Put it to the Test: Getting Serious about Explanation in Explainable Artificial Intelligence[*]

Florian J. Boge[†]& Axel Mosig[‡]

May 14, 2025

Artificial Intelligence (AI) has become a topic of major interest to philosophers of science. Among the issues commonly discussed is AI's *opacity*. To remedy opacity, scientists have provided methods commonly subsumed under the label 'eXplaibable Artificial Intelligence' (XAI) that aim to make AI and its outputs 'interpretable' and 'explainable'. However, there is little interaction between developments in XAI and philosophical debates on scientific explanation. We here improve on this situation and argue for a descriptive and a normative thesis: (i) When suitably embedded into scientific research processes, XAI methods' outputs can facilitate genuine scientific understanding. (ii) In order for XAI outputs to fulfill this function, they should be made *testable*. We will support our theses by building on recent and long-standing ideas from philosophy of science, by comparing them to a recent framework from the XAI community, and by showcasing their applicability to case studies from the life sciences.

1 Introduction

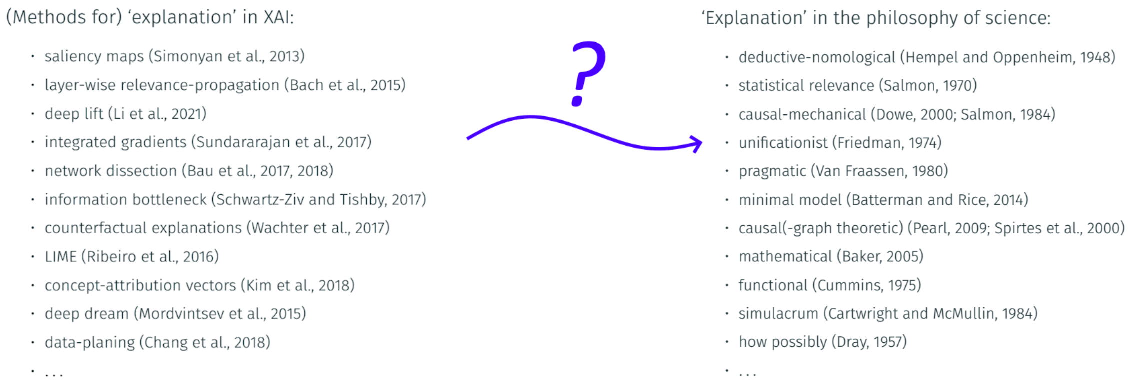
Artificial Intelligence (AI) has become a topic of major interest to philosophers of science. Among the issues commonly discussed is AI's *opacity* (Creel, 2020; Sullivan, 2022b; Räz and Beisbart, 2022; Boge, 2022): The fact that it is intransparent on various levels and in various respects to

[*] The version of record of this article will be published in *Minds & Machines*. A link/doi will be provided when available.

[†] Institute for Philosophy and Political Science (IfPP), TU Dortmund University
& Lamarr Institute for Machine Learning and Artificial Intelligence

[‡] Ruhr University Bochum, Center for Protein Diagnostics, Gesundheitscampus 4, 44801 Bochum

both users and developers how common AI methods work. To remedy this situation, software engineers, mathematicians and working scientists have provided methods that aim to make AI and its outputs ‘interpretable’ and ‘explainable’, these efforts being commonly subsumed under the label ‘eXplaibable Artificial Intelligence’ (XAI). At the same time, philosophers of science from at least Hempel and Oppenheim (1948) onwards have distilled various forms of explanation from scientific practice. Yet, there is little interaction between debates on explanation and developments in the field of XAI.



(Methods for) ‘explanation’ in XAI:

- saliency maps (Simonyan et al., 2013)
- layer-wise relevance-propagation (Bach et al., 2015)
- deep lift (Li et al., 2021)
- integrated gradients (Sundararajan et al., 2017)
- network dissection (Bau et al., 2017, 2018)
- information bottleneck (Schwartz-Ziv and Tishby, 2017)
- counterfactual explanations (Wachter et al., 2017)
- LIME (Ribeiro et al., 2016)
- concept-attribution vectors (Kim et al., 2018)
- deep dream (Mordvintsev et al., 2015)
- data-planing (Chang et al., 2018)
- . . .

‘Explanation’ in the philosophy of science:

- deductive-nomological (Hempel and Oppenheim, 1948)
- statistical relevance (Salmon, 1970)
- causal-mechanical (Dowe, 2000; Salmon, 1984)
- unificationist (Friedman, 1974)
- pragmatic (Van Fraassen, 1980)
- minimal model (Batterman and Rice, 2014)
- causal(-graph theoretic) (Pearl, 2009; Spirtes et al., 2000)
- mathematical (Baker, 2005)
- functional (Cummins, 1975)
- simulacrum (Cartwright and McMullin, 1984)
- how possibly (Dray, 1957)
- . . .

Figure 1: How do XAI methods relate to extant notions of explanation?

This situation is not entirely unwarranted. First of all, there is major diversity on both sides: ‘explainability methods’ comprise saliency maps (Simonyan et al., 2013), layer-wise relevance-propagation (Bach et al., 2015), network dissection (Bau et al., 2017, 2018), the information bottleneck-framework (Schwartz-Ziv and Tishby, 2017), local approximators like LIME (Ribeiro et al., 2016), concept-attribution vectors (Kim et al., 2018), integrated gradients (Sundararajan et al., 2017), data-planing (Chang et al., 2018), and many, many more. Similarly, Woodward and Ross’s oft-cited (2021) SEP-article on scientific explanation already mentions Hempel and Oppenheim’s (1948) deductive-nomological model of explanation, Salmon’s (1970) statistical relevance model, causal-mechanical explanation in the spirit of Salmon (1984) and Dowe (2000), and unificationist (Friedman, 1974) and pragmatic (Van Fraassen, 1980) approaches, but does not yet discuss minimal model (Batterman and Rice, 2014), causal-graph theoretic (Spirtes et al., 2000; Pearl, 2000), mathematical (Baker, 2005), functional (Cummins, 1975), simulacrum (Cartwright and McMullin, 1984), how possibly (Dray, 1957), and mechanistic explanations in the sense of Machamer et al. (2000) or Craver (2006). How, if at all, do these two sets relate (see Fig. 1)?

Indeed, several critical voices (Páez, 2019; Krishnan, 2020; Buijsman, 2022) have urged that the term 'explanation' in XAI is largely misguided: The ouputs of XAI methods are nothing like the claims or propositions involved in explanations in the sense familiar to philosophers. However, others (Räz, 2022; Erasmus et al., 2021; Baron, 2023; Watson and Floridi, 2021) have been more optimistic, in part based on detailed investigations of individual methods and philosophical accounts of explanation.

But *even if* an XAI method can be said to explain, how much can it thereby support the progress of science? For isn't the target of XAI methods just the behavior of an individual *Machine Learning (ML) model* or some class thereof, whereas scientific explanations target *real-world phenomena*?

Following recent developments in XAI (Murdoch et al., 2019; Roscher et al., 2020; Schuhmacher et al., 2022) as well as philosophical accounts of ML's potential in fostering scientific understanding (Räz and Beisbart, 2022; Sullivan, 2022b; Boge, 2022), we will here argue for two distinct theses – one descriptive, the other normative. Our descriptive thesis is that (i) when suitably embedded into a scientific research process, an XAI method's outputs can facilitate genuine scientific understanding. That is, when thus embedded, the method may facilitate an explanation of real-world phenomena, not just of a given ML model.[1] Crucially, this presupposes a view of XAI whereupon it is not *solely* concerned with the workings of ML models – a

[1] Note that the embedding into a research process is necessary, and that 'facilitating' is not the same as *constituting* or *being* an explanation. This is similar as in Lawler and Sullivan's (2021) account of model-*induced* explanations. In other words: We are neither claiming that XAI outputs are themselves typically explanations, nor that they can deliver such explanations as stand-alone devices. However, at least the second point only makes for a difference in degree: Any scientific model or representation needs to be interpreted by appeal to background knowledge to facilitate understanding.

view we shall establish in this paper.

As for the correspondence between types of explanations and XAI methods, we believe the relation is (in general) many to many: If one aims to explain a system mechanistically, a saliency map might turn out to be as helpful as a set of Shapley values.[2] Similarly, a saliency map might support both mechanistic explanations and the discovery of a law-like connection in the service or a deductive-nomological explanation. But explanation is generally a pluralistic and contextual matter: even a single phenomenon may admit of different explanations (Bokulich, 2018) and the success of explanation generally depends on the particular aims and interests of explainer and explanation-seeker (Potochnik, 2016; Van Fraassen, 1980). Hence, this should be unsurprising and per se implies nothing about XAI's ability to aid explanation.

However, following especially Schuhmacher et al. (2022), we also put forward the normative thesis that (ii) in order for XAI outputs to fulfill the function as described in (i), they (or the explanations they facilitate) should be treated on a par with genuinely scientific explanations. In particular, this means that they should be made *testable* (Douglas, 2009, 12).

The structure of the paper is as follows: In Sect. 2.1 we offer a brief account of the importance of explanation to science, putting scientific understanding at center stage. Then, in Sect. 2.2, we consider in what sense and how XAI could possibly help to advance science by promoting the understanding of real-world phenomena. We will here distinguish two senses of 'XAI', one of which allows the relevant targeting of real-world phenomena. Our normative thesis, which emphasizes the testability of decidedly scientific explanations, will be supported in Sect. 3.1 by motivating why and in what ways testability matters in science. However, testability is well-known to be a thorny issue. We will hence discuss ensuing problems in Sect. 3.2, to then turn to a nuanced account (Schupbach, 2016) that also puts specific emphasis on the testing of *explanations*. In Sect. 4.1, we show how these ideas relate to scientific practice, turning to an approach from the XAI community that realizes several related ideas in terms of a concrete

[2] We will go into concrete examples in Section 4.

framework. We shall also comment on the notions of ‘representation’ and ‘interpretability’ in this context. All this will be supplement by case studies from the life sciences in Sect.s 4.2 and 4.3, showing how the theses defended here can be (and is already being) applied in practice.

