them (n = 6), while others in addition to this provide what is called a 'pathway analysis', namely a study to contextualize mutated genes within already characterized biological pathways[37]. (n = 10). Other studies provide also

[35] http://www.nature.com/tcga/

[36] The reason for doing this is that, when mechanistic sketches in this field are outlined, usually genes and mutations at the genetic level are considered central, and hence we decided to focus only on this important level because the epistemic processes applied to it are similar to the ones applied to other levels. Hence the analysis of other levels may be redundant

[37] On the difference between 'pathways' and 'mechanisms' see (Ross 2018) and (Boniolo and Campaner 2018)

attempts to sketch mechanistic descriptions, though by involving only few of the entities found to be associated to the phenomenon of interest, and in this way they identify specific biological processes that may be important for the development of specific types of cancer (n = 15). Studies from 32-43 have different aims. They are mostly computational analyses of big data sets across different types of cancer. The most minimal are focused just on 'oncogenic signatures' across cancers (n = 4), while others try to identify possible high-level abstract mechanisms that may be conserved across tumor types (n = 6). Finally, few studies are devoted to the development of specific algorithms (n = 2).

Fourth, we report the number of machine learning algorithms used. We observe that in general several algorithms are used within the same study, and this indicates the fulfillment of different sub-goals within each screening. These sub-goals are identified in two sheets of Supplementary Table 1, named 'Algorithms' and 'Taxonomy'. 'Algorithms' provides a list of the algorithms used and these are associated to the sub-goals identified. On the other hand, 'Taxonomy' quantifies the number of times such sub-goals have been achieved in the screenings analyzed.

4.2 Predictions, explanations and the aims of molecular biology

If our epistemological analysis about explanation and machine learning in the mechanistic context is correct, then we should observe specific things in analyzing the outputs of a project like TCGA. In particular, we should observe that (a) *there is no fully-fledged mechanistic explanation of cancer* in each and every paper of TCGA and (b) *predictive models serve goals that are different (and independent) from the typical explanatory-based aims of molecular biology*. Let us now see whether these two claims that have been derived in principle hold in the actual practice of some data-intensive biology.

First of all, we hasten to add that when we say 'fully-fledged mechanistic explanation of cancer' we do not mean a 'theory' of cancer. There can be a general theory of cancer in the sense of a family of models explaining how instances of cancer may manifest one or more hallmarks that are usually produced in certain ways[38]. What we mean by 'fully-fledged' mechanistic explanation of cancer is that a biologist, starting from the characterization of a phenomenon (a particular type of cancer in x samples), tries to instantiate those discovery strategies identified by mechanistic philosophers (Craver and Darden 2013; Bechtel and richardson 2010) in order to explain (i.e. elaborate a causal narrative) of how that particular

[38] Consider for instance the so-called 'hallmarks of cancer' (Hanahan and Weinberg 2000; 2011)

phenomenon has been produced. For example, one may start to identify entities and activities that are likely to be involved in the production of the phenomenon to explain, and then recompose these components into a narrative of how these components produce exactly the explanandum (Glennan 2017).

Let us now turn to the extent to which (a) applies to actual biological practice. On the basis of the analysis of TCGA publications, we argue that, in each paper analyzing a specific cohort of samples of a specific type of tumor, the data gathered is not used to explain mechanistically how that particular cancer has been produced. However, one may say that in the past it has been already observed that data-intensive studies in molecular oncology follows a mechanistic recipe. For instance, Ratti (2015) emphasizes that the structure of the discovery strategies of studies such as TCGA's is similar to the ones emerged in the mechanistic tradition (Bechtel and Richardson, 2010). However, the study of Ratti (2015) is not very specific in spelling out the explanatory dimension of data-intensive studies in molecular biology. For instance, consider the phase of validation of computational analysis (Ratti 2015, pp. 206-209). This is the phase when a few mutations/genes that are found associated to cancer are 'experimentally' validated to see if they play a causal role in the phenomenon of interest. Experimental validation – also known as the 'functional test' (Bertolaso, 2016) – can be interpreted as being part of the discovery strategies of the mechanistic tradition (Bechtel and Richardson 2010; Craver and Darden 2013)[39]. However, even though discovery strategies of the mechanistic tradition aim at explanatory goals, a mere experimental validation of the causal role of an entity is just one piece of a very complex puzzle. Numerous studies of TCGA experimentally validate some cancer genes, and they also mention the general processes these entities may be involved into. For instance, in Supplementary Table 1 we emphasize that studies number 6, 8, 11, 13, 16, 17, 18, 19, 20, 21, 22, 23, 24,30, 31 investigate *possible mechanisms* of a few genes/mutations. Instead of trying to explain through a mechanistic explanation a phenomenon such as how from certain mutations a specific type of tumor will show this and that phenotypes, these studies focus on a few entities that machine learning analysis strongly associates with tumors. This means that there is not an attempt to explain the phenomenon of interest (i.e. a specific type of tumor), but only an exploration of the information about *few* entities that may play important causal

