How Theories of Induction Can Streamline Measurements of Scientific Performance

Slobodan Perović & Vlasta Sikimić

Journal for General Philosophy of Science

ISSN 0925-4560
Volume 51
Number 2

J Gen Philos Sci (2020) 51:267-291
DOI 10.1007/s10838-019-09468-4

Springer

Journal for General Philosophy of Science (2020) 51:267–291
https://doi.org/10.1007/s10838-019-09468-4

ARTICLE

How Theories of Induction Can Streamline Measurements of Scientific Performance

Slobodan Perović[1] · **Vlasta Sikimić**[1]

Published online: 7 August 2019

Abstract
We argue that inductive analysis (based on formal learning theory and the use of suitable machine learning reconstructions) and operational (citation metrics-based) assessment of the scientific process can be justifiably and fruitfully brought together, whereby the citation metrics used in the operational analysis can effectively track the inductive dynamics and measure the research efficiency. We specify the conditions for the use of such inductive streamlining, demonstrate it in the cases of high energy physics experimentation and phylogenetic research, and propose a test of the method's applicability.

Keywords Induction · Formal learning theory · Scientometrics · Bibliometrics · High energy physics · Phylogenetics

1 Introduction

There are two broad approaches to identifying optimal conditions for generating scientific knowledge: the operational approach (OA) (e.g. Martin and Irvine 1984a, b; Perovic et al. 2016) and the inductive approach (IA) (e.g. Genin and Kelly 2015; Kelly et al. 2016; Baltag et al. 2015; Kelly 2004; Schulte 2000, 2018). The former seeks to identify the optimal organizational structure of agents of scientific knowledge-production, such as individual researchers, research groups, laboratories, etc. It studies relations between their properties (e.g. the size of groups, their social and cognitive diversity, their hierarchy, the relations between researchers within labs, etc.) and the outcomes of their work as these are revealed in relevant meta-data (publication rates, citations, patents, educational achievements, etc.). The latter approach focuses on identifying optimal and suitably formalized inductive procedures used by the agents or communities to generate reliable knowledge. In other words, OA explores how to optimize scientific groups or networks to reach the best conclusions at

✉ Slobodan Perović
sperovic@f.bg.ac.rs

Vlasta Sikimić
Vlasta.sikimic@gmail.com

[1] Department of Philosophy, University of Belgrade, Čika Ljubina 18-20, FIlozofski fakultet, 11000 Belgrade, Serbia

the group level. In contrast, IA explores how to optimize the scientific learning process in general.

The approaches have developed independently in terms of their focus and methodology. The results of OA are usually published in science policy and research policy journals, with some recent overlap into social epistemology journals, while IA has developed mainly within the field of philosophy of science in symbiosis with relevant methods and conceptual insights in computer sciences.

The main benefit of OA is its immediate applicability to science policy, something enthusiastically exploited by policy makers. Its scope is limited but precise, making its results immediately applicable. For instance, studies based on citation metrics, one of the main tools of OA, have powerful policy implications for relevant institutions. Consequently, in recent years, both the funding and the organizational structure of scientific institutions have been predicated on them to a large extent.

We explore the inductive streamlining of OA that can result in tracking the actual inductive reasoning process by using suitable meta-data, i.e. citation metrics. Via inductive analysis (Formal Learning Theory), we can establish whether the field *justifiably* leads to efficient and reliable convergence patterns and why, by revealing the underlying inductive method. More specifically, the citation metrics effectively trace the exact patterns of this underlying inductive activity within the network of agents (labs and researchers) and identify the properties that make the agents more efficient in pursuing the activity.

This should provide a qualitatively new kind of insight.[1] Alongside the insights into the agents' efficiency in producing scientific knowledge[2] that OA typically aims to provide, our hybrid approach delivers concurrent internal inductive insights into the domain in which OA is applied. In other words, operational analysis based on citation metrics tracks the reasoning patterns in the inductively analyzed scientific pursuit.

It is important to note that this sort of analysis can be applied only to suitable domains wherein agents follow a regular inductive pattern identified by IA. Only in such domains can the application of OA based on citation metrics reflect patterns of reliable convergence on the results (that is, if certain external circumstantial conditions are met for the citation metrics; these are specified in following sections). Thus, even though specific metrics and inductive analysis can be adequately applied only to particular kinds of domains, the example we provide should encourage further exploration of suitably combined inductive and operational analysis (Sect. 5). It is possible that different scientific pursuits may be better analyzed using different inductive approaches. We focus on Formal Learning Theory as it fits the analysis of particular cases we discuss, but, for instance, the much more common Bayesian confirmation approach may be fruitful in other cases.

We demonstrate this hybrid analysis in our two case studies, high energy physics (Sect. 3) and phylogenetics (Sect. 4). In the analysis of the first case, first, we start from an obvious fact: conservation laws constitute the baseline principle of inductive reasoning in High Energy Physics (HEP) and, as such, constitute the baseline of both preliminary informal, as well as various levels of formal inductive analysis of relevant scientific pursuits. This includes inductive reconstruction based on Formal Learning Theory (FLT) as it suitably utilizes machine learning. On the one hand, FLT studies algorithms and rules of inference which constitute reliable inductive methods. In particular, Kelly et al. (1997)

[1] At the same time, meeting the conditions for achieving it will take care of some of the difficulties OA typically encounters.

[2] We define this sort of efficiency more precisely in the next section.

and Schulte (2018) demonstrate that algorithms *favouring simplicity* are more reliable than alternatives. On the other hand, such a process can be reconstructed by modelling algorithms and running machine learning computer programs also based on the principle of parsimony. This symbiosis is the kind of IA we focus on. Second, in this area of experimental physics, where the convergence on results is fast, stable, and relevant over long periods of time (decades), the patterns of agents' (e.g., scientists and experimental teams) production of knowledge can be accurately tracked by the citation metrics (due to the external conditions we specify). Thus, one can trace and compare the efficiency of various research units (e.g. experimental teams) in producing results on which other teams converge.

This sort of analysis is predicated on the fact that data used for the inductive analysis of the method including machine learning reconstructions (i.e. results of the experiments and other relevant physical parameters) and for the operational analysis based on citation metrics (citation counts, size of teams, and duration of the experiments) are data concerning the same domain (units), that is, the same set of experiments. On the one hand, the inductive analysis suggests that the researchers and teams in HEP laboratories are following a reliable method of inference from data, one that ought to result in stable convergence on particular experimental results. The analysis is based on the actual general dataset (the results produced by the experimental groups in HEP in a given period) and the method of inductive reasoning is tested on the key results in the dataset. On the other hand, the citation metrics indicate the exact patterns of convergence on particular teams' experimental results *within the same set of experiments*, i.e., the patterns produced by the same method as that tested by FLT.

The IA utilizing FLT in symbiosis with machine learning reconstructions based on the principle of parsimony demonstrates that this state of affairs is a result of a reliable inductive pursuit: the quick and stable convergence on a particular team's (or lab's) results in the network of experimental teams or labs, traceable by OA based on citation metrics, turns out to be a result of the inductive reasoning of agents as essentially inductive-computing devices, which could be reconstructed via relevant computer models. Thus, in this case, the citation metrics effectively measure the efficiency of pursuing the inductive process. The efficient and inefficient laboratories or teams identified by relevant citation analyses are, in effect, efficient or inefficient at performing the inductive process characterizing the pursuit.