2 From Explainable AI to Scientific Explanation

2.1 Why Value Explanation?

Understanding is a major goal of science (Elgin, 2017; de Regt, 2017; Potochnik, 2017). While there is disagreement on whether understanding is always connected to explanation (Khalifa, 2017; Dellsén, 2020; Elgin, 2017; de Regt, 2017), there is unanimous agreement that explanation is one major vehicle for understanding. But why exactly do we value understanding? Again, opinions differ, but it seems clear that at least one benefit from understanding is that it typically equips us with greater skill for action[3] than mere factual knowledge:

> Understanding is widely held to be a matter of grasping. [...] an important element of grasping is knowing how to exploit the information or insight one’s understanding provides. (Elgin, 2017, 33)

An illustrative case is the role of explanation in the cure of the scurvy disease. As British naval surgeon James Lind could demonstrate, including lemons in the diet of sailors provided an effective remedy against scurvy. Lind’s (1757) study is commonly considered the first randomized controlled intervention study in medicine, and thus meets what is nowadays commonly accepted as the gold standard for the effectiveness of medical treatments. However, the discovery of an effective remedy was far from putting an end to sailors’ suffering from scurvy.

Even after the British navy introduced a daily ration of lemons for all sailors in 1797, unexplained failures of the lemon remedy occurred for more than a century to come (Harvie,

[3] Even though our main example involves physical actions such as interventions, we here intend to also subsume cognitive actions like the drawing of inferences or model-development under ‘action’.

2002). As it turned out over the course of decades, the effectiveness of lemons was severely compromised through procedures attempting to preserve lemons by boiling lemon juice or by storage in copper tubes. Similarly so for the replacement of Mediterranean lemons by West Indian limes. Moreover, it remained an unsolved riddle why most animals would not suffer from scurvy, while humans as well as guinea pigs did. In short, unexplained failures put the remedy into question throughout more than a century.

Explanation of essentially all failures and riddles that accompanied the lemon remedy came about with the identification of vitamin C as the antiscorbutic factor by Szent-Györgyi and Haworth (1933). The discovery of vitamin C in turn elucidated its role in the biosynthesis of collagen (Jeffrey and Martin, 1966), thus uncovering the biochemical disease mechanism of scurvy as a nutritional deficiency disease. The availability of a scientific explanation had an impact beyond resolving unexplained failures: It facilitated the necessary skills to wield one's knowledge in the service of action, as it led to the cheap production of synthetic vitamin C, thus further improving the treatment.

AI is nowadays being deployed in high-stakes contexts such as medicine or pharmacology. Hence, explanations of AI systems should satisfy the same stringent conditions as scientific explanations, in analogy to the case of vitamin C. However, how is this even possible, if XAI targets ML models, whereas scientific explanations target real-world phenomena such as scurvy and vitamin C deficiency?

2.2 How could XAI help to explain real-world phenomena?

There is a well-known distinction between understanding *a model or theory* and understanding *with* a model or theory:[4] Strevens (2013, 513), e.g., distinguishes a genuine mode of understanding wherein "the object of [...] understanding is [...] a theory rather than a phenomenon

[4] Models and theories are distinct in many ways, but we will here treat them on the same footing: as representational devices for making predictions and generating explanations in science.

or state of affairs." This sense of understanding is a "precondition[...] for every explanation" (ibid.), and thus for understanding real-world phenomena (ibid., 512).[5] de Regt (2017, 23) very similarly holds that understanding a phenomenon "necessarily requires [...] the (pragmatic) understanding of the theory that is used in the explanation."

However, since understanding scientific phenomena is usually considered a primary aim of science (de Regt, 2017; Elgin, 2017), there is an apparent tension between the use of ML and science's aims: If we take ML models to be opaque in the sense that it is difficult to understand why their outcomes arise (Beisbart, 2021, 11643), then insofar as understanding of phenomena presupposes the understanding of a theory or model, a science based on ML models falls short of achieving one of its core aims.

There is a debate about the extent to which the conclusion of the preceding paragraph is true: Sullivan (2022b, 128) argues that it is not the opacity of an ML model that potentially hinders understanding, but rather the 'link uncertainty' attached to it, that is, "the amount, kind, and quality of scientific and empirical evidence supporting the link connecting the model to the target-phenomenon". As a main example, Sullivan discusses Schelling's model of segregation: Here, it seems irrelevant how the model is implemented for it to promote understanding of segregation phenomena in cities; all that matters is that we know how to connect its in- and outputs to evidence about people's housing behavior.

It is certainly right that, in order to provide understanding, a model needs to be connected to evidence. However, we doubt that the ability to link an ML model to even lots of high-quality evidence is per se *sufficient* to generate understanding. For example, DeepMind's AlphaFold2, the Deep Neural Network (DNN) whose successful development and deployment was recently honored with a Nobel Prize, allowed scientists to master a task that was unsolved for half a century (predicting protein structures from amino acid sequences). But despite a tight link to masses of existing high-quality evidence about amino acids and proteins, there is reason to

[5] Confusingly, Strevens (2013) calls the *former* kind of understanding 'understanding with'.

doubt AlphaFold2's value in promoting understanding:

> AlphaFold [...] says nothing about the mechanism of folding, but just predicts the structure using standard machine learning. It finds correlations between sequence and structure by being trained on the 170,000 or so known structures in the Protein Data Base: the algorithm doesn't so much solve the protein-folding problem as evade it. (Ball, 2020)

For similar reasons, Räz and Beisbart (2022, 1) have suggested a qualified reading of Sullivan's account:

> If we employ a weak notion of understanding, then her [Sullivan's] claim [...] that understanding with DNNs is not limited by our lack of understanding of DNNs themselves [...] is tenable, but rather weak. If, however, we employ a strong notion of understanding, particularly explanatory understanding, then her claim is not tenable.

As Räz and Beisbart (2022, 12) argue, scientists might gain *objectual* understanding directly from the use of ML models like AlphaFold2, where objectual understanding concerns a subject matter or phenomenon on the whole, as in '*S* understands *P*'; *P* being the relevant topic, subject matter, or phenomenon.[6] Thus, scientists might understand 'the protein universe' better, *directly* from AlphaFold2's outputs. But this contrasts with understanding *why* proteins fold the way they do, where 'understanding why' is usually taken to be intimately connected to explanations.

Stated differently, even though AlphaFold2 has increased our understanding of protein *folds*, it hasn't rendered protein *folding* any more understandable, by virtue of falling short of providing a *mechanism* – something that could *explain* why proteins fold the way they do. Hence, even if some things may be understood more or less directly with the help of models like AlphaFold2, it remains fair to say that the amount of understanding that transpires from opaque ML models is in general rather limited, and to some extent even 'off-target'.

Plausibly, XAI methods are needed to remedy this situation. This is also the conclusion supported by Räz and Beisbart (2022).[7] However, our *reasons* for embracing XAI-based explana-

[6] This might consist in an ability to map out (and 'grasp') dependencies pertaining to a phenomenon (Dellsén, 2020), or in the establishment of a coherent set of beliefs about *P* by means of exemplification (Elgin, 2017).

[7] It is also somewhat tacitly admitted by Sullivan (2022b, 122) when she discussed the relevance of saliency maps. To the extent that Sullivan hence always meant that most of the understanding comes from ML *combined with* XAI (see also Sullivan, 2022a, 1072), we find ourselves largely in agreement with her.

tions as key to understanding real world phenomena slightly diverge from theirs. For Räz and Beisbart (2022, 14–5; emph. added), the core problem is that:

> researchers do not fully understand which features the DNN picks up on, nor *how these features are combined* to produce the final classification. However, understanding how this works means understanding *how the model as such behaves* in general [...] and not how the model relates to a particular [...] target.

We agree that it is the 'features picked up on' that matter, but we disagree that we need to understand "how this works" or "how the model as such behaves in general". For instance, Lipton (2018) argues that it is possible to understand what features an ML model has picked up on without thereby gaining any insight into what goes on inside the model. Similarly, Boge (2022) has recently offered an account according to which the question of *what* the model *learns* and *how* the model *does it* are conceptually entirely distinct.

As an example, consider how particle physicists have experimented with explicitly adding physically inspired information, inferred by appeal to physical laws, to a DNN's training data (Baldi et al., 2014), or even statistically removing such information from the data (Chang et al., 2018). Both procedures provided evidence that a DNN trained to distinguish signal from background data had somehow *autonomously* managed to acquire said physical information: Adding the relevant quantities hardly improved its performance, while eradicating the information on them from the data spoiled the performance entirely. But this did not at all show *how* the DNN managed to acquire or store this information – it did not shed any light on the model's *inner workings*.[8]

[8] We agree, though, that *very often* investigating the model itself is the most promising way to understand what it has learned. This was the case, for instance, in a study by Iten et al. (2020): Plotting the activations of nodes in a specific layer of an autoencoder against certain physical quantities, one can show that the DNN has learned these quantities, but also how they are encoded into it. Hence, while questions about the explanatorily relevant information discovered by the model are *conceptually* distinct from questions about the model's functioning, individual XAI methods may happen to shed light on both (also Boge, 2022)

2.3 Two Views of XAI

How to reconcile all this with (a) the fact that, on the face of it, XAI is concerned with the explanation of ML models, and (b) that understanding a model is necessary for understanding real-world phenomena? To gain clarity on (a), we propose a distinction: As we see it, 'XAI' is the name of a diverse research field, broadly dedicated to the explainability of AI systems. However, there are at least two distinct sets of explanatory targets in XAI, and hence two ways of looking at the field in its entirety: On a *narrow* construal, XAI is plainly concerned with explaining the AI itself. That is, on the narrow construal, XAI provides us with detailed insight into the workings of algorithms used in the field of AI; or in Rudin's (2019, 206) words: "'explanation' here refers to an understanding of how a model works, as opposed to an explanation of how the world works."

On a *broader* construal, though, XAI is concerned more generally with *things to do with* the *use* of AI. Thus, Lipton (2018, 11) writes: "An interpretation may prove informative even without shedding light on a model's inner workings. [...] The real goal might be to explore the underlying structure of the data [...]." In turn, this 'underlying structure' might be indicative of mechanisms in the real world that give rise to specific patterns picked up on by the machine (Raghu and Schmidt, 2020). Thus, an XAI method may not even target an ML model specifically; it may target the *real-world reasons* for the model's success – that is, the processes, mechanisms, and relations 'out there' in the world, leading to specific patterns in the data picked up on by the machine.