[39] Please note that experimental validation is different from the validation phase of machine learning procedures, where an algorithmic model is chosen according to its performance on a validation set, which is a subset of the available dataset.

roles in one or more tumors. The attention shifts from the phenomenon to explain to some entities that may be involved[40]. In other studies, rather than explaining a phenomenon, information about mutated genes is contextualised in so-called mechanistic schemas (Levy 2014). This is when TCGA practitioners try to make sense of the causal role of one or more genes by showing that what the genes are supposed to do is consistent (by means of computational analysis) with it being part of an already known biological pathway whose disruption in cancer has been observed in the past, even though the role of the genes in the pathway is not clarified at the level of detail that a mechanistic explanation would require (studies number 4, 5, 7, 9, 10, 12 14, 25, 27, 28 only focus on pathway analysis, while the previously mentioned studies have both the focus on new cancer genes and pathway analysis). Schemas are employed notably in some studies of the Pan-cancer analysis of TCGA (studies number 33, 38, 39, 41, 42, 43), where selected entities that are associated with several types of tumors are then investigated with respect to their possible mechanistic contexts.

Therefore, while no study of TCGA does provide a fully-fledged mechanistic description of the phenomenon of interest, some may provide either sketches or schemas but limited to a few entities. This is consistent with the introduction of ML's methodology in molecular biology and our analysis of explanation, understanding and prediction; *when big data sets with an overwhelming number of variables are analyzed, no heavily vetted mechanistic explanation is attempted*, but rather biologists limit their analyses to a few entities[41]. The fact that studies try to provide evidence for new cancer genes is also related to the biomedical dimension of the project; its findings can be used for later phases of drug discovery (Ratti 2016). However, our claim is that if you do molecular biology with machine learning techniques, and if you want to have the best machine learning performances, then

[40] One may also say that this is the point where 'data-intensive' studies meet mechanistic studies, in the sense that studies such as the ones of TCGA are only one step towards the elaboration of mechanistic explanations. However, this observation may miss the importance and the role of bioinformatics tools. These are not just tools aimed at selecting a few entities by means of eliminative inductive strategies, but tools that are used to characterize the complexity of biological systems. In fact, we may use these tools to prioritize a few entities and elaborate a simple mechanistic models, but by doing this we would miss the complexity of biological systems and the other analyses that can be done on biological complexity without necessarily narrowing down just a small and local part of it, as we do when we just focus on a few cancer genes as part of the complexity of a particular type of tumor.

[41] This aspect may be interpreted as being related to the pathway concept, as Ross (2018) points out when she says that "instead of identifying a particular explanatory target and 'drilling down', these maps [i.e. pathways representation] involve identifying a set of entities in some domain and 'expanding our by tracing their causal connections" (p 13)

you cannot even in principle elaborate fully-fledged mechanistic explanation[42]. It is important to notice that this is not due technological limitations; the argument spelled out in Section 2 and in great detail in (Ratti and López-Rubio, 2018) is that the more the size of the model increases, the less the human mind is able to organize the model's components into a causal narrative, which forms the backbone of any mechanistic description with explanatory force. Moreover, we would also like to resist the objection that the explanation - even if it cannot be elaborated right now by humans - is still there, present in the complexity of the biological data sets. Our appeal to the capacity of the agents elaborating explanations implies and advocates a sort of epistemic conception of explanations - the explanations is not contained in the how-possibly model and has to be 'discovered', but it is a product of the cognitive abilities of the agent. Therefore, if there are no cognitive agents that can elaborate the explanation, then there is no explanation at all.