Thus, the FLT kind of inductive analysis argues for epistemically meaningful efficiency analysis (via suitable citation metrics). The citation metrics should not be applied without previous meta-analysis, while the meta-analysis can be directly applied to the assessment of the reliability of citation metrics. The citation metrics represent a quantitative tracing of patterns of convergence among agents (there could be other measures)[3] and can be correlated with various properties of agents (e.g. the structure of teams), but they do not address the deeper reasons for this convergence. The reasons are identified by inductive methodological analysis. The inductive analysis tells us whether and why the method in the field justifiably leads to efficient and reliable convergence patterns, but how exactly this happens in the network of labs and researchers is traced by the citation metrics. Thus, when operational analysis identifies properties of agents (e.g. team structure or funding structure) that turn out to be beneficial to their efficiency (measured by high citation counts), they

[3] Citation patterns happen to supervene on the patterns of reasoning in the network in HEP case because of the external conditions we will specify. That is not always the case, because citation metrics can be messy and out of tune with the actual patterns of reasoning.

are beneficial precisely because the agents better perform inductive process thanks to that property.

In biology, the time scales of convergence on the experimental results are typically much longer than in HEP. This, generally speaking, makes the proposed hybrid analysis much harder and, in some cases, impossible to pursue. Yet analogous to the conservation principle in HEP, the parsimony principle is a baseline principle in phylogeny research. It states that all other things being equal, the best hypothesis is the one that requires the fewest evolutionary changes. As biologists already use a streamlined computing analysis based on the principle of parsimony to parse their data, then, OA might fruitfully be applied to this methodologically (inductively) streamlined pursuit.

Finally, we suggest two general tests—an FLT/inductive test and a general OA test—to determine whether a scientific pursuit can be justifiably assessed by OA in this way. As a final note, we discuss implications of the potential convergence of FLT-based inductive analysis with other inductive approaches, including statistical analysis.

2 Operational and Inductive Approaches to Epistemically Optimal Organization of Scientific Networks

2.1 Operational Approach

The days of a lone observer who publishes her results after a long solitary process of experimentation and deliberation are mostly gone. In modern science, especially in modern laboratories, the researcher constantly acquires, updates, and revises her beliefs based on her relationship with other researchers in a local or a larger network of researchers. This has motivated fairly recent social-epistemological examinations of science in philosophical literature (Kitcher 1990; Zollman 2010; Weisberg and Muldoon 2009). Much earlier, however, the operational studies of science in science policy research embraced the subject. As we will see shortly, we can fruitfully examine and assess the inductive procedures pursued by scientific agents, understood as a process pursued by a lone researcher or by a team of deliberating researchers. In operational analysis, however, the focus is different: instead of generalizing the patterns of reasoning and inferences themselves, the structure of and relations within the networks of researchers are examined as preconditions for generating knowledge. Thus, as a strand of social-epistemological and science policy analysis, OA focuses on identifying optimal ways to organize scientific networks of agents by studying types of connections between agents, structure of networks, their size, and the extent of their centralization and seeks to identify the operating procedures most likely to yield efficiency (in e.g. producing reliable results).

The focus of such analysis is not on a reasoning process and its patterns per se, but on the structure of networks and their different properties.[4] This represents a broad quantitative approach to the analysis of a scientific community, rather than an abstract analysis of the reasoning process. Its upside is the derivation of a quantitative metric of efficiency. It

[4] The results of this sort of research are typically published in science and research policy journals with some recent overlaps with social epistemology. Notable examples, relevant to our argument, include Maruyama et al. (2015), Carillo et al. (2013), Corley et al. (2006), and Martin and Irvine (1984a, b). All these methods of analysis, including computer simulations, were originally developed in Organization Theory in industrial economics (Peltonen 2016).

makes use of the tools of analysis and insights from various quantitative analyses of the organization of scientific institutions.

In citation metrics, perhaps the most powerful and widely used tool of OA, knowledge production and, hence, the optimality of the organization or agent can be measured through relevant metrics of efficiency: the numbers of publications of the results, the citation of the results by others, and their impact in various domains. Yet influential studies often suffer from problems typical of other social sciences research. In fact, the re-examination of the methodology of OA, primarily the use of citation metrics, is ongoing (Bornmann 2017; Alexander et al. 2015; Warner 2000; MacRoberts and MacRoberts 1989) and clear methodological guidelines are lacking (Braun 2010).

First, there are problems with transparency. Murky metrics are frequently applied (Van Noorden 2014) or little qualitative analysis is provided. It is sometimes hard to see what justifies the use of citation metrics in a particular domain other than the sheer availability of data, such as citation records and/or desired research goals. In such a situation, the citations represent little more than a particular property suitable for data extraction, because no further internal coherence of the domain selected for the analysis, e.g. methodological coherence or coherent research goals, has been identified at all. Generally speaking, there seems to be an unexamined assumption in the background of such studies that all researchers within the chosen domain of analysis pursue more or less the same kind of activity in terms of method and goals, e.g. researchers within a particular sub-field working on different tasks or even across diverse scientific institutions, so their production (publication and citation counts) can be justifiably compared.

Second, on a more practical level, the analysed citations can be dispersed across fields without really indicating an expert assessment of the papers or the value of their results. The researchers can take longer periods (years or even decades) to agree on the value of the results, but citations around the time the results were published do not reflect this. In many fields, the convergence on the results does not even occur, and the research remains atomized. On an even more mundane level, the number of published papers and citations can be overwhelming or hard to track, citations themselves can be unreliable for a number of reasons, the relevant papers may have multiple authors who may not equally contribute (Allen et al. 2014), and so on.

Now, although we go on to address some of these difficulties using citation metrics, this is not our main goal; instead, we address them indirectly through the (external) conditions we suggest for achieving a new level of epistemic transparency in OA.

We should mention some fairly recent uses of simulations and decision theory which aim at testing various network structures as models of real scientific networks of researchers and labs. They purport to test reliability and efficiency (time of solving a task) of a scientific pursuit. Results are taken by proponents as informative of the actual properties of scientific networks. Although operational in terms of testing various hypotheses on efficiency and structure of scientific networks, this approach is abstract much like inductive analysis.[5] Obviously, the results are not directly related to concrete target networks, the way they are in OA with the use of citation metrics, so they cannot be directly used to advance science policy aims (Zollman 2010). The improvement of the simulation-target relationship can rely on ever-more detailed simulations (Rosenstock et al. 2017; Borg et al. 2017) or on

[5] In other words, these hypothesis-driven simulations are based on theoretical considerations and can be used to show that a hypothesis about the efficiency of a scientific network is plausible. They stand in contrast to data-driven models, which are calibrated and tested with data.

empirical calibrations of the simulated model. Our proposal to bring together inductive analysis and the relevant meta-data (citation metrics) should provide a more empirically-based matching between abstract analysis and the actual networks to which it should apply.

2.2 Scientific Agents as Inductive Computing Devices

The inductive approach we use here is an agent/belief-centered exploration of epistemic networks. In this approach, it does not matter whether the computing agent is a deliberating collective or a solitary individual, a computer, or a network of computers. The focus is on computing and logical procedures in hypothesis-formation for sorting out data obtained by experiments or observations, whatever the structure of the agents. It is not surprising then that, as we have briefly pointed out, the view of inductive procedures and reasoning optimality in this sort of IA, usually labeled Formal Learning Theory (FLT) (or simply Learning Theory) can benefit from a suitable use of machine learning methods as it has been effectively developed as a theoretical branch of Machine Learning Theory focusing on inductive procedures and parsimony. Thus, the IA framework we focus on is as follows: FLT as a general inductive framework furnishes a theoretical reconstruction of an empirical problem and gives an a priori demonstration of the reliability of various methods of generating hypotheses. Meanwhile, machine learning based reconstructions generate hypotheses on actual datasets based on the principle of parsimony in a certain field (e.g. in particle physics) to try to recover the hypotheses actually obtained in the field, or even make further predictions—i.e. generate the results in accord with the FLT account of inductive reasoning.