As an illustrative toy example (and nothing more), suppose that a classifier is used on a range of test-cases and investigated using LIME (Ribeiro et al., 2016). As is well known, LIME locally approximates the given Machine Learning (ML) model, f, by a distinct model, g, and then provides an interpretable output, such as an image wherein all but the most crucial pixels have been blackened, or a list of salient features within the input, weighted by their relative importance according to g. Since we are working under the assumption that LIME is being used on a range of well-understood test cases, it would thus primarily aid in understanding individual decisions of the machine. Furthermore, patching these together, a general pattern might emerge that might then yield some insight into how f works as a whole.

However, assume that we instead take the model to simply be a highly reliable prediction tool and are not interested in its inner workings. Rather, we might be interested in workings within the domain on which the data have been taken, and of which we have little understanding. Using LIME in similar ways as above, we might be able to find individual features of the data that prompt a given prediction. This information could in turn be used to discern distinctive patterns, allowing us to infer back to a mechanism that explains the presence of these patterns.

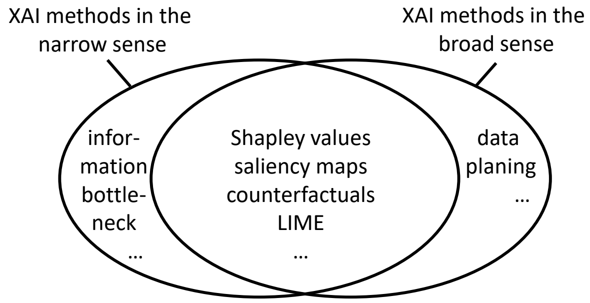


Figure 2: Relation between, and examples of, XAI in the broad and the narrow sense.

However, not all methods fall under XAI in both senses like LIME does (see Fig.2). Chang et al.'s method of removing particle physics information statistically from the data and then checking the ML model's performance, which they called 'data planing', did not explain anything about the model per se: It only explained which information the model was exploiting, not at all how the model did this. Thus, we submit it should fall under XAI in the broad but not the narrow sense. In contrast, the information bottleneck framework (Schwartz-Ziv and Tishby, 2017), discussed also by Räz (2022), can explain how models learn to generalize, without shedding light on what it is that they find. It should thus count as XAI narrowly, but not broadly, construed.

As the LIME-example shows, information gained from XAI methods in the broad, but not necessarily in the narrow sense, can ideally be used to facilitate human-understandable representations or models of aspects of the target domain. These may then be used to generate *explanatory hypotheses*, where a hypothesis is a claim connecting some model or other representation to a class of real-world systems (Giere, 2010, 80). A hypothesis may count as explanatory of a phenomenon if it can figure centrally in the explanation of the phenomenon on (at least) one of the extant accounts of explanation.

We submit that it is this sort of model or representation which needs to be intelligible (needs to be grasped). It may then facilitate understanding of phenomena, by means of facilitating explanatory hypotheses against the backdrop of further domain knowledge. I.e., it is such a representation which we need to understand in order to then understand *with* it—*not*, in the first instance, the *ML model itself*. This is our response to (b): that understanding a model is necessary for understanding real-world phenomena. Clearly, this is conditioned on our response to (a): that XAI must not be construed overly narrowly as being concerned solely with the inner workings of ML models.

Speaking in the abstract, we move from ML model to understanding by using an XAI method, broadly construed, to reach an interpretable representation of some important information the model has extracted from the data to successfully predict. Combining this representation with background knowledge may lead to an explanatory hypothesis which may yield the sought-for understanding. What can be understood depends on the specific combination of XAI method, target system, domain knowledge available and explanatory question asked. We will return to these issues in Section 4.1.

Note also that we assumed a many-to-many connection between XAI methods and types of explanation, as well as explanatory pluralism about even individual phenomena. Thus, what can be explained and by what method will be as case-by-case matter but what is clearly common to all cases is that, if an XAI method is supposed to serve the purpose of explaining real-world phenomena, it cannot be an XAI method purely in the narrow sense.

We will provide some detailed case studies later, but a slightly more realistic example might already be helpful. Turning back to the case of AlphaFold2, there is evidence (Roney and Ovchinnikov, 2022) that it learns at least part of the physics underlying the dynamical generation of possible protein configurations during its training. Assuming that AlphaFold2 somehow transcribes the physics information into its activations, reading out these activations in skillful ways could hence help to significantly constrain the physically realistic models of the folding mechanism. It would then promote understanding of protein-folding, by aiding the genesis of explanatory hypotheses, informed by the skillfully read-out activations.[9]

More concretely, Guo et al. (2022, 2) could recently demonstrate that AlphaFold2 offers about the protein dynamics, when the model's own confidence measure for a given part of the structure is interpreted as a measure of that part's flexibility, and compared with its mobility as predicted by simulation models (Guo et al., 2022, 2). However, this was only possible by means of matching AlphaFold2's outputs to simulation models, which provide explanatory information on part-flexibility.[10]

In sum, we have here argued that in order for thesis (i) to be true – in order for XAI to promote scientific understanding at all – we need to (a) construe XAI broadly: As also comprising methods that render understandable patterns in the data which the ML model exploits, and which may also indicate the real-world reasons for the model's success. And, (b), we argued that it is XAI-based models or representations that need to be grasped in order to facilitate

[9] Methods for skillfully reading out activations in relevant ways are still in their infancy. Iten et al. (2020) simply plotted activations of an autoencoder against known quantities determining the underlying equations of motion, but of course no additional understanding of the real world system can be generated in this way. Wetzel (2025) recently proposed a more advanced method, combining the latent activations of a DNN with symbolic regression. The idea is to embed the DNN in an equivalence class of functions that can be given as an invertible transformation of some closed form expression, and to find the intersection between that class and a relevant class of closed-form expressions. This is a fairly 'theory-free' and innovative approach, but of course, defining the inventory of relevant symbols and configuring the regression algorithm still requires background knowledge, and only a limited set of equations can be effectively generated in this way.

[10] For accounts of how simulations can explain, see Durán (2017); Boge (2020); Schweer and Elstner (2023).

explanatory understanding of phenomena, not ML models per se: If such representations are embedded into a broader research context, this can lead to explanatory hypotheses that provide understanding, whereas the understanding generated by means of ML directly is usually fairly limited.

3 Put it to the Test

3.1 Untestable Explanations are Suspect

What makes an explanation scientifically credible? As is well known, in developing their DN-account of explanation, Hempel and Oppenheim (1948) *identified* explanation with *prediction*. The sole difference for Hempel and Oppenheim (1948, 138) was that in a prediction, a relevant statement was to be derived from a law and antecedent conditions *before* the occurrence of the predicted phenomenon, whereas in an explanation, the prediction came after the fact. This identification was heavily challenged, for instance by Scriven (1962); among other things because it did not seem to fit the explanatory practices in evolutionary biology.

However, more recently, Heather Douglas (2009) has defended a more modest proposal in the broad spirit of Hempel and Oppenheim. According to Douglas (2009, 457) an explanation need not *be* a prediction; in order to count as scientific, it need only make testable predictions *available*. Accordingly, explanations without testable predictions are “scientifically suspect” (ibid., 446) – a view shared by several philosophers of explanation (e.g. de Regt, 2017; Khalifa, 2017).

But why value predictivity? Of course, the idea is that the prediction might come out *wrong* (Popper, 1963; Barnes, 2022; Vickers, 2019). Hence, what makes an explanation credible is, in part, that it sticks out its neck and risks getting refuted. In line with these ideas, our normative thesis thus is that any explanations generated with the help of XAI should be made testable, especially when the target is not the ML model but the reality behind the data (when XAI is construed broadly): In order for XAI to facilitate credible explanations of the real-world reasons for a set of ML predictions, such explanations need to make testable predictions available.

A case in point is the debate over 'just so stories' in evolutionary biology, usually traced to Gould and Lewontin's criticism of (naïve) adaptationism. As Gould and Lewontin argued, it is fully sufficient for the presence of some detectable feature to not have been sufficiently harmful for reproduction and survival so that the Darwinian process did not 'select it away'. However, adaptationists instead seek out the evolutionary *utility* behind each and every trait they see in a biological specimen. The key problem attested to adaptationism by Gould and Lewontin is that it could thus not possibly be proven wrong: With enough effort, there always is *some* evolutionary story one may confabulate (Gould and Lewontin, 1979, 153–4).

Hence, two (interdependent) conditions are crucial for the testability of a given hypothesis, H: (I) That there be definable conditions under which H is refuted or rejected, and (II) that H be precise enough to allow for the formulation of such conditions.

Should we thus reject Darwinian explanations as *unscientific*? Such a conclusion would be too strong and unwarranted: Orzack and Sober (1994, 367 ff.) discuss successful tests of the implications of certain optimality models, which provide explanations in terms of adaptive processes. Whether this amounts to an indirect test of adaptationism itself is another question (see Sterelny and Griffiths, 2012). But this is exactly the point: One should not *assume* that there always is an adaptationist story, but confront each and every individual adaptationist explanation with further data instead.

Similarly, we suggest that one should not take the seemingly most plausible explanation based on an XAI-output at face value, but subject it to further testing. In other words, if we want our XAI-based explanations to be more than 'just so stories' about AI systems or their outputs, we better find ways to subject them to rigorous empirical testing.

3.2 Testability as Explanatory (Dis-)Confirmation

Testability is well known to be a thorny issue. Popper (1959) famously thought that science demarcated itself from pseudo-science by having falsifiable consequences. On his account, theory T would be falsified if a reproducible effect contradicted some consequence of T (Popper, 1959, 66). Since we have put testable predictions and the refutability of explanatory hypotheses at center stage, it might seem that we here follow a Popperian route.[11] But Popperian ideas have long been known to be problematic, due to the arbitrariness of falsification-thresholds (e.g. Spanos, 2019) and difficulties in even defining falsifiability (e.g. Genin, 2022).