Let us turn to (b). By connecting molecular biology to the mechanistic tradition, philosophers of biology have emphasized its explanatory goals. Even though prediction, control and other aims are recognized, these have value only to the extent that they increase the chances of elaborating explanations. However, it seems that TCGA hardly fits this framework. While it is true that predictive models may function as a way to select entities and activities that can be central in mechanistic descriptions (Ratti 2015), this ignores the fact that predictive models play other important roles, which are not necessarily related to explanation and cause-effect narratives. Among the others, predictive models are useful for disease classification. For these purposes, an unsupervised clustering algorithm is first employed in order to discover clusters of tumors according to their genetic signature. This does not necessarily mean that the members of a cluster share the same oncogenic mechanism, but only that the observed mutations are similar, i.e. no detailed explanation of a common mechanism is proposed for the members of a tumor cluster. After that, new tumors can be classified by supervised learning algorithms into these previously discovered clusters. Again, the classification is not dependent on any explanation of the oncogenic mechanisms, since the algorithm classifies the tumors according to their genetic signatures. In turn, this is related to the use of these studies as powerful tools for diagnosis. In order to diagnose a specific disease, you do not need an explanation of the disease itself. In the case of these screenings, we do not need to connect causally or mechanistically a mutational signature to a

[42] One may argue that mechanistic philosophers' requirements for a good explanation are in tension, but this is beyond the scope of the present paper

specific type of tumor. We just need to know that anytime there is a specific pattern of mutations, then there is a precise quantification of the probability of having a specific type of tumor, even though plenty of associations will turn out to be just spurious correlations (as the study number 34 seems to suggest) or just indirectly related to difference makers. This is part of the aims of TCGA and its focus on the so-called *precision medicine*. By creating more and more specific subgroups of the same tumor by means of machine learning analyses, TCGA aim at elaborating highly specific genome profiles that can be next further tailored around the needs of individual patients or very small groups of patients.

Therefore, the analysis of the TCGA publications is consistent with what one could have inferred from our in principle analysis of molecular biology and machine learning, namely that machine learning pushes molecular biology towards slightly different epistemic aims and concerns. To say this more loud and clear, the introduction of machine learning in molecular biology has introduced a change in the molecular biology system of practice (Chang 2014)[43], by shifting the field from purely explanatory practices to a variety of predictive activities that may be even in a tradeoff relation with explanations.

5. CONCLUSION

In this article, we have analyzed the claim according to which the introduction of data science in molecular biology has somehow changed its epistemic aims.

First, we described the epistemic aims of molecular biology, and we have identified explanatory goals achieved in terms of the elaboration of mechanistic models/descriptions (Section 2; 2.1). We have also connected the task of building explanatory models to the intelligibility of such models, such that extremely complex mechanistic sketches are unlikely to be turned in good explanatory models.

Next, we described the epistemic aims of data science, and machine learning in particular. We have emphasized that machine learning points towards the elaboration of predictive models (Section 3.2). Moreover, we have also added that machine learning can produce more effective predictive models as the data volume it analyses grows, both in terms of number of samples and number of input variables. This, together with the remarks about

[43] This of course does not mean that more traditional forms of molecular biology cannot possibly coexist with machine learning-driven molecular biology

intelligibility of models, led to the identification of a tradeoff relation between the ability to elaborate mechanistic-like explanation and predictive models.

We finally showed that in practice molecular biologists have to deal with such a tradeoff; when they attempt to elaborate mechanistic explanations, they necessarily have to narrow their focus to a few of the variables considered to investigate the phenomenon under scrutiny (Section 4). However, the predictive force of such narrow explanations will be very limited if compared to the predictive force of predictive (and non-explanatory) models generated by taking fully into account the complexity of the data sets about a phenomenon.