The FLT approach treats epistemic agents as computing agents rather than as ideal epistemological agents the way traditional epistemology does. Computing agents are presumed to be governed by inductive rules when they test hypotheses and derive conclusions. FLT asks whether "epistemic utilities, […] personal probabilities, conformational commitments for how to maintain […] [relevant] probabilities, and the rules of hypothetical reasoning" (Kelly et al. 1997, 248) reliably converge on the truth. Optimality is tied to reliability and convergence on the truth. Thus, "[a]n important learning theoretic project' and certainly the key to our argument 'is to determine whether a proposed methodological norm prevents inquiry from being as reliable as it could have been" (ibid., 247). More recent work in FLT seeks to justify methodological norms (like parsimony) by proving that they are necessary conditions for optimal convergence to the truth, rather than surmountable impediments.

Belief revision is treated as a continuous process dependent on contingencies and, thus, as essentially unpredictable. But eventually, in the long run, science corrects itself (Schulte 2000). It is therefore appropriate to search for a rationale for the reliability of the method; although the analysis can never tell us when inquiry converges on the truth, it can identify a more or less reliable way to converge on the truth. The key presupposition is that over time a process should lead to convergence, despite short-run errors. The task is to identify procedures that ensure this. *The goal, then, is to identify general principles and inference rules and demonstrate their reliability, i.e. that they are better at converging on the truth than the alternatives.*

The agents never know what sort of evidence item they will obtain next; and thus they do not know when they have arrived at the truth, but they can still assess the reliability of the pursuit. Following an appropriate method—identified as a set of suitable inductive principles and rules—they will be more likely to arrive at the truth in a substantially more direct and faster fashion than if they are using other alternative methods. Thus, even though

we can never know the final outcomes in specific cases, i.e. whether a scientific pursuit arrived at the truth, FLT approach can determine whether the inductive methods of a scientific field are reliable, relative to the goal of efficient convergence to the truth.

The tools for modelling scientific procedures of reasoning and the search for optimality include causal and neural networks. The cases used in analysis are usually tentative examples or hypothetical tests applied to particular cases (Thagard et al. 1990) or sometimes even to datasets (Chickering 2002) but, generally speaking, this sort of study of the optimality of scientific networks is conducted at an abstract level of analysis. Yet crucial for our aims is that the FLT and a particular application of machine learning based on the principle of parsimony are inherently concurrent: the reliable inductive procedures are identified by FLT (and generated by agents-as-computing-devices). They are composed of principles of parsimony and inference rules and can be verified in concrete cases by generating appropriate algorithms and running computer programs that reconstruct this presumed inductive process. Hence, the implementation of inductive procedures as characterized by FLT, can be effectively tested by suitable computer reconstructions.

3 Case 1: Inductive Streamlining of Operational Analysis of Experimentation in High Energy Physics (HEP)

3.1 Convergence on Actual Experimental Results and Convergence on Truth

In HEP, belief revisions are substantially minimized compared to other scientific fields. The convergence on the results is fast, stable, and relevant over long periods (decades, so far). Even the convergence on major discoveries such as J/psi, top quark, or Higgs boson occurred in a matter of days, weeks, or very rarely, months.[6] And retractions are rare and memorable events worthy of media attention in the HEP community.

The HEP experiments are either unique or almost unique—there is either only one or at most only a handful of similar detectors and experimental machines. Peers take into account the results of a handful of experimental centres, sometimes one or two laboratories, and a limited number of experimental groups. It is thus practically impossible to avoid citing relevant published papers. In addition, the citations of the published experimental results occur almost without exception in journals within the specialized peer group of HEP experimentalists,[7] as they are rarely of interest outside this already very streamlined field. This means there is virtually no failure in tracking the impact of the results in publications. In this respect, despite immense resources, the structure of the experimental HEP network may be like that of experimental science in the 17th century, with its small closed circles.

In other words, the judgement of peers on experimental results is as reliably tracked as it gets by publication and citation rates. The weighted citation metrics straightforwardly indicate which experiments are deemed inadequate, adequate, or fruitful, and, most importantly, without significant danger of divergence or polarization of the produced results in

[6] See e.g. historical accounts of the major discovery of J/psi in the 1970s (Ting 1977), or W and Z bosons in the 1980s (Darriulat 2004), or those of a number of other particles and their properties.

[7] The only recent significant exceptions are journals in astroparticle physics where HEP results are relevant and cited by physicists outside HEP laboratories.

the near or distant future, as is common in other scientific fields, including some other subfields of physics.

In fact, scientometrics, bibliometric analysis in particular, do not often closely track the pattern of peer agreement in the scientific community that produced and assessed the results. The citations are usually dispersed across fields, so they reflect the views of the results by the scientists external to the field in which they were produced. Moreover, certain expert communities fail to take the relevant results into account; i.e. they do not cite them, but they cite the same results produced by competing groups. In such (rather frequent) cases, the bibliometric patterns could be heavily influenced by various social factors external to the actual expert reasoning process. This means the inductive analysis could not possibly be matched properly with the citation metrics.

Thus, HEP has extraordinary traits compared to some other scientific fields, so we can reliably assess and compare the efficiency of the organization of laboratories and experiments based on citation metrics. The HEP laboratories and their organization have been the target of policy studies based on citation metrics (Martin and Irvine 1984a, b; Perović et al. 2016), taking advantage of the fact that this fast and stable convergence is reliably reflected in citation counts.[8] Yet we still need an independent argument that the quick and stable convergence is the result of a reliable scientific pursuit and that the patterns of convergence are not spurious, accidental, or an artefact of some peculiar traits of the scientific network in HEP. If inductive analysis of the method applied in the domain (over the same dataset) can provide such an argument, the citation metrics reveal the inductive process itself and enable the comparison of agents in the efficiency of their inductive pursuits.

If we analyze a field using OA based on citation metrics without previously considering IA, we can end up with inadequate results, even when the data set is big enough. To use a caricature example, if we simply applied citation analysis to publishing in general, assuming that the number of citations correlates with the quality of the output, we might conclude that the British newspaper *The Sun* is the best publication in the UK. More realistically, in some fields, the number of retracted highly cited papers is not negligible. For example, a study by Voinnet et al. (2003) was cited 900 times before it was retracted. In addition, in-depth analyses of 101 medical claims supported by at least one publication with more than 1000 citations show that only one claim led to extensive clinical use (Contopoulos-Ioannidis et al. 2008). Finally, results with a subsequently high impact can be initially rejected by top-tier journals, precisely because the findings are not expected by the scientific community. For instance, the first publications of the Krebs-cycle and the radioimmunoassay for which the authors later received the Nobel Prize were initially rejected by high-impact journals (Campanario 1993).[9]

The cases summarised in the previous paragraph are not suitable candidates for OA, as they are not likely to pass an adequate independent IA test (e.g. of the sort we offer below) in the first place. In any case, the analysis prior to citation metrics and other OA ought to include a thorough analysis of the inductive patterns of reasoning in the field. Once the field successfully passes the IA test, we can responsibly use citation counts, as we now know exactly what we are measuring. OA and IA *in concert* do not eliminate all hidden

[8] It is also significant that the citations are tracked in the most advanced tracking system of that sort; INSPIRE-HEP categorizes citations into six categories, and has been in place for decades, preceding any currently used citation trackers such as Google or Thomson Reuter's WoS.

[9] See also Bornmann and Daniel (2008) for various reasons researchers cite papers for reasons other than acknowledgement of the quality of the results.

possibilities of spuriousness, but they certainly provide independent checks on the analysis based on citation counts.

3.2 FLT-Based Inductive Test of the Pursuit in HEP

IA based on FLT aims at identifying whether the method used in a pursuit could have been more reliable. The reliable method—identified as a set of principles and generated rules of inference—is the one that ensures convergence on the truth better than the alternatives, or is perhaps even the *only one* that provides such convergence and, as such, validates the convergence on the actual experimental results. The method instructs that the agent must have enough evidence items and justifiably believe that the theory is adequate whatever future experimental results throw at her (Schulte 2000).