We hence instead follow a broadly Bayesian approach (e.g. Earman, 1992; Sprenger and Hartmann, 2019), wherein both confirmation and disconfirmation become gradual, incremental processes. Since we are here interested in the testability of *explanatory* hypotheses, it will be most helpful to adopt a framework wherein hypotheses get rejected for the very reason that they serve as a poor *explanations* of the available evidence. This is the case in Schupbach's (2016) recent account of *robustness analysis*, construed as a competition between different rival explanatory hypotheses.

For our purposes, the core elements of Schupbach's account can be summarized as follows. Schupbach (2016, 292 ff.) presupposes a notion of explanatory power, defined as

$$\mathcal{E}(E, H|B) = \frac{P(H|E \wedge B) - P(H|\neg E \wedge B)}{P(H|E \wedge B) + P(H|\neg E \wedge B)}, \tag{1}$$

which has been extensively justified in (Schupbach and Sprenger, 2011). Here, P is a (regular) probability measure, determining a hypothetical agent's rational credence, E is some new evidence, H some explanatory hypothesis, and B some background condition, consisting of all the past evidence. $\mathcal{E}$ ranges between -1 and 1, where 1 means that H explains E perfectly

[11] Buchholz and Raidl (2022) have recently also applied falsificationsim to ML, following ideas by Gillies (1996). Our investigation is thus orthogonal in topic (and maybe complementary in effect) to theirs, as we are interested in the testability of XAI- based explanations, not in the correspondence between falsification and ML training.

well, whereas −1 means that H would explain the *absence* of E perfectly well, and 0 means explanatory irrelevance. For Schupbach, robustness analysis now is a competition in the sense of explanatory power, where several means of detecting a result have the power to exclude different explanations.

For example, in the famous case of Brownian motion, the first experiments conducted by Robert Brown were compatible with the explanation of the granulated pollen's dancing motion on water in terms of vital forces. However, when Brown also used inanimate materials and the same behavior was seen, this ruled out the vital force-explanation (Schupbach, 2016, 276). Similarly, in the case of Scurvy and Vitamin C, using West Indian Limes and boiled lemon juice ruled out the explanation that scurvy could be cured simply by means of ingesting juices from citrus fruits; it had to be some specific compound found, in sufficient quantities, only in fresh lemons.

We are here not per se interested in the details of robustness analysis in the context of ML or XAI (Freiesleben and Grote, 2023, for an account of ML robustness). For us, the important point is that one explanatory hypothesis, H', can be ruled out by evidence E in favor of another hypothesis, H, if the following conditions hold:

$$\mathcal{E}(B,H) > 0, \quad \mathcal{E}(B,H') > 0 \tag{2}$$

$$P(H \wedge H') = 0 \quad \text{or} \quad \mathcal{E}(E,H|H') \leq 0 \tag{3}$$

$$\mathcal{E}(E,H|B) \approx 1, \quad \mathcal{E}(\neg E,H'|B) \approx 1 \tag{4}$$

Thus, both H and H' may explain the past evidence, B, but H and H' are mutually incompatible; either in the sense that both cannot be reasonably assumed true together, or at least in the sense that assuming one takes away the explanatory power of the other. Finally, and this is the crucial point, H is supposed to explain the new piece of evidence, E, whereas H' explains its absence – just as, in the Brownian case, vital forces could explain the absence of motion in inanimate granules, whereas intermolecular forces inside the water would explain its presence.

Two consequences of the formal analysis are crucial: (A) The above conditions imply that $P(E|H' \wedge B) \approx 0$ and $P(E|H \wedge B) \approx 1$ (Schupbach, 2016, 298). That is, H can very reasonably be said to *predict* E, and H' to predict the *absence* of E. Furthermore, (B), we generally also obtain $P(H'|E \wedge B) < P(H'|B)$, since H''s explaining B will typically ensure that $P(H'|B)$ is sufficiently high, $P(H'|B)/P(E|B)$ is even larger so long as E is not implied by B, and $P(E|H' \wedge B) \approx 0$ will rescale $P(H'|B)/P(E|B)$ by approximately zero via Bayes theorem. Similarly, it can be shown that $P(H|E \wedge B) > P(H|B)$, so these conditions ensure that H receives confirmation from E, whereas H' is disconfirmed.

We submit that this offers a nice way of spelling out Douglas' intuition that explanations should make testable predictions available to count as scientific: There should be some pieces of evidence, E, that H has the power to explain, whereas no available rival H' does. In virtue of E's obtaining, H thus receives a boost in confirmation, and the relevant rivals H' get disconfirmed. If a model in particle physics, containing some new particle, say, explains the measurement of some quantity V within open interval Δ, whereas all rival hypotheses about the particle landscape not presupposing the particle suggest that $V \notin \Delta$, then we can see that measurement $V \in \Delta$ (incrementally) confirms the presence of the particle and disconfirms its absence.

Furthermore, our normative thesis suggests that the same standard should apply to any explanation created with the aid of XAI, broadly construed: For an explanation, H, based on XAI methods to rise to the rigor of a *scientific* explanation, there should be discernible predictions, E, associated with H, which have the power to discriminate between H and its rivals in their capacity to account for E. However, as explained above, for XAI to generate any scientific explanations at all, its methods have to be embedded into broader research processes. It is this aspect that we shall return to in detail next. We shall do so with a special focus on XAI in the life sciences, because several recent developments in this field nicely support our case.

4 Testing XAI-Based Explanations in the Life Sciences

4.1 Views from the XAI Community: The ‘FXAI Framework’

If explanations of an ML model’s success in terms of real-world reasons are to serve as serious scientific explanations, we need to find ways to subject them to rigorous testing by appeal to diverse kinds of evidence, beyond the original training and testing data of the relevant ML model. A framework which takes this message to heart has recently been proposed by Schuhmacher et al. (2022) in the context of medical imaging. This framework is called ‘FXAI’ by Schuhmacher et al., where ‘F’ stands for ‘falsifiable’ (though as we have urged above, this needs to be taken with a grain of salt).

In the usual ML pipeline, as illustrated[12] in Fig. 3 a), data, x_i, are obtained from an experiment or observation $\mathbf{E}_\mathrm{I}$ on some object of study i.[13] These data are used to define the ground truth; either by explicit labeling (as in supervised ML) or by defining, say, a target distribution as a function of x_i (as in unsupervised learning). Crucially, in the usual ML pipeline, the testing thus never escapes the scope of existing data, as $\mathbf{E}_\mathrm{I}$ is not supplemented by additional experiments.

Suppose now that the ML pipeline thus construed was supplemented with an XAI method. It would be highly problematic to infer a purported explanation of the model’s success in terms of real-world reasons by means of the XAI outputs: Maybe the data from $\mathbf{E}_\mathrm{I}$ are peculiar in such a way that, when combined with background knowledge, the XAI output leads in a completely wrong direction. Given the connection between explanation, understanding, and an enriched scope for action established in Sect. 2.1, this should be worrisome: If XAI was used to, say, infer features of tumors from medical images coined tumorous by an ML model, this might lead to the suggestion of treatments that are unsuccessful, as in the scurvy case, or in the worst case even harmful.

Within the FXAI framework, as illustrated in Fig. 3 b), the relevant ML model is directly

[12] Similar illustrations are found in Murdoch et al. (2019) or Roscher et al. (2020).

[13] See Boge (2024) for a recent account of the differences in data-taking between experiment and observation.

supplemented by an XAI method that produces interpretable representations h_i alongside the ML model's outputs, y_i, where h_i stems from some suitable space $\mathcal{I}$ of representations (such as a space of heatmaps, lists of relevant features, simplified local surrogate models, and so on). These representations, h_i, then need to be combined with domain-relevant background knowledge (or 'domain knowledge', for short) in order to generate an explanatory hypothesis, H.