ACKNOWLEDGEMENT

The authors would like to thank David Teira, Enrique Alonso, and the participants to the workshop "Making sense of data in the sciences" in Hannover, and in particular Federica Russo and Sara Green, for their valuable comments and suggestions. They are also grateful to the editor and four anonymous reviewers for their constructive feedback.

REFERENCES

Akbani, R., et al. (2015). Genomic Classification of Cutaneous Melanoma. *Cell*, *161*(7), 1681–1696.

Alberts, B. (2012). The End of "Small Science"? *Science*, *337*(September), 1230529.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: a mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(2), 421–441. http://doi.org/10.1016/j.shpsc.2005.03.010

Bechtel, W., & Richardson, R. (2010). *Discovering Complexity - Decomposition and Localization as Strategies in Scientific Research*. Cambridge, Massachusetts, and London, England: The MIT Press.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Boem, F., & Ratti, E. (2016). Towards a Notion of Intervention in Big-Data Biology and Molecular Medicine. In G. Boniolo & M. Nathan (Eds.), *Foundational Issues in Molecular Medicine*. London: Routledge.

Boniolo, G., & Campaner, R. (2018). Molecular pathways and the contextual explanation of molecular function. *Biology & Philosophy*

Callebaut, W. (2012). Scientific perspectivism: A philosopher of science's response to the challenge of big data biology. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 69–80. http://doi.org/10.1016/j.shpsc.2011.10.007

Carrier, M. (2014). Prediction in context: On the comparative epistemic merit of predictive success. *Studies in History and Philosophy of Science Part A*, *45*(1), 97–102. http://doi.org/10.1016/j.shpsa.2013.10.003

Cox, D. R. (2001). Comment to 'Statistical modeling: The two cultures'. *Statistical Science, 16*(3), 216–218.

Craver, C. F. (2006). When mechanistic models explain. *Synthese*, *153*(3), 355–376. http://doi.org/10.1007/s11229-006-9097-x

Craver, C., & Darden, L. (2013). *In search of Mechanisms*. Chicago: The University of Chicago Press.

De Regt, H. W. (2009). The Epistemic Value of Understanding, *76*(December), 585–597. http://doi.org/10.1086/605795

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

De Regt, H. (2017). *Understanding Scientific Understanding*. Oxford: Oxford University Press.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56(12), 64–73.

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

Frické, M. (2015). Big data and its epistemology. *Journal of the Association for Information Science and Technology, 66*(4), 651–661.

Gerlee, P., & Lundh, T. (2016). *Scientific models*. Basel, Switzerland: Springer.

Glennan, S. (2017). *The New Mechanical Philosophy*. Oxford University Press.

Golub, T. (2010). Counterpoint: Data first. *Nature*, *464*(7289), 679. http://doi.org/10.1038/464679a

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems, 24*(2), 8–12.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning*, 2nd edn. New York: Springer.

Hempel, C., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, *15*(2).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York: Springer.

Junqué de Fortuny, E., Martens, D., & Provost, F. (2013). Predictive modeling with big data: Is bigger really better? *Big Data*, *4*(1), 215–226.

Kaplan, D. M., & Craver, C. F. (2011). The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective. *Philosophy of Science*, *78*(4), 601–627. http://doi.org/10.1086/661755

Keller, E. F. (2002). *Making Sense of Life: Explaining Biological Development with Models, Metaphors and Machines*. Cambridge, Massachusetts, and London, England: Harvard University Press.

Latour, B. (1987). *Science in action*. Cambridge, Massachusetts: Harvard University Press.

Leonelli, S. (2011). Packaging Data for Re-use: Databases in Model Organism Biology. In P. Howlett & M. S. Morgan (Eds.), *How Well Do Facts travel? The Dissemination of Reliable Knowledge*. Cambridge, MA: Cambridge University Press.