The limit case of an infinite inductive process tells us nothing about the specific flow of a sequence of evidence items and how it influences inductive judgement. On the one hand, for Bayesian agents, acting within a long enough period, there is always some wiggle room for the convergence on the truth to eventually emerge. On the other hand, in real cases, an actual relatively long-lasting convergence on a particular hypothesis in light of certain evidence items is not simply an empirical fact about a case, but may also tell us something substantial about the stable nature of the sequence of the flow of evidence items in the case. Thus, there is a *critical time constraint* on any particular hypothesis testing: which method is reliable depends on the specific inductive problem for a particular critical length of pursuit.[10] Now, if one researcher can show that her method is more reliable than any alternatives, she can be deemed to justifiably converge on the truth within the critical time. This requires identifying (e.g. by suitable reconstruction) that the principles and rules governing her inductive inferences are demonstrably more reliable than the alternatives.[11] The key to this demonstration in our case will be to show that the rules of inference based on the core principles are restrictive enough over the dataset (actual experimental data); i.e. there are few alternatives or indeed no possible ones whatsoever.

The practicing experimental particle physicists constrain their derivations from data (i.e. their hypotheses) using conservation principles (conservation of the momentum, energy, spin, charge). They choose the conservation principles which effectively rule out as many unobserved particles as possible, the existence of which would violate them. Thus, in practice, the analysis of particle trajectories is based on the conservation laws; e.g. different potential identities of particles are calculated based on the assumption of the conservation of the momentum of the in-coming and out-going particle tracks. In other words, these physicists opt for the closest fit with the data given a strong simplicity preference.[12] The most likely outcomes are selected based on the obtained data and phenomenological (rather than high-level theoretical) models, used for simulation runs, essentially predicated on the conservation principles.

Although it is hard to deny that "the research program for searching for selection rules [in particle physics] has justified itself by its success so far" (Schulte 2000, 776), the IA analysis should independently reveal the link between the convergence in practice and the

[10] This is analogous to the statistical significance in Neyman–Pearson hypothesis testing. This fact could be exploited further, but it is not one of the goals of our analysis.

[11] In disciplines in which several inductive methods are formally justified, the disagreement in the field will be justified as well. Thus, we will not be able to talk about reliable convergence of opinions.

[12] See e.g. Dissertori et al. (2003).

inductive reliable convergence on the truth. The discovery of new particles, even very surprising ones, is always possible, but the point is that the stream of actual discoveries based on the sequence of evidence (particle interactions) in the pursuit is the product of an inductively reliable method.

In fact, without the parsimonious use of conservation principles it would be hard to imagine modern HEP experimentation. The methodology of the field reduces to it in a fairly straightforward way. This is why physicists themselves have been motivated to reconstruct the inductive method driving the field, and this is why it is a perfect toy-model for our argument. The conservation principles are the baseline principles of HEP practice and also can be identified as the baseline principles of the inductive process. Thus, both real-world practical derivation procedures and IA based on FLT will make recommendations through inference rules based on insights bounded by this same baseline. Now, the baseline principles, in effect, generate suitably applicable selection rules over the experimental dataset by providing a restrictive system of inferential rules.

The computable inferential mechanisms that adequately grasp the actual inductive process in particle physics have been investigated (Schulte and Drew 2010; Valdés-Pérez and Żytkow 1996; Valdés-Pérez and Zytkow 1996; Kocabas 1991). Suitable algorithms and models have been constructed and even used for the discovery process, where the inductive process is modelled and computed based on little more than the conservation principles over a dataset. Thus, the pursuit has not only been modelled as an inductive process in agreement with the FLT inductive framework but also been implemented in the actual pursuit.

These models were either supplied with given constraints or built from scratch. Using the conservation principles, if the simplest model does not capture a hypothesis or a set of data, a more complicated one is used.[13] As an example, the standard quark model was reconstructed through such computations, but "[p]robably the most significant result is that an exhaustive search in the space of quark models for baryons followed by the mesons reveals the standard quark model stands out nearly uniquely as the simplest, when the constraints of complementary pairs is imposed" (Valdés-Pérez and Żytkow 1996, 2109).

In fact, the key part of this formalization is the proof of the restrictive selection of the rules: "Under pure induction (i.e. without additional assumptions)" other than those provided by conservation principles applied to the dataset, "more than one selection rule and quantum property are never needed to distinguish any set of allowed [particle] reactions from any set of prohibited ones" (Valdés-Pérez and Zytkow 1996, 172). This characterization applies to a somewhat simplified model, but computing based on more robust models shows unique determination as well, as demonstrated by Schulte and Drew (2010). In fact, the constraint on the selection rules is strong in all models: in general, assuming conservation laws, the number of selection laws that are not redundant turns out to be small.

As Schulte (2000) points out, it is precisely this restrictiveness that warrants physicists 'projecting the theory': based on it, they are justified in expecting that the theory will be valid for some future expected evidence. Now, since the reconstructed methods of selection based on conservation principles warrant this expectation, the fast and stable convergence on the results we encounter in practice is warranted. The models thus reconstruct an inductive method that generates the procedures and results concurrent with those used by

[13] Simplicity is defined as the number of constituents and the number of constituents per particle (Valdés-Pérez and Żytkow 1996, 54).

physicists for discovering particular particles (and one could even formally show this[14]). It would indeed be difficult to imagine a realistic (in terms of base-line principle and inferences) reconstruction of the pursuit that veers far from such models. Thus, given the results of the reconstructions, the pursuit is based on a reliable inductive method, and projecting the theory is justified in the actual pursuit, as we have the same baseline and parameters (conservation rules and evidence items) in both IA and practice.

We can, in fact, generalize this case with the IA (FLT) test, i.e., a test of the inductive coherence of the pursuit. The conditions for judging whether a pursuit is FLT-inductively coherent are the following:

1. There are computable models *matching* the actual pursuit (over a relevant set of actual experimental data).
2. There is a common core to these various models: a base-line inductive principle and a set of restrictive rules of inference.
3. By successfully computing, i.e. providing successful retrodictions and predictions of key results—over the dataset via restrictive rules based on the postulated principle—the models *warrant* and *explain* the actual fast and stable convergence of researchers on the results.

3.3 Citation Metrics and the Efficiency of Scientific Networks in Pursuing Inductive Processes

In a nutshell, if a pursuit passes the IA test, the OA analyst is reasonably justified in treating the fast and stable convergence on the results as an indicator of the use of a reliable method. Moreover, the OA of the team structure in HEP labs, applied within the same dataset, can provide deeper insight beyond the operational traits immediately captured by such analysis. In other words, the reasons for the convergence on the results are identified by inductive methodological analysis: it identifies, in general terms, whether and why the method in the field justifiably leads to efficient and reliable convergence patterns. Yet it cannot identify how exactly this happens within the network of labs and researchers. Instead, the operational analysis based on citation metrics quantitatively traces the exact patterns of convergence among agents: it identifies the efficient experiments that produce results on which other teams will converge, as well as the inefficient ones, and correlates the efficiency of experiments with the properties of the agents performing them (e.g. the size of teams and division into sub-teams, or their funding structures).[15] Whatever properties of agents turn out to be beneficial to efficiency, as identified in operational analysis, we know they are beneficial because the agents perform a better inductive process thanks to that property.

Let us gradually unpack this argument, starting with relevant examples of bibliometric analysis.

[14] There is no need to spell out the proofs here; they can be found in Schulte (2000).

[15] We can thus identify a temporal constraint on the applicability of the citation metrics and the reasons behind it: the long expiry dates of citation-metric analysis in certain cases (e.g. HEP) are determined by the justifiably long-term convergence on the results in the pursuit, as the revision of beliefs is justifiably minimized. Apart from establishing reliability of the results, IA has the potential to establish the computational properties of a scientific pursuit. For instance, Schulte has investigated the NP hardness of finding a simplest linear causal network from conditional correlations.