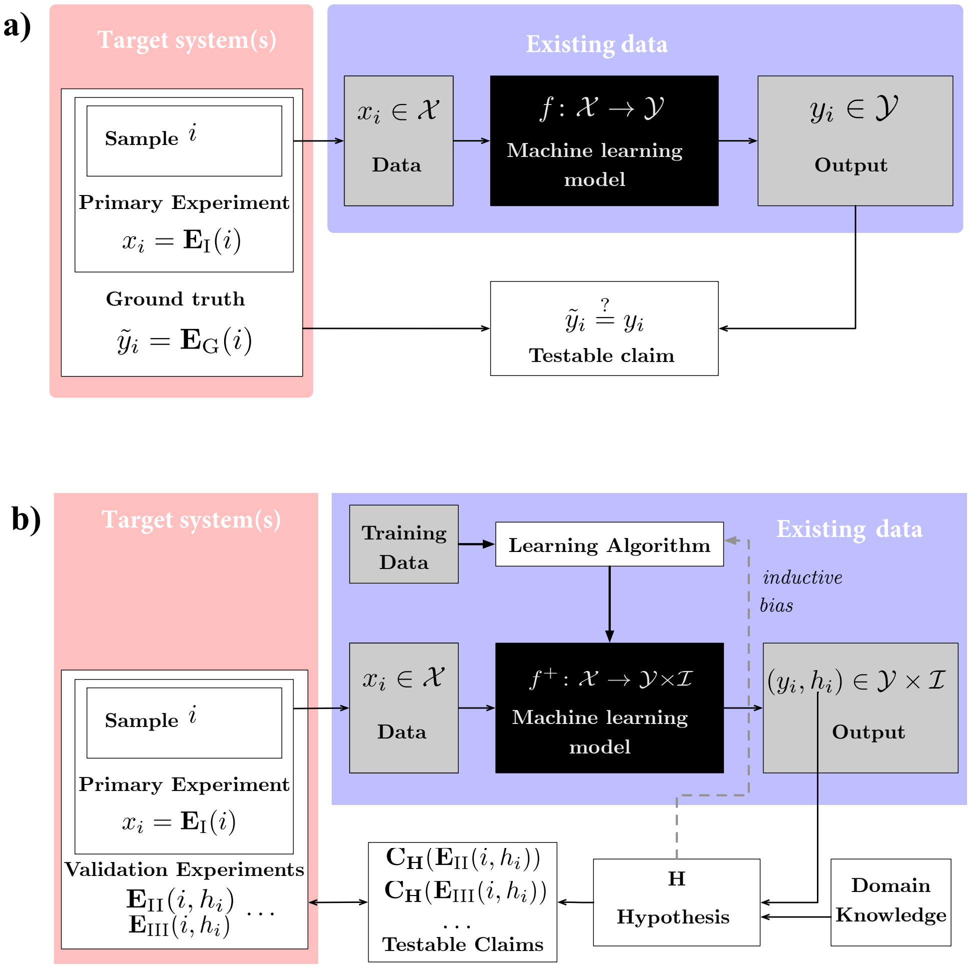

a)
Target system(s)
Sample i
Primary Experiment
$x_i = \mathbf{E}_{\mathrm{I}}(i)$
Ground truth
$\tilde{y}_i = \mathbf{E}_{\mathrm{G}}(i)$
Existing data
$x_i \in \mathcal{X}$
Data
$f : \mathcal{X} \to \mathcal{Y}$
Machine learning model
$y_i \in \mathcal{Y}$
Output
$\tilde{y}_i \overset{?}{=} y_i$
Testable claim
b)
Target system(s)
Training Data
Learning Algorithm
Existing data
inductive bias
$x_i \in \mathcal{X}$
Data
$f^+ : \mathcal{X} \to \mathcal{Y} \times \mathcal{I}$
Machine learning model
$(y_i, h_i) \in \mathcal{Y} \times \mathcal{I}$
Output
Sample i
Primary Experiment
$x_i = \mathbf{E}_{\mathrm{I}}(i)$
Validation Experiments
$\mathbf{E}_{\mathrm{II}}(i, h_i)$ ...
$\mathbf{E}_{\mathrm{III}}(i, h_i)$
$\mathbf{C}_{\mathbf{H}}(\mathbf{E}_{\mathrm{II}}(i, h_i))$
$\mathbf{C}_{\mathbf{H}}(\mathbf{E}_{\mathrm{III}}(i, h_i))$
...
Testable Claims
H
Hypothesis
Domain Knowledge

Figure 3: FXAI framework, reproduced after (Schuhmacher et al., 2022). **a**): In the usual ML pipeline, a primary experiment $\mathbf{E}_{\mathrm{I}}$ on an object i yields data x_i. These may be paired up with targets (labels) y_i for supervised learning, where y_i may be obtained either from a ground truth experiment on i or by annotating x_i. An ML model is then trained to produce outputs that are compared to the targeted ones. The only testable claim derivable from the output y_i is whether it matches the ground truth. **b**) If, by means of an XAI method, an interpretable representation h_i is provided alongside y_i, researchers can generate testable explanations, H, targeting real-world reasons for the model's outputs, by appel to domain-knowledge. Such explanation are only scientific if they predict testable claims C, which make reference also to further experiments $\mathbf{E}_{\mathrm{II}}$,$\mathbf{E}_{\mathrm{III}}$,... on relevant objects i. In principle, a well-validated explanation, H, can then also be used to improve the original ML model via 'inductive biases' (dashed, gray arrow).

By means of such hypotheses, based on XAI together with domain knowledge, further, testable claims may become available; claims, C, that are either probabilistically ($P(C|H) \approx 1$) or deductively ($H \vdash C$) predicted by H. In order for H to even possibly count as a scientific explanation, these claims need to make reference to the outcomes of further experiments, $\mathbf{E}_{\mathrm{II}}, \mathbf{E}_{\mathrm{III}}, \ldots$ on i or a relevantly similar specimen from a pre-defined class of objects (e.g., tissue samples, pharmaceuticals, proteins...). Further, as we argued in Sect. 3.2, the different experiments should in principle have the power to exclude H as an explanation of relevant phenomena, in favor of some alternative H'. Only if it withstands such a test, can H be accepted as 'the' explanation of the relevant phenomena.

For example, Schuhmacher et al. (2022) trained a DNN to classify tumorous and non-tumorous images, but in addition equipped it with a space $\mathcal{I}$ of activation maps to highlight the pixels most important for the classification. This led to certain coherent regions being highlighted which, against the backdrop of relevant domain knowledge, suggested that the DNN had learned to identify cancerous cells. Given, however, that activation maps are known to be problematic as credible representations of the features that prompt successful predictions (Adebayo et al., 2018), an alternative explanation of the DNN's focusing on these regions would have involved, say, bright pixels rather than actually cancerous regions. Hence, to test the hypothesis that activations indeed correspond to cancerous regions, Schuhmacher et al. (2022) compared the XAI-based representations to results from a second experiment, $\mathbf{E}_{\mathrm{II}}$, wherein the highlighting of tumorous regions came from hematoxylin and eosin staining, and this led to results consistent with the tumor-activation identification.

We will return to the example (and also some of its limitations) in more detail below. But as we can see already at this stage, the FXAI framework instantiates a concrete proposal from the XAI community that satisfies our normative thesis: The explanations generated by means of XAI are made testable, in the sense that their explanatory compatibility with a battery of experiments can be probed, which may either lead to confirmation or disconfirmation. Our suggestion is that something along the lines of this framework ought to be accepted if XAI is to serve the overarching aims of science: To explain and understand.

We have grounded XAI's ability to facilitate explanations partly in the 'interpretable representations' provided by them (similarly Fleisher, 2022). So what is meant by this term? Representations are a heterogeneous bunch, but the following aspects are central to all scientific representations (Frigg and Nguyen, 2021, 2022): (i) *Targetedness*: A representation 'stands in for' some other system, the target of the representation. Thus, a model ship may stand in for an actual ship in a lab experiment. (ii) *Asymmetry*: The representation represents the target, but not vice versa. Thus, the model ship does represent the ship, but the ship does not represent the model. (iii) *Inferential surrogacy*: By investigating the representation, we gain insights about the target system. Thus, toying with the model ship in the lab, we may find out things about, say, the real ship's behavior in a canal. (iv) *Graded accuracy*: A representation typically has ways in which it accurately represents, ways in which it *mis*represents, and ways in which it *fails* to represent its target. Thus, the model ship might represent the spatial relations pertaining to parts of the original ship or its mass-distribution, misrepresent such things as material strengths or the workings of the engine, and will fail to represent the actual size and weight of the real ship. (v) *Contextuality*: Whether and how accurately a representation serves its representational functions is relative to a set of aims and purposes of human agents. Thus, for a bunch of scientists, the model ship may represent the real ship but for a kid, it might just be a nice toy.

It is easy to see how (i)–(v) typically apply to the outputs of XAI methods: A saliency map will stand in for the features recognized by an DNN, but not vice versa; by investigating it, we may gain insights into the domain from which the data originate; some things may be well-represented by means of saliency maps, some less well, and some will be neglected; and for a non-expert, a saliency map superimposed on the original image may just be a freaky image. We encourage the reader to consider for herself how (i)–(v) equally apply to, say, lists of features suggested by LIME or the information plane of Schwartz-Ziv and Tishby (2017).

The term 'interpretation' is also used diversely across philosophy and science: Just compare different uses of the term across, say, logical empiricism (e.g. Carnap, 1939), mathematical model theory (e.g. Hodges, 2023), or the philosophy of scientific models (e.g. Hughes, 1999). What uses of 'interpretation' have in common is that they involve one thing that becomes understandable in virtue of it getting mapped to elements of something else: The symbols of a language are equipped with meaning by being mapped to observable things or elements of some models; the elements in a mathematical model are mapped to observable, non-observable, or even fictive things, and we understand what the model 'says' by virtue of that mapping.

It might seem that we have gone full circle, because explanations were identified as something that promotes understanding, and could sometimes be based on XAI outputs. But if XAI outputs are representations that are *understood* by means of an interpretation (similarly Sullivan, 2024, 7), then we are ultimately basing understanding on understanding. However, it is important to distinguish different *forms* of understanding: Strevens (2013, 511) distinguishes 'understanding why' from 'understanding that', and traces the sort of 'grasping' necessary for understanding-why to understanding that something explains something else. Thus, for Strevens, understanding-why presupposes understanding *that* certain explanatory relations pertain. Similarly, de Regt (2017, 40) builds on a notion of *intelligibility*, defined by qualities of a theory that facilitate the theory's use in generating explanations.[14] However, intelligibility essentially involves grasping how the theory works (de Regt, 2017, 102).

The notion of interpretability we have in mind is a close cousin of de Regt's notion of intelligibility. We consider XAI-representations interpretable to the extent that scientists can use them for the sake of drawing inferences about a targeted system, based on an investigation of that representation. It might seem that 'interpretable representation' is thus a pleonasm,

[14] Reference to scientists *use* is well in line with various *stakeholder*-accounts of XAI (Páez, 2019; Langer et al., 2021; Zednik, 2021; Buchholz, 2023).

but it is vital to realize that a representation can fare better or worse in respects (i)–(v), and partly due to the ‘depth’ of the interpretation.[15] For example, an ML model is itself a representation of some connection in the targeted domain. But it is usually not suited to human aims and purposes, in efforts to infer further things about the target. The way we have defined ‘interpretability’, it means a quality of a representation that makes it easy for scientist (but not necessarily for laypeople) to draw inferences based on it. Hence, an XAI-output may equally serve the inferential aims of scientists better or worse, and the degree to which it does defines its interpretability.

4.2 Case Study I: Tumor Localization in Pathology

To show how XAI may indeed become integrated into research processes in such a way as to facilitate explanations and understanding of the underlying subject matter, we now turn to two case studies from the life sciences. The first one deals with image analysis in pathology, and hence extends the example briefly discussed in Sect. 4.1.

In the relevant kind of study, tissue samples, sliced into thin-sections and subsequently captured as a microscopic image, are used for diagnosing cancer and its different subtypes. These image-based diagnoses are then commonly the basis for a treatment decision (Van der Laak et al., 2021). The first task in assessing the status of a tissue sample as cancerous or not is to identify whether the sample contains tumor regions and, if so, to localize these. As deep learning systems commonly assign a disease status to whole images, or at least fixed-sized parts of images, XAI approaches are needed to localize which regions in the image have been identified as tumorous. There are different approaches that can localize which pixels in the input image were most relevant for classifying the image as tumorous, e.g. through local approximations using LIME (Ribeiro et al., 2016), or weakly supervised learning approaches (Campanella et al.,

[15] This too is consistent with Strevens’ (2013, 514) ideas on grasping and understanding-that: “there are degrees of grasping [...]: if you are not completely clear on how the correct explanation of a phenomenon goes, but you have a good grasp of most of the explanation’s elements, then you understand it pretty well but not perfectly.”