Leonelli, S. (2012). Introduction: Making sense of data-driven research in the biological and biomedical sciences. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 1–3. http://doi.org/10.1016/j.shpsc.2011.10.001

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

Levy, A. (2014). What was Hodgkin and Huxley's achievement? *British Journal for the Philosophy of Science*, *65*(3), 469–492. http://doi.org/10.1093/bjps/axs043

Levy, A., & Bechtel, W. (2013). Abstraction and the Organization of Mechanisms. *Philosophy of Science*, *80*(2), 241–261. http://doi.org/10.1086/670300

Lombrozo, T. (2011). The Instrumental Value of Explanations. *Philosophy Compass*, *6*(8), 539–551. http://doi.org/10.1111/j.1747-9991.2011.00413.x

Love, A. C., & Nathan, M. J. (2015). The Idealization of Causation in Mechanistic Explanation. *Philosophy of Science*, *82*(December), 761–774. http://doi.org/10.1086/683263

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

Matthewson, J., & Weisberg, M. (2008). The structure of tradeoffs in model building. *Synthese*, *170*(1), 169–190. https://doi.org/10.1007/s11229-008-9366-y

Morange, M. (1998). *A History of Molecular Biology*. Cambridge, Massachusetts, and London, England: Harvard University Press.

Morgan, M., & Morrison, M. (Eds.). (1999). *Models as Mediators*. Cambridge University Press.

Pietsch, W. (2015). Aspects of theory-ladenness in data-intensive science. *Philosophy of Science*, *82*(5), 905–916.

Press, G. (2013). A Very Short History Of Data Science. *Forbes*. http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/

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., & López-Rubio, E. (2018). Mechanistic Models and the Explanatory Limits of Machine Learning. In*: [2018] PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association (Seattle, WA; 1-4 November 2018)* <http://philsci-archive.pitt.edu/view/confandvol/confandvolPSA2018.html>.

Rice, C. C. (2016). Factive scientific understanding without accurate representation. *Biology and Philosophy*, *31*(1), 81–102. http://doi.org/10.1007/s10539-015-9510-2

Ross, Lauren N. (2018) Causal concepts in biology: How pathways differ from mechanisms and why it matters. [Preprint] URL: http://philsci-archive.pitt.edu/id/eprint/14432 (accessed 2018-03-13).

Sloan, P. (2000). Completing the Tree of Descartes. In P. Sloan (Ed.), *Controlling our Destinies - Historical, Philosophical, Ethical, and Theological Perspectives on the Human Genome Project*. Notre Dame: University of Notre Dame Press.

Spirtes, P., Glymour, C., & Scheines, R. (1993). Causation, Prediction, and Search. New York: Springer.

Stevens, H. (2013). *Life out of sequence - A data-driven history of bioinformatics*. Chicago: Chicago University Press.

Stevens, H. (2015). Networks: representations and tools in postgenomics. In S. Richardson & H. Stevens (Eds.), *Postgenomics - Perspective on Biology After the Genome*. Durham and London: Duke University Press.

Stevens, H. (2017). A Feeling for the Algorithm: Working Knowledge and Big Data in Biology. *Osiris*, *32*(1), 151–174. http://doi.org/10.1086/693516

Strasser, B. (2011). The Experimenter 's Museum - GenBank , Natural History , and the Moral Economies of Biomedicine. *Isis*, *102*(1), 60–96.

Shmueli, G. (2010). To explain or to predict? *Statistical Science,* 25(3), 289–310.

Strevens, M. (2008). *Depth - An Account of Scientific Explanation*. Harvard University Press.

Sugiyama, M. (2015). *Introduction to Statistical Machine Learning*. Burlington, Massachusetts: Morgan Kaufmann.

Tabery, J., Piotrowska, M., & Darden, L. (2015). Molecular Biology. In *Stanford Encyclopedia of Philosophy*.

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression. *Proceedings of the National Academy of Sciences of the United States of America, 99*(10), 6567–6572.

Vaske, C.J., Benz, S.C., Sanborn, J.Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., & Stuart, J.M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics, 26*(12), i237–i245.

Weinberg, R. A. (1985). The molecules of life. *Scientific American*, *253*(4), 48–57. http://doi.org/10.1038/scientificamerican1085-48

Weinberg, R. (2010). Point: Hypotheses first. *Nature*, *464*(7289), 678. http://doi.org/10.1038/464678a

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

Weisberg, M. (2006). Forty Years of "The Strategy": Levins on Model Building and Idealization. *Biology and Philosophy*, *21*(5), 623–645. https://doi.org/10.1007/s10539-006-9051-9