Table 1 Numbers of highly cited papers across HEP laboratories within one year in the period 1969–1978; *n* stands for the number of citations

	n ≥ 15	n ≥ 30	n ≥ 50	n ≥ 100
CERN	111 (26%)[a]	31 (26%)	9 (19.5%)	1 (9%)
DESY	20 (4.5%)	9 (7.5%)	3 (6.5%)	0 (0%)
Brookhaven	37 (8.5%)	6 (5%)	2 (4.5%)	1 (9%)
Fermilab	106 (24.5%)	37 (31%)	17 (37%)	1 (9%)
SLAC	75 (17.5%)	21 (17.5%)	11 (26%)	6 (54.5%)
Others	80 (19%)	15 (13%)	4 (8.5%)	2 (18%)
World total	429	119	46	11

Data based on Martin and Irvine (1984a, b)

[a]All percentages are rounded to the nearest 0.5%

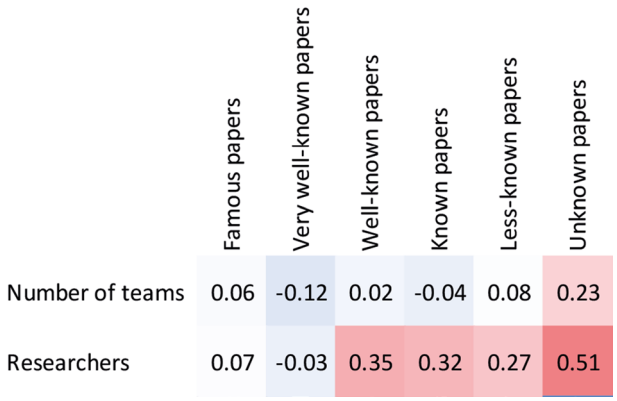

Fig. 1 Results of the data envelopment analysis comparing efficiency within a series of experiments performed at Fermi National Laboratory based on weighted citation counts (data based on Perović et al. 2016). Positive correlations between weighted citation counts and size of teams are represented numerically and coloured in red. i.e. darker tones, while negative ones in blue, i.e. in gray and white tones. The results show that the number of unknown papers (based on weighted citation counts) increases with the number of researchers. However, more researchers do not produce more famous or more very well-known papers

In perhaps the most comprehensive study of its kind to date, a three-part assessment (Martin and Irvine 1984a, b) of the performance of CERN with respect to other HEP laboratories, as well as the performance of individual accelerators of the laboratory, offered various quantified results with the ambitious normative intention of improving the performance of experimentation in HEP as a whole (Table 1). The number of published papers and citations were used as a key metric in this extensive comparative study of the performance (production) of major HEP laboratories. Thus, various sort of operational analyses based on these citation metrics can be performed, tallying, for instance, the efficiency of individual laboratories (e.g. their production of highly cited papers indicating fruitful experimental results) with the extent of their hierarchical structure, or with the resources their teams spend.

A more recent study (Perović et al. 2016) conducted on data from Fermi National Laboratory was based on the data from 27 large similar experiments[16] with the goal of computing their efficiencies in relation to the team sizes (Fig. 1). The most inefficient experiments in the quantitative study turned out to be those of the largest teams in the group; they either stalled at the level of realization, or the protocols were so flawed that the data analysis could not be completed. The most efficient teams, those who excelled in weighted citation counts of the publications based on the results in the experiments performed by the teams, were small. These results concur with other similar studies across scientific fields (van der Wal et al. 2009; Bonaccorsi and Daraio 2005).

Regardless of the exact details and exact implications of the results of published studies, what really renders the use of weighted citation rates valuable is the fast and stable convergence on the results by real experimental networks and in HEP in general. As noted earlier, the citation metrics indicate fast (relative to other fields) and very stable peer agreement on the experimental results. In addition, the field is unusually isolated: researchers publish and cite others in their own journals, they are not cited by external sub-fields, and the experimental centers are few and far between. Retractions are very rare in HEP labs, so disagreements are never that deep. They may concern the precision of the results or a deeper explanation of a phenomenon at stake, but this only testifies to the fruitfulness of the paper being cited. Thus, the convergence is *accurately reflected* in the citation metrics. (Let us call this *an external condition* of the OA.)

Now, on the one hand, the FLT analysis suggests that the researchers and teams in HEP are following a reliable method of inference from data, one that ought to result in stable convergence on the results. The analysis and the models produced are based on the actual dataset (the results of the experimental groups in HEP), and the method of reasoning is tested on the key examples in it (the key results). On the other hand, the citation metrics indicate the exact pattern of convergence within the same general dataset produced by the same method the FLT identified via its models and their testing. Thus, the citation metrics will track convergence in that same environment by selecting a set of experiments from the larger dataset.

The extent to which the dataset of a particular citation metrics matches the dataset used in the actual IA test can vary. Overall, however, it would be hard to argue that the fast and stable convergence is an accidental outcome unrelated to the inductive pursuit of the outlined sort: based both on the informal assessment of the inductive process and on the computable models broadly matching its various aspects, for all practical purposes, the agents in the pursuit act as inductive-computing devices of specific traits that operational analysis identifies based on citation counts.

What we aim at with our argument is to point out their general agreement over the same general dataset. The semi-empirical FLT models we have described above test the method with respect to some of the key results produced by a set of experiments in HEP, dating from 1960s to mid-1980s. The operational analyses rely on the citation metrics metadata over the datasets of the experiments in that same period or, in the case of Martin and Irvine (1984a, b), almost the entire set of results of those experiments. They pivot on the convergence patterns of the same dataset used for the inductive analysis. This very general matching of two kinds of analyses over the dataset constitutes a preliminary argument for hybrid analysis: the method in HEP passes the IA test over the general dataset (the

[16] Experiments are similar—i.e. homogenous in terms of techniques and other traits of the experimental process—yet varied in terms of their efficiency.

inductive models generate the key results on which teams converged), while the citation patterns mirror the patterns of convergence of selected data regions (i.e. selected sets of experiments) in that same dataset.

The existing inductive models (reconstructions) we have cited are concerned with the abstract level of relevant physical theory. The actual experimental searches in laboratories, however, use concrete phenomenological models constructed in accord with the Quantum Chromodynamics (QCD) (postulating quarks as elementary units of hadrons, e.g. protons and neutrons) and the Standard Model of particle physics.[17] Now, the existing inductive models that we discussed reconstructed the process that led to the discovery of some of the key properties in QCD; i.e. they reconstructed quarks from the actual experimental results. These experiments belong to a wide domain of the experiments assessed by Martin and Irvine (1984a, b); the inductive analysis reconstructs some of the exemplary cases in the domain. Thus, the results of the existing inductive models of the pursuit based on computer reconstructions are a relevant indication of the kind of inductive process taking place in the entire pursuit.

In the case of more narrow studies such as Perović et al. (2016) the analyzed experiments explored particle dynamics within the already established QCD framework rather than pursuing the discovery of the core properties of QCD (i.e. quantum numbers). Thus, the inductive reconstructions and models warrant the core inductive strategy in the pursuit; that is, they encompass the actual experiments which are subject of operational (citation-based) analysis. We could certainly create more custom-made matching inductive models that suit the pursuit within the specific dataset in the operational (citation-based) studies.

This is the first and admittedly abstract preliminary step of this sort of hybrid analysis, but we argue it is valid. The inductive test applies very generally across the dataset, and it is hard to see what other method could have been employed in any of the experiments in the selected segments of the dataset for citation analysis instead of the one identified by the IA test. Analyses with more exact matching datasets that will tune the two kinds of analysis to the desired precision would certainly be the ultimate goal, with the potential to offer the strongest argument in favour of hybrid analysis. It should be noted, however, that even at the level of this preliminary general matching between reasoning patterns in inductive models on the one hand, and citation metrics and identified beneficial operational properties on the other, the exactness and the justification of the analysis exceed those of the usual modelling and simulating of scientific institutions with very abstract models and computer simulations, since in our case, both the inductive models and the citation patterns are supplied with empirical data from the same dataset.