2019).

Whichever XAI approach will be used, it will commonly yield a heatmap in the coordinate system of the input image, and it is straightforward to hypothesize that *activation in the heatmap localizes tumor* in the input image. This tumor-localization hypothesis indeed explains the output of an underlying ML model by linking to mechanisms, specifically those that are well-established as the 'hallmarks of cancer' (Weinberg et al., 2000), which here constitute the body of domain knowledge on which the tumor-localization hypothesis builds. However, given especially the problematic status of heat- and saliency maps (Adebayo et al., 2018), it is important to test the tumor localization hypothesis, i.e., whether the regions with high activation in the heatmap exhibit the mechanisms that are understood as the hallmarks of cancer. Clearly, re-inspecting the input image is of limited relevance, as it does not contain molecular evidence about the presence of tumor-driving mutations or the expression of tumor proliferating genes. If the input is a generally un-inspectable hyperspectral image (Schuhmacher et al., 2022), then re-inspecting the input image is even categorically excluded.

In order to facilitate testability, Schuhmacher et al. (2022) suggest to refer the tumor-localization hypothesis to the sample underlying the input image, rather than the input image itself. This facilitates the possibility to conduct further experiments on the sample that explicitly test different traits of the hallmarks of cancer. For example, one can microdissect the image regions with high activation and then profile genomic mutations or gene expression patterns in the dissected regions. However, while several rival hypotheses as to the explanation of the ML model's output can thus be tested on real-world data, the explanation itself refers to the model's reasons for a given prediction rather than some as yet ill-understood mechanisms. In other words: The understanding promoted in this case is indeed understanding of the model rather than the target.

The role of explanatory hypotheses becomes more pronounced in more complex classification tasks in pathology. One such task is to distinguish microsatellite stable (MSS) from microsatellite instable (MSI) tumors, which is crucially important for predicting the success of immune therapy (Le et al., 2015). Microsatellite instability is relatively well understood in terms of the underlying cellular and molecular mechanisms: It is generally accepted that MSI is caused by defects in the genes involved in the mismatch repair mechanism, so that microsatellites as genomic defects accumulate in the coding regions of several genes. Some of the defective genes in turn constitute neoantigens to the immune system, leading to a distinctively altered immune reaction, which leads to a distinctive pattern of tumor infiltrating lymphocytes (Kather et al., 2019). In short, MSI is associated with cellular and molecular processes that involve diverse subregions of tumorous tissue regions and their microenvironment.

Current DNNs that classify MSS vs. MSI achieve clinical grade predictive accuracy (Wagner et al., 2023). They also provide interpretable outputs along with the classification results, usually through heatmaps at a coarse grained level that are obtained from the weakly supervised learning approaches used for training. The most elaborate work in this direction has been conducted by Wagner et al. (2023), who obtain eight different heatmaps through a multi-head attention mechanism. The patterns highlighted by these are consistent with morphological patterns expected on the basis of the current mechanistic understanding of the genesis of MSI-tumors, as described above. For example, a heatmap showed that the DNN by Wagner et al. (2023) used mucinous regions to predict the presence of MSI, which is causally linked to MSI (see figure 4).

The explanation here invokes a mechanism, and is hence *mechanistic* (Machamer et al., 2000): The tissue becoming tumorous is explained in terms of cellular and subcellular processes and properties. Furthermore, it does ‘go out to the world’, as it relates to mechanisms inside the body. However, almost all the work is done by the background knowledge here, which is prior knowledge of a possible mechanism: It is only against this prior mechanistic understanding of

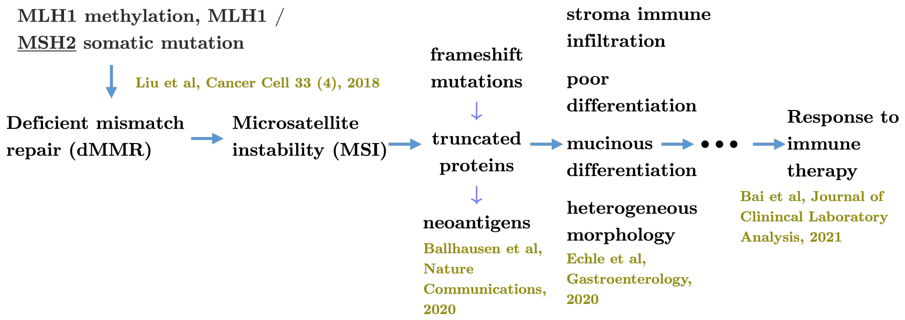

MLH1 methylation, MLH1 / MSH2 somatic mutation
Liu et al, Cancer Cell 33 (4), 2018
Deficient mismatch repair (dMMR)
Microsatellite instability (MSI)
frameshift mutations
truncated proteins
neoantigens
Ballhausen et al, Nature Communications, 2020
stroma immune infiltration
poor differentiation
mucinous differentiation
heterogeneous morphology
Echle et al, Gastroenterology, 2020
• • •
Response to immune therapy
Bai et al, Journal of Clinincal Laboratory Analysis, 2021

Figure 4: Present understanding of the mechanism underlying microsatellite instability in cancer.

MSI-cancers that the heatmaps can be understood as supporting that same explanation of MSI-tumors. A different way to put it is that the explanation makes no *use-novel* predictions (Worrall, 1985): The prediction that these regions should identify MSI-cancers, as they relate to underlying causes, is used to devise the explanation itself; it cannot serve as a test of that explanation, but only whether the prediction involves factors that comply with theory 'already on the table'. Finding and testing such predictions would only have been possible had the heatmap suggested the importance of factors *different* from those recognized in the MSI-mechanism as currently understood (fig. 4).

We shall hence turn to a third study in which the following three factors are met: (I) explanations concern the target, not the model; (II) they are successfully facilitated by the integration with domain knowledge; and (III) they are, and can be, subjected to additional testing by means of additional (and potentially disconfirming) evidence.

4.3 Case Study II: Single- and Dual-Target Behavior in Pharmacology

The third study that we want to look into here comes from pharmacology (Feldmann et al., 2021). In pharmacology, molecules that bind to biological macromolecules and alter their function are usually called 'ligands' or 'compounds', whereas the relevant biological macromolecules themselves are then referred to as 'receptors' or 'targets' (e.g. Salahudeen and Nishtala, 2016; Talevi, 2015). The interaction between compound and target is often conceptualized in terms of a lock-and-key-metaphor:

> The general idea is that the ligand (the key) and the target (the lock) should have complementary features to efficiently interact and trigger some biological response (open the lock). (Talevi, 2015, 2)

However, since in vivo-interactions are complex, this is not generally the case. Cases in which compounds show a rather selective 'multiple target-behavior' (i.e., affect various different targets) could prove beneficial, whereas 'promiscuous' behavior seems generally undesirable (ibid.). Thus, a crucial task in pharmacology is to find out which compounds will exhibit (which sort of) multiple target-behavior. Furthermore, understanding the underlying molecular features that give rise to this kind of behaviour would certainly vastly simplify the task.

In an effort to improve this understanding with the aid of ML, Feldmann et al. (2021) trained a range of balanced random forests and analyzed them using XAI-methods to identify features that relate to dual-target behavior. In a prior study (Feldmann and Bajorath, 2021), it could be shown that no *global* structural features seem to exist that underlie such behavior, i.e., structures that generally determine dual-target behavior. To show this, Feldmann and Bajorath (2021) first trained and tested a range of ML classifiers to distinguish single- from dual-target compounds for a specific combination of compound and target-pair. They then tested these classifiers again on a different target-pair, but could observe a performance typically no better than random guessing. Thus, in case there are structural molecular features that determine dual-target behavior, these appear to be "'local' in nature, i.e. confined to individual target combinations" (ibid., 2)

The second study (Feldmann et al., 2021) then investigated the possibility of such local structures. The key part here was to analyze the importance of certain features at the atomic level (i.e., the presence of specific atoms with inter-atomic bindings) by means of Shapley Values (Shapley, 1953). Recall that Shapley values treat features as 'players' in a game, where the 'payout' corresponds to the ML model's prediction at a given instance minus its average prediction. More concretely, Shapley values are computed by fixing a certain feature to a specific value, combining it with all possible combinations of other features at their various allowed values, and investigating the difference in prediction when omitting the feature in question (Molnar, 2020, Sect. 9.5, for an illustration). Feature-contributions can be negative, meaning that the presence of these features makes the given prediction less likely.

The fact that Shapley values probe for the importance of a feature by investigating its *absence* was crucial for the results of Feldmann et al.'s second study. In a first step, Shapley values were used to select the N most relevant features of a given dual target-compound for the classification as 'dual target'. Then, a smaller number M of most frequently occurring features across all these compounds were selected. These were ranked according to frequency, and the resulting rank was superimposed on the structural feature-representation at the atomic level.

In several cases, this resulted in highly interpretable representations. Figure 5 displays two distinct dual target compounds that both have caffeine as a 'coherent substructure' (highlighted by the Shapley-based ranks). Furthermore, the small number of single target-compounds that do feature caffeine as a coherent substructure as well were almost all incorrectly classified. Thus, against the backdrop of known chemistry, the XAI method strongly suggested that the relevant ML models used the presence of caffeine (and similar coherent substructures) to predict dual target-behavior.