The cited research follows a strict inductive pattern. Therefore, these citation metrics are not simply measures of social dynamics in the field but trace the teams' inductive patterns of reasoning. The citation metrics in this case, then, can be considered a reliable measure of productivity, i.e. the efficiency of experimental groups (within a laboratory or across laboratories) in producing results that will guide new research. In HEP, the weighted citations track peers' views, and this is informative for truth-tracking in this discipline. The context of quick and stable convergence, namely the inductive reasoning dynamics in the network, is *the internal factor* streamlining the agreement. It indicates that *the experimental teams identified as efficient outliers or more productive laboratories—based on weighted citation analysis in the above-mentioned studies, directly reflecting the converging peer view—are*

[17] Both are constructed in accord with even higher level of physical theory, Quantum Field Theory and Quantum Electrodynamics.

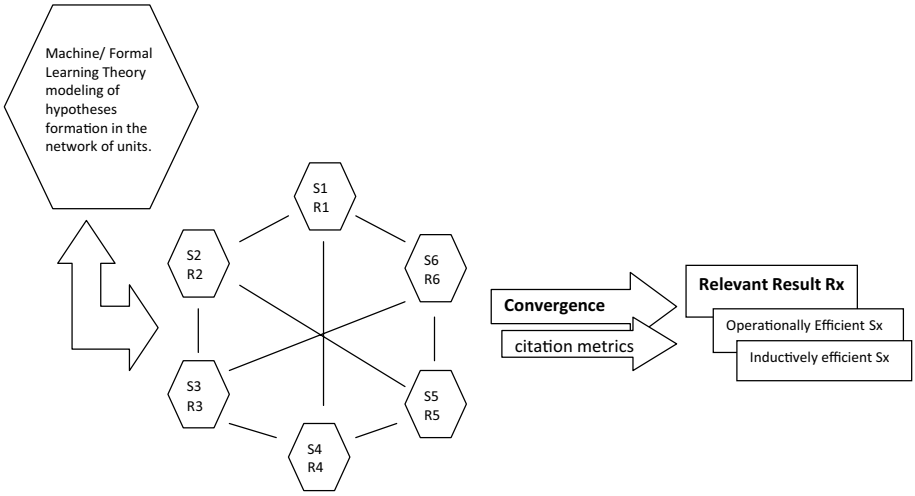

Fig. 2 (1) A network of experimental teams, different in terms of their structure S (number of researchers, hierarchy, knowledge background, etc.), each producing results R. Eventually, they stably converge on the result Rx. (2) Operational analysis identifies the exact convergence patterns (or the lack of them) on the results with *suitable* meta-data (e.g. citation patterns in a highly isolated field) identifying efficient unit S. (3) The inductive analysis (Formal Learning Theory including machine learning reconstructions) generates models of the hypothesis-testing method implemented in the network of the teams that, if successful, produces convergence on the actual result Rx. (4) The convergence patterns on R in the network identified by operational analysis are due to the appropriate method producing R. The operational analysis tracks the team (of structure S) most efficient in implementing inductive reasoning dynamics in the network

significantly better than the other teams at the performance of the inductive process that characterizes the pursuit in question. Thus, suggesting that other teams should be more like the most efficient teams in terms of the measured operational parameters (team size, number of teams, etc.) is not an operational 'shot in the dark', and the convergence result of spurious or accidental correlation, an artefact of the network, or simply a result of an unknown parameter, since we now know that the pursuit is inherently a specific inductive process and, as such, is effectively tracked by citation metrics.

We can now list the following conditions that a scientific pursuit should satisfy to be deemed suitable for the OA test:

1. Internal condition: The pursuit passes the FLT test.
2. Empirical condition: Fast and stable convergence on the results in the pursuit.
3. External condition: Convergence is suitably reflected in publication and citation counts.

If a pursuit passes the OA test, the citation metrics effectively measure the efficiency of the inductive process in the scientific network (Fig. 2).

Finally, it should be noted that many experimental outcomes are not reported, as vast numbers of runs of the colliders are based on a great variety of triggers (algorithms that determine the conditions under which the events in the detector will be recorded), the vast majority of which turn out to be of no significance. This is the only reasonable solution given that only particle 'signatures' (decays), not energy scales at which novel particle interactions take place are predicted by the models, so the experimental task is to

relentlessly comb vast backgrounds (i.e. known interactions) across the energy scales.[18] For the results deemed worthy of publication, although the efficiency metric based on weighted citations in HEP is first and foremost a measure of how reliable the results are judged to be by peers, it is also a direct indicator of the fruitfulness of the results. It tells us how excessively the peers relied on and were motivated by the results of the experiments to further their own work.[19] Fruitfulness gives the edge to efficient outliers over other reliable results. Experiments are fruitful when they confirm or explore a cornerstone of the model so the experiments succeeding them are bound to rely on their results. Thus, they become an indispensable part of the background knowledge of future experiments. In other words, they act as essential data constraints on the event selections within the relevant framework of the conservation laws, both reinforcing and projecting the theory.

It is possible that further parameters not identified by the study we cite are relevant to the outcome of small teams being efficient and large teams inefficient. Further qualitative study would deepen the explanation, based on an adequately identified meta-property (small teams). It is up to further research to identify the reasons why smaller teams are efficient (e.g. beneficial flattening of the hierarchy in smaller teams or particularly able scientists tending to congregate in smaller teams).

What IA clearly contributes to OA is the ability to identify what exactly the teams are supposed to be good at, while OA identifies the key meta-property (small teams) that makes them good at that activity. The small teams were successful in pursuing the inductive process (that IA identified) in the field. Arguably, however, the most able scientists picked the small teams and that is why the small teams were successful; the same scientists could have made large teams successful as well. If so, good for them, because they knew what teams to pick to perform the tasks, but clearly nothing in the data suggests they could have achieved the same rate of success with large teams—quite the contrary. We could call this the 'Nelson-effect' by invoking a loose but instructive analogy.

Captain Horatio Nelson was extremely successful in sea battles, an outlier in every conceivable sense, because he was particularly skilled. Yet he was skilled, among other key things, in deliberately picking small fast ships instead of large ones with lots of firing power but slow (e.g. he initially used such a boat to intercept and stop an enemy line of large ships that his line of ships caught up with and destroyed). The skill of picking an adequate ship, i.e. small and fast, made him extremely successful. Perhaps Nelson could have achieved the same rate of success with a large ship, but that sort of scenario would require imagining a very different actual context and changing the key parameters of the sea battles; small ships sailed by others could have ended up being inefficient when used the way Nelson used them, thus showing Nelson was efficient independently of the small-ship choice.

[18] Most experiments do not purport to establish the existence of new particles; rather, they explore properties of the known ones. The Standard Model is a null hypothesis in the vast majority of experiments; it provides the expected background interactions, so the exploratory experiments that do not turn up new particles will be null experiments—but they will also provide important information on their properties (e.g. energy scales) that the model does not deliver. Even if an experiment that does not have any results of significance is reported, it will not result in the number or quality of citations that accompany experiments with confirmatory results.

[19] This was certainly true of the citation patterns of the experiments from the late 1960s to the mid-1990s—the period analysed by the above-outlined studies; now research has become so centralized that essentially all particle physicists are engaged in one mega-project.

To come back to our key example, at the very least, the data suggest the identified (with FLT) inductive process is more likely to be performed in an adequate way if research is done with a smaller team. As far as the empirical data can reasonably show us, small teams matter as a meta-property, although further qualitative analysis is needed and even though fictional and increasingly unrealistic scenarios, counter to the actual data, could be always concocted. The burden is on further research, as precise and as data-driven as the type at stake, to show that such scenarios are plausible, let alone supported by the data.

Could it be, however, that the production of the results that turn out to be fruitful, or in other words, the measured efficiency, is a result of serendipity rather than a particularly efficient pursuit of the inductive process? Perhaps the successful efficient laboratories or teams simply stumble on fruitful hypotheses, while the inefficient ones are merely unlucky in their choice.

There are, in fact, three different levels of inefficiency tracked by citation metrics. The least efficient experiments have a problem at the level of 'cables': they do not operate the equipment well and never really take off even though they consume lots of resources (Perović et al. 2016; Martin and Irvine 1984a, b). In other words, the experimental team as an inductive-computing unit has a hardware problem. Then there are those who get stuck at the level of data analysis for various reasons. In such cases, essentially the unit cannot compute well—its deficiency is analogous to software deficiency in performing a computation. Finally, some teams never reach the highest level of efficiency because the hypothesis they test is not fruitful enough.

So it seems only in the last case could we justifiably suppose serendipity is a major factor. Yet perhaps a team makes the initial choice of the hypotheses to be tested as part of its inductive-computing process. The question, then, is to what extent the choice and formation of the hypothesis (or rather a set of hypotheses) for testing is: (1) part of the overall inductive-computing process, and (2) shaped by parameters entirely external to the inductive pursuit.

Either way, the difference between stellar and mediocre results is decided at the higher level of the inductive-computing process; in other words, at the level of the choice of algorithms, i.e. the choice of generating rules. The selection rules are generated over an existing data set, so the formation of the list of potential hypotheses for testing is very restricted. We would really need to see how each list was created (along the lines of e.g. Maruyama et al. 2015) to address the possibility of serendipity at this stage of testing; e.g. it is crucial to know how sub-hypotheses are produced from a very general master-hypothesis delivered by the Standard Model or any of the alternative models in particle physics. In any case, smaller teams turn out to produce more fruitful results so it may be that smaller teams are better conditioned for superb computations at the level of picking algorithms/hypothesis, possibly because they are demonstrably better at handling hardware and software.

Finally, when the phenomenon of significance is discovered in the experiments—e.g. a substantial evidence for a new particle (e.g. Higgs-boson)—, the physicists do not jump to the conclusion what exact particle or property they have discovered (e.g. Higgs-boson of the Standard Model or a Higgs-boson-like particle of Super Symmetry model), as the particle or property may be accounted for by competing models given the evidence. The inductive process leads them to converge on the discovery of a particle or a property of significance, but convergence on its exact nature as it is characterized by a specific model is a more arduous process. One could suggest that the quick and stable convergence we see in the laboratories is a result of the reasoning process more akin to deductive reasoning, or a low-level induction working with simpler data-sets and models, and that the truly inductive process never results in such quick convergence. This is certainly possible especially

because the reasoning in these cases concerns phenomenological models. Yet it could be that the inductive process that leads to the convergence on a particular model is of the same sort as the one leading to fast and stable convergence on the phenomena of significance. It is just that it takes longer for new experiments to update the physicists' beliefs and turn their higher-level dilemmas into a search for new phenomena of significance.

3.4 Conclusion

To sum up, the research in HEP follows inductive rules and patterns stemming from the baseline conservation principles. This inductive process, in turn, guarantees a broad and reliable convergence on the results. Based on the inductive convergence on the reliable results, the impact of the results can be measured by weighted citations and taken as representative. In this way, IA justifies the OA identification of optimal organizational structure. If IA cannot guarantee the reliability of the results, then we are not justified in applying OA. In this way, IA based on FLT streamlines OA, although it is possible that there are other internal justifications of this sort. The success of the FLT at reconstructing the inductive method in HEP should not be surprising: the method of data gathering and analysis in HEP is, generally speaking, along the lines of FLT induction. Even a superficial glance at the field, let alone a detailed analysis, suggests this. This is the case in some subfields of contemporary experimental biology as well.

4 Case 2: Phylogenetic Research

4.1 Machine Learning Techniques and Parsimony in Phylogeny

In biology, generally speaking, consensus on results is not fast and reliable. The time scale is much longer than in HEP, even if consensus on the results and their relevance eventually occurs. And it is often difficult to find a coherent set of inductive rules governing the research in biology. Yet along the lines of our previous analysis, we can find subfields of biology governed by inductive principles based on the FLT, wherein the pursuit passes the inductive test and gives the green light to operational analysis.

Phylogenetics, a subfield of evolutionary biology identifying trees of evolutionary relations between species (phylogeny), is particularly suitable for such analysis. Analogously to the conservation principle in physics, in phylogenetics, the usual baseline principle is the principle of parsimony. The principle states that all other things being equal, the best hypothesis concerning an evolutionary relationship is the one that requires the fewest evolutionary changes (Yang and Rannala 2012). Analogously to an inductive principle used as a baseline in FLT analysis, it makes the process of reaching a conclusion efficient, even though it does not guarantee its truth.

Over the years, the concept of 'fewest evolutionary changes' has been interpreted in various ways. In the beginning, researchers compared the set of properties of organisms. They gradually moved on to focusing on the common development of species. Finally, they started calculating similarities between sequences of genes. The principle for all three kinds of reconstruction remained the same, however, i.e., the closeness relation.

Now, in the case of HEP, as explained above, we rely on machine learning analysis of the inference procedures independent of the actual experimental process other than sheer experimental results. The machine learning application is an afterthought of sorts that

Table 2 Differences between sequences in phylogenetic analysis

	AAA	AAB	BBA
AAA		1	2
AAB	1		3
BBA	2	3	

produces the models distilling the inductive process in accordance with the FLT behind the pursuit, even though it could be subsequently utilized in it. But in phyolgenetics, based on the guiding principle of parsimony, biologists run much reduced models as the primary tool of analysis. So to assess the suitability for the OA, we need to assess the actual application of relevant methods and see to what extent they agree with the FLT-based inductive procedures.

Evolutionary relationships are established based on sequence similarities between genes, and the inductive principle ('fewest changes') suggests that closely related organisms share a higher degree of sequence similarity. To give an example, some amino-acids are more similar than others; therefore, not every difference in the protein sequence of genes indicates the same evolutionary distance. To account for this, matrices employing observed amino-acid changes between homologous proteins in large data sets have been designed to calculate expected exchange frequencies between genes with similar evolutionary distance. Numerical scores are assigned to differences in the nucleotide, or amino acid sequence, based on the frequencies of the differences. The greater the frequency (in large data sets) the smaller the number assigned. The calculation gives an optimal tree, i.e., the one with the smallest number of differences between branches.

For example, take three sequences with the same length, AAA, AAB, BBA. If we set the expected frequencies to 1 for all differences, the resulting scores are the following: AAA-AAB:1, AAA-BBA:2, and AAB-BBA:3 (Table 2).

To create an optimal tree, the algorithm searches for the minimal total score, i.e. the smallest sum. Thus, it will place AAB and AAA together, with BBA as an out-group, resulting in an overall score of 3, with the leaves on the resulting tree grouped as follows: (AAA-AAB)-BBA.

Note that, in general, this approach results in a reliable tree, depending on the adequacy of other assumptions. However, several obstacles can prevent researchers from finding the correct solution[20] (Yang et al. 2016). When sequencing approaches became cheaper, for example, whole genome sequences were suddenly available; the resulting tree depended on which sequencing method was selected. The problem of homogeneity continues today, albeit to a lesser extent. Now, protein sequences are mainly used to generate trees, as similar homologous proteins can be found in very high amounts. To establish their relationship, researchers use matrices of the frequency of amino acid exchanges. These matrices contain data on how frequently a specific exchange occurs in sequences with a specific similarity; for example, BLOSUM62 holds the observed frequencies for proteins with a similarity of 62% (Henikoff and Henikoff 1992). In addition, inserts and deletions are scored with a specific value. All these scores can be chosen by the researcher, and this affects the result (the tree), at least with respect to some details.

[20] Historically, researchers constructed trees solely based on the 19S RNA, because of the difficulties obtaining sequence information (Yang et al. 2016).