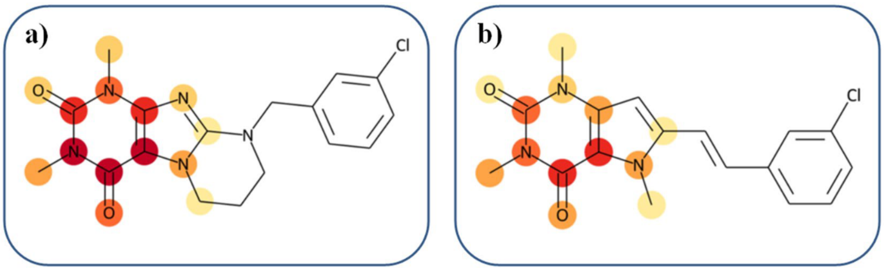


Figure 5: Two distinct dual target-compounds with coherent caffeine substructures highlighted by means of Shapley values. Taken and modified from Feldmann et al. (2021) under a CC BY 4.0 Deed license (https://creativecommons.org/licenses/by/4.0/). Color available online

However, more importantly, it thus also becomes a reasonable hypothesis that caffeine is *causally responsible* for dual target behavior in a certain class of chemicals containing it as a substructure. This is an explanatory hypothesis about the causes of dual targe-behavior, which makes reference to the target domain, and not to the ML model per se. Furthermore, this hypothesis can be (dis-)confirmed using further evidence. Indeed, Feldmann et al. (2021, 7) performed a literature search and found studies which independently confirmed that caffeine derivatives act against both monoamine oxidase B and the adenosine A2A receptor. Specifically, in the relevant study (Pretorius et al., 2008), a broad class of molecules with caffeine substructure was identified that showed dual target behavior, which class was clearly distinct from the sample investigated by Feldmann et al. (2021). In this way, the only reasonable explanation that remained was that dual target behavior is indeed connected to caffeine – and neither an artifact of the investigation nor due to some other factor related to the specific molecules investigated by Feldmann et al. (2021).

Shapley values are known to not always admit of a causal interpretation (Heskes et al., 2020), but in the first place establish a correlation between predictive factors and prediction outcomes for the training, testing, and validation sets. However, the explanation provided in this particular case *is* causal: Caffeine (among others) is identified as causing dual-target behavior, when embedded into some chemical as a coherent substructure. To further confirm this hypothesized causal link, one might – depending on chemical realizability – intervene on the substructure to alter it, and test whether dual-target behavior still occurs. More realistically, one might approximate an intervention by collecting a large group of diverse chemicals with caffeine as a substructure, as well as a diverse control group of chemicals without caffeine as a substructure, and look for dual-target effects across these groups (cf. Woodward, 2003, 95).

The domain-knowledge needed to establish this causal hypothesis concerns knowledge of chemical structures, general knowledge of the relevance of molecular composition to pharmacological behavior, and maybe even more specific knowledge about the potential relevance of coherent substructures to dual- and single-target behavior, for sorting out the causal meaning of the Shapley values in this particular case. Nevertheless, we submit that the thus-interpretable XAI-representation played a non-negligible role in forming the causal hypothesis, and can hence be said to have facilitated the explanation.

In sum, what we have shown with these case studies is that three factors important for the successful facilitation of an explanation of real-world phenomena by XAI can be met in practice: (I) the method is XAI broadly construed, so it may concern the target, not the model; (II) the representation delivered can be successfully integrated with domain knowledge; and (III) there is additional testing. In the tumor-localization case, (I) was lacking in the study on tissue-sample images, and (II) was problematic in the MSI/MSS study, thus also impairing (III). In the study presented here, all three factors are arguably met and we thus consider it an actual implementation of the ideas we have put forward in this paper. Furthermore, the fact that attribution-methods, such as saliency maps, were used in the tumor studies to identify potential causal factors and Shapley values were used in the pharmacology study to do the same thing further supports our claim that the relation between XAI and types of explanation is many-to-many, and that the association between both must be established case-by-case.

5 Conclusions

In this paper, we have argued for two theses, one descriptive and the other normative: that (i) when suitably embedded into a scientific research process, XAI methods' outputs can facilitate genuine scientific understanding. And (ii) that in order for XAI outputs to fulfill this function, they should be made *testable*. We have supported these theses by building on ideas from philosophy of science and from XAI, as well as by showcasing case studies from the life sciences in which relevant implementations of them have recently shown major potential for scientific progress (see, especially, Feldmann et al., 2021).

To defend (i), we have suggested that there are two ways to construe the term 'XAI': On a narrow construal, XAI is plainly concerned with explaining the AI itself (Rudin, 2019), but on a broader construal, it is concerned more generally with things to do with the use of AI (Lipton, 2018). Thus, focusing on the outputs of XAI methods *broadly construed*, we have argued that they correspond to representations of features of the data that are more easily interpretable than ML models themselves, and that the integration of these with background knowledge can give rise to explanatory hypotheses.

Furthermore, building on ideas by Douglas (2009), as well as on the debate on 'just so stories' in evolutionary biology, we have suggested that, (ii), the testability of explanations is key to establishing their scientific status, where 'testing' may be construed as a process of (dis-)confirming rival explanations of some phenomenon by diverse lines of evidence (Schupbach, 2016). As we have shown, these ideas can be applied, to varying degrees, in the life sciences (Schuhmacher et al., 2022; Wagner et al., 2023; Feldmann et al., 2021), and they have recently been implemented as a concrete framework in XAI (Schuhmacher et al., 2022).

Overall, we suggest that our conclusions support a view on XAI and the philosophy of science in which both may profit from closer engagement: By paying close attention to practices in XAI and the sciences making use of it, philosophers can get a clear view of changing and constant practices of explanation and understanding in the age of AI. In turn, practicing scientists may profit from paying attention to the epistemic subtleties associated with scientific explanation and testing, in order to find epistemologically grounded ways for extracting genuine explanations and understanding from XAI, and to ultimately foster progress in science.

Funding: FJB received generous funding from the German Research Foundation (DFG), as part of his Emmy Noether group *UDNN: Scientific Understanding and Deep Neural Networks* (grant 508844757). AM received funding through the *Center for Protein Diagnostics* (PRODI), which is funded by the Ministry of Culture and Science (MKW) of the State of North-Rhine Westphalia, Germany (grant number: 111.08.03.05-133974). Part of his research for the paper was also conducted within the *Slide2Mol* project funded by the Computational Life Sciences program of the German Federal Ministry of Education and Research, grant number 031L0264.

Acknowledgements: We have profited from comments by audiences of the Münster *Ethics of AI (Un-)Explainability* workshop, the annual *Philosophy of Science Conference* at the Inter University Centre Dubrovnik, FJB's Workshop *Epistemological Issues of Machine Learning in Science*, the 2025 *Lamarr Lab Visits* and two anonymous referees. FJB would also like to acknowledge helpful personal conversations with Jürgen Bajorath and Konstantin Genin.

Declarations

Competing interests: None

Ethical Approval: Not applicable

Consent to Participate: Not applicable

Consent to Publish: Not applicable

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). Sanity checks for saliency maps. *Advances in neural information processing systems*, 31.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Baker, A. (2005). Are there genuine mathematical explanations of physical phenomena? *Mind*, 114(454):223–238.

Baldi, P., Sadowski, P., and Whiteson, D. (2014). Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5:4308.

Ball, P. (2020). Behind the screens of alphafold. *Chemistry World*, 9 December. https://www.chemistryworld.com/opinion/behind-the-screens-of-alphafold/4012867.article.

Barnes, E. C. (2022). Prediction versus Accommodation. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2022 edition.

Baron, S. (2023). Explainable ai and causal understanding: Counterfactual approaches considered. *Minds and Machines*, 33(2):347–377.

Batterman, R. W. and Rice, C. C. (2014). Minimal model explanations. *Philosophy of Science*, 81(3):349–376.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. *arXiv preprint arXiv:1704.05796*.

Bau, D., Zhu, J.-Y., Strobelt, H., Zhou, B., Tenenbaum, J. B., Freeman, W. T., and Torralba, A. (2018). Gan dissection: Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1811.10597*.

Beisbart, C. (2021). Opacity thought through: on the intransparency of computer simulations. *Synthese*, 199:11643–11666.

Boge, F. J. (2020). How to infer explanations from computer simulations. *Studies in History and Philosophy of Science Part A*, 82:25–33.

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

Boge, F. J. (2024). Re-assessing the experiment / observation-divide. *Philosophy of Science*, page 1–18.

Bokulich, A. (2018). Searching for noncausal explanations in a sea of causes. In Reutlinger, A.

and Saatsi, J., editors, *Explanation Beyond Causation: Philosophical Perspectives on Non-Causal Explanations*, pages 141–63. Oxford, New York: Oxford University Press.

Buchholz, O. (2023). A means-end account of explainable artificial intelligence. *Synthese*, 202(2):33.

Buchholz, O. and Raidl, E. (2022). A falsificationist account of artificial neural networks. *The British Journal for the Philosophy of Science*. https://doi.org/10.1086/721797.

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

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., and Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309.

Carnap, R. (1939). *Foundations of logic and mathematics*. Chicago: Chicago University Press.

Cartwright, N. and McMullin, E. (1984). How the laws of physics lie.

Chang, S., Cohen, T., and Ostdiek, B. (2018). What is the machine learning? *Physical Review D*, 97(5):6.

Craver, C. F. (2006). When mechanistic models explain. *Synthese*, 153(3):355–376.

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

Cummins, R. E. (1975). Functional analysis. *Journal of Philosophy*, 72(November):741–64.

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

Dellsén, F. (2020). Beyond explanation: Understanding as dependency modelling. *The British Journal for the Philosophy of Science*, 71(4):1261–1286.

Douglas, H. E. (2009). Reintroducing prediction to explanation. *Philosophy of Science*, 76(4):444–463.

Dowe, P. (2000). *Physical Causation*. Cambridge: Cambridge University Press.

Dray, W. H. (1957). *Laws and Explanation in History*. Greenwood Press.

Durán, J. M. (2017). Varying the explanatory span: scientific explanation for computer simulations. *International Studies in the Philosophy of Science*, 31(1):27–45.

Earman, J. (1992). *Bayes Or Bust? A Critical Examination of Bayesian Confirmation Theory*. MIT Press.

Elgin, C. Z. (2017). *True Enough*. MIT Press.

Erasmus, A., Brunet, T. D., and Fisher, E. (2021). What is interpretability? *Philosophy & Technology*, 34(4):833–862.