A further question is which exact data are relevant for the tree reconstruction. Do we compare conserved proteins or domains, and how do we weigh exchanges in these conserved positions in contrast to variable regions? All these decisions are based on the researchers' former experience and might therefore vary. Horizontal gene transfer further complicates analysis in some cases (Koonin 2016). Even though researchers tend to construct binary trees, horizontal transfer of DNA between different branches can occur. In such cases, the real tree is not binary but a net. And because scientists can only access information about the current specimen, they can only infer the sequence of the last common ancestor based on probabilities; they cannot know if they are correct. New information might force recalculation, leading to changes in the tree.

Nonetheless, the principle of parsimony in phylogeny is clearly an efficient method for generating adequate rules of evolutionary relationship.[21] The core tasks and analyses are results of a streamlined inductive-computing process around which the entire scientific process is organized, precisely the way FLT framework suggests. In other words, the reduced models based on parsimony are what the actual pursuit consists of, so the pursuit inherently satisfies the internal condition of the OA test. Whether it satisfies the external one (for an adequate use of citation metrics) is less clear: as the results of phylogenetic research are of a wider significance, citation counts will be spread across various fields, much more so than in the case of HEP results. Hence, we need to be able to identify and extract expert-based citations if we are to draw conclusions concerning the inductive efficacy of various elements of the network (teams, sub-teams, labs, individuals, etc.). This requires more research.

4.2 Applicability of Inductive Analysis in Other Areas of Biology

Phylogeny is one of myriad research topics in contemporary experimental biology. As different principles and approaches are applied in different subfields, the application of a hybrid of IA and OA across biology is a non-trivial task. There are various reasons why results in biology are, in general, not as quickly agreed upon and as reliable as they are in HEP. First, results that cannot be replicated are published in journals with high impact factors and get a high number of citations (Pusztai et al. 2013). Second, the majority of all retractions of the papers published by 2015 mention fraud or other kinds of scientific misconduct (Brainard and You 2018). Third, there may be a problem deciding what constitutes sufficient evidence for a hypothesis, especially if the hypothesis is non-parsimonious, i.e. when the hypothesis is not the simplest explanation of the phenomenon. Fourth, an expectancy bias appears in reports on the results. These negatively influence the replicability of biological experiments and slow down consensus (Goodman et al. 2016).

In modern phylogenetics, however, data are numerically expressed, and this makes the field suitable for machine learning analysis based on parsimony. In many other branches of experimental biology, such as cell biology, pictures are the main data. To analyze them, interpretation is crucial. But how these pictures are interpreted is heavily dependent on the prior knowledge and beliefs of the scientist. Another relevant issue is that experimental conditions in biology are not as clearly set as they are in experimental particle physics. Even though efforts are made to provide similar conditions, it is hard to do so when it comes to, for instance, the quality of soil, light, or

[21] This use accords with an account of parsimony in Kelly (2004, 2007).

humidity in plant biology. For example, unless the bulbs in plant growth chambers are simultaneously exchanged and equally used in different laboratories, the light quality will not be exactly the same, and this may affect the results. In particle physics, it is substantially easier to provide equal conditions, especially since the same experimental machines (accelerators and detectors) are often simply recombined to perform different experiments.

4.3 Non-parsimonious Results

To understand which evidence is sufficient for the acceptance of a hypothesis by the biological community, we need to consider the expected likelihood of the hypothesis. In the case of non-parsimonious results, acceptance is much slower than for parsimonious ones. For example, consider Koch's second postulate: all infectious diseases are caused by an organism. After showing that protease-resistant proteins, prions, cause Scrapie disease, Koch's second postulate was abandoned (Soto 2011). It took some time for the argument that a protein can cause an infectious disease and influence the folding of other proteins to be accepted. The first experiments were conducted in 1967, and the protein hypothesis was formulated. Acceptance was bolstered by the famous results of Prusiner (1982), but the scientific community was not persuaded until 1990 when mice were infected with the disease in a laboratory (Soto 2011).

The discovery of human papillomavirus as the main cause of cervical cancer took a similar path (zur Hausen 2009). It was already known that viruses could cause cancer, yet it was not accepted that a virus could be the main cause of a specific type of cancer. At the time, the disease was not considered infectious, thus a substantial number of argumentative steps was needed for establishing the correlation between the virus and cervical cancer. In the end, a uniform hypothesis that cancers cannot be caused by infectious diseases was defeated.

When it comes to hypotheses in line with the parsimony principle, the scientific community has fewer acceptance requirements. For instance, results that are in line with Koch's second postulate are accepted by the life science community quicker. Koch's proof that a bacillus causes anthrax required only two argumentative steps. In the first step, the presence of the microorganism in patients was established, while in the other step subjects were infected with the microorganism grown in pure culture.

Generally speaking, in the case of disease-causing agents, we can point to some general criteria, but we cannot find regular principles such as the conservation principle in physics. Yet as illustrated by previous cases, the acceptance of unexpected and/or non-parsimonious hypotheses takes time and requires many argumentative steps, so we cannot talk about concomitantly fast and reliable conclusions. Thus, in these cases, the relevant research line is not predicated on a baseline principle. It isn't that data aren't available (e.g. testing the Higgs boson hypothesis waited for three decades because the experimental apparatus was not available) but the community was always divided on the relevance of the existing data. The citation data might reflect this division, but we could not use them to decide which labs were efficient and which ones inefficient, because in this and other cases, those whose results are not cited but denigrated might emerge winners in the end, albeit after a long period of time.

5 Wider Inductive Convergence and Adequacy of Operational Analyses

An inductive analysis of a scientific pursuit of the sort we discuss here provides at least minimal assurance of the methodological coherence required for operational analysis to yield transparent methodological insights of the pursuit. Yet we need not limit our analysis to the FLT-based inductive account.

Generally speaking, philosophers and theoreticians of induction are selective when choosing cases to illustrate or assess their inductive models. This means that IA is not as open-ended as we may like to think; for example, each scientific pursuit may find a fitting inductive analysis, as several inductive accounts have been developed. In fact, the few, often identical, cases invoked in discussions of IA are simply drops in an ocean of cases and represent those displaying coherence of the pursuit based on at least one inductive model. This provides at least minimal assurance that a selected domain exhibits methodological coherence, as explicated by at least one inductive method. This is much more than operational analysts typically offer in the way of epistemically transparent use of their citation metrics—which is often nothing. Even checking for basic coherence of a pursuit requires a model. And checking a sophisticated pursuit like the one in HEP or research on phylogeny requires sophisticated inductive models and tests.

The hybrid of IA and OA may not be applicable to all pursuits. For instance, exploratory pursuits often do not reflect inductive coherence and are characterized by divergent results. Other research pursuits, as we have demonstrated, have no overarching parsimonious streamlining. The application of OA cannot be inductively justified in such cases, as the inductive process behind the pursuit is not effectively computable—there is no unifying principle or restrictive rules, nor are there computable models of the pursuit. Moreover, it may be problematic, from this standpoint, to apply an epistemically transparent OA across pursuits at all. It is not clear what inductive efficiency a citation metric could track in this case. In general, only very streamlined (inductively reduced) or mature pursuits pass the IA test. Perhaps the major challenge is to develop clear criteria for exploratory scientific pursuits and determine the inductive baseline in such cases, if there is one.

Inductive assessment introduces a substantial measure of transparency to operational analysis but, at the same time, puts substantial restrictions to it. Perhaps OA should not be applied prior to identifying methodological coherence of some sort within the domain of citation metrics application. This is a baseline constraint on OA that prevents spurious analysis and undesired side-effects stemming from lack of understanding of the operationally analysed pursuit: insofar as the IA of the pursuit is adequate, such effects are not likely to occur. Thus, OA will not suggest anything that will undermine or go against the methodology that made the pursuit successful in the first place.

The inherent inductive process behind the analysed domain guarantees the success of inductive analysis and, thus, ensures the transparency of the operational analysis. Besides a lack of desired inductive streamlining, however, other problems might occur when applying specific types of operational analysis. Not every type of operational analysis can