Feldmann, C. and Bajorath, J. (2021). Machine learning reveals that structural features distinguishing promiscuous and non-promiscuous compounds depend on target combinations. *Scientific Reports*, 11(1):7863.

Feldmann, C., Philipps, M., and Bajorath, J. (2021). Explainable machine learning predictions of dual-target compounds reveal characteristic structural features. *Scientific Reports*, 11(1):21594.

Fleisher, W. (2022). Understanding, idealization, and explainable ai. *Episteme*, 19(4):534–560.

Freiesleben, T. and Grote, T. (2023). Beyond generalization: a theory of robustness in machine learning. *Synthese*, 202(4):109.

Friedman, M. (1974). Explanation and scientific understanding. *The Journal of Philosophy*, 71(1):5–19.

Frigg, R. and Nguyen, J. (2021). Scientific Representation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition.

Frigg, R. and Nguyen, J. (2022). *Scientific Representation*. Elements in the Philosophy of Science. Cambridge University Press.

Genin, K. (2022). On falsifiable statistical hypotheses. *Philosophies*, 7(2):40.

Giere, R. (2010). *Explaining Science: A Cognitive Approach*. University of Chicago Press.

Gillies, D. (1996). *Artificial intelligence and scientific method*. Oxford University Press.

Gould, S. J. and Lewontin, R. C. (1979). The spandrels of san marco and the panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 205(1161):581–598.

Guo, H. et al. (2022). Alphafold2 models indicate that protein sequence determines both structure and dynamics. *Nature Scientific Reports*, 12:10696. https://doi.org/10.1038/s41598-022-14382-9.

Harvie, D. I. (2002). *Limeys: the true story of one man's war against ignorance, the establishment and the deadly scurvy*. Sutton Pub Limited.

Hempel, C. G. and Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, 15(2):135–175.

Heskes, T., Sijben, E., Bucur, I. G., and Claassen, T. (2020). Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4778–4789. Curran Associates, Inc.

Hodges, W. (2023). Model Theory. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2023 edition.

Hughes, R. I. (1999). The ising model, computer simulation, and universal physics. In Morgan, M. and Morrison, M., editors, *Models as Mediators*, pages 97–145. Cambridge University Press.

Iten, R., Metger, T., Wilming, H., Del Rio, L., and Renner, R. (2020). Discovering physical concepts with neural networks: Supplementary materials. *Physical Review Letters*. available online at https://journals.aps.org/prl/supplemental/10.1103/PhysRevLett.124.010508/Supplementary_information.pdf.

Jeffrey, J. J. and Martin, G. (1966). The role of ascorbic acid in the biosynthesis of collagen i.

ascorbic acid requirement by embryonic chick tibia in tissue culture. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 121(2):269–280.

Kather, J. N., Pearson, A. T., Halama, N., Jäger, D., Krause, J., Loosen, S. H., Marx, A., Boor, P., Tacke, F., Neumann, U. P., et al. (2019). Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature medicine*, 25(7):1054–1056.

Khalifa, K. (2017). *Understanding, Explanation, and Scientific Knowledge*. Cambridge University Press.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, pages 2668–2677. PMLR. https://proceedings.mlr.press/v80/kim18d.html.

Krishnan, M. (2020). Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, 33(3):487–502.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., and Baum, K. (2021). What do we want from explainable artificial intelligence (xai)?–a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. *Artificial Intelligence*, 296:103473.

Lawler, I. and Sullivan, E. (2021). Model explanation versus model-induced explanation. *Foundations of Science*, 26(4):1049–1074.

Le, D. T., Uram, J. N., Wang, H., Bartlett, B. R., Kemberling, H., Eyring, A. D., Skora, A. D., Luber, B. S., Azad, N. S., Laheru, D., et al. (2015). Pd-1 blockade in tumors with mismatch-repair deficiency. *New England Journal of Medicine*, 372(26):2509–2520.

Lind, J. (1757). *A Treatise on the Scurvy*. A. Millar.

Lipton, Z. (2018). The mythos of model interpretability. *Queue*, 16:31–57. https://doi.org/10.1145/3236386.3241340.

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

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Orzack, S. H. and Sober, E. (1994). Optimality models and the test of adaptationism. *The American Naturalist*, 143(3):361–380.

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (xai). *Minds and Machines*, 29(3):441–459.

Pearl, J. (2000). *Causality*. Cambridge university press.

Popper, K. (1959). *The Logic of Scientific Discovery*. London: Routledge.

Popper, K. (1963). *Conjectures and Refutations*. London: Routledge.

Potochnik, A. (2016). Scientific explanation: Putting communication first. *Philosophy of Science*, 83(5):721–732.

Potochnik, A. (2017). *Idealization and the Aims of Science*. University of Chicago Press.

Pretorius, J., Malan, S. F., Castagnoli, N., Bergh, J. J., and Petzer, J. P. (2008). Dual inhibition of monoamine oxidase b and antagonism of the adenosine a2a receptor by (e,e)-8-(4-phenylbutadien-1-yl)caffeine analogues. *Bioorganic & Medicinal Chemistry*, 16(18):8676–8684.

Raghu, M. and Schmidt, E. (2020). A survey of deep learning for scientific discovery. *arXiv preprint arXiv:2003.11755*.

Räz, T. (2022). Understanding deep learning with statistical relevance. *Philosophy of Science*, 89(1):20–41.

Räz, T. and Beisbart, C. (2022). The importance of understanding deep learning. *Erkenntnis*. https://doi.org/10.1007/s10670-022-00605-y.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Roney, J. P. and Ovchinnikov, S. (2022). State-of-the-art estimation of protein model accuracy using alphafold. *Phys. Rev. Lett.*, 129:238101.

Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *Ieee Access*, 8:42200–42216.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Salahudeen, M. S. and Nishtala, P. S. (2016). An overview of pharmacodynamic modelling, ligand-binding approach and its application in clinical practice. *Saudi Pharmacological Journal*, 25(2):165–175.

Salmon, W. C. (1970). Statistical explanation. In Colodny, R., editor, *The Nature and Function of Scientific Theories*, pages 173–231. University of Pittsburgh Press.

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.

Schuhmacher, D., Schörner, S., Küpper, C., Großerueschkamp, F., Sternemann, C., Lugnier, C., Kraeft, A.-L., Jütte, H., Tannapfel, A., Reinacher-Schick, A., et al. (2022). A framework for falsifiable explanations of machine learning models with an application in computational pathology. *Medical Image Analysis*, 82:102594.

Schupbach, J. N. (2016). Robustness analysis as explanatory reasoning. *The British Journal for the Philosophy of Science*, 69(1):275–300.

Schupbach, J. N. and Sprenger, J. (2011). The logic of explanatory power. *Philosophy of Science*, 78(1):105–127.

Schwartz-Ziv, R. and Tishby, N. (2017). Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*.

Schweer, J. and Elstner, M. (2023). Dealing with molecular complexity. atomistic computer simulations and scientific explanation. *Perspectives on Science*, 31(5):594–626.

Scriven, M. (1962). Explanations, predictions, and laws. In Feigl, H. and Maxwell, G., editors, *Scientific Explanation, Space and Time*, page 170–230. Minneapolis: University of Minnesota Press.

Shapley, L. S. (1953). 17. a value for n-person games. In Kuhn, H. W. and Tucker, A. W., editors, *Contributions to the Theory of Games (AM-28), Volume II*, pages 307–318. Princeton University Press, Princeton.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Spanos, A. (2019). *Probability theory and statistical inference: Empirical modeling with observational data*. Cambridge University Press.

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

Sprenger, J. and Hartmann, S. (2019). *Bayesian Philosophy of Science*. OUP Oxford.

Sterelny, K. and Griffiths, P. E. (2012). *Sex and death: An introduction to philosophy of biology*. University of Chicago press.

Strevens, M. (2013). No understanding without explanation. *Studies in history and philosophy of science Part A*, 44(3):510–515.

Sullivan, E. (2022a). Inductive risk, understanding, and opaque machine learning models. *Philosophy of Science*, 89(5):1065–1074.

Sullivan, E. (2022b). Understanding from machine learning models. *The British Journal for the Philosophy of Science*, 73(1):109–133.

Sullivan, E. (2024). Do machine learning models represent their targets? *Philosophy of Science*, 91(5):1445–1455.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328. PMLR. https://proceedings.mlr.press/v70/sundararajan17a.html.

Szent-Györgyi, A. and Haworth, W. N. (1933). 'hexuronic acid"(ascorbic acid) as the antiscorbutic factor. *Nature*, 131(3297):24–24.

Talevi, A. (2015). Multi-target pharmacology: possibilities and limitations of the "skeleton key approach" from a medicinal chemist perspective. *Frontiers in pharmacology*, 6:156790.

Van der Laak, J., Litjens, G., and Ciompi, F. (2021). Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5):775–784.

Van Fraassen, B. (1980). *The Scientific Image*. Clarendon Library of Logic and Philosophy. Clarendon Press.

Vickers, P. (2019). Towards a realistic success-to-truth inference for scientific realism. *Synthese*, 196(2):571–585.

Wagner, S. J., Reisenbüchler, D., West, N. P., Niehues, J. M., Zhu, J., Foersch, S., Veldhuizen, G. P., Quirke, P., Grabsch, H. I., van den Brandt, P. A., et al. (2023). Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. *Cancer Cell*, 41(9):1650–1661.

Watson, D. S. and Floridi, L. (2021). The explanation game: a formal framework for interpretable machine learning. *Synthese*, 198:9211–9242.

Weinberg, R., Hanahan, D., et al. (2000). The hallmarks of cancer. *Cell*, 100(1):57–70.

Wetzel, S. J. (2025). Closed-form interpretation of neural network classifiers with symbolic gradients. *Machine Learning: Science and Technology*, 6(1):015035.

Woodward, J. (2003). Experimentation, causal inference, and instrumental realism. In Radder, H., editor, *The Philosophy of Scientific Experimentation*. University of Pittsburgh Press.

Woodward, J. and Ross, L. (2021). Scientific Explanation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition.

Worrall, J. (1985). Scientific discovery and theory-confirmation. In Pitt, J. C., editor, *Change and Progress in Modern Science*, pages 301–331. Dordrecht: D. Reidel.

Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*, 34(2):265–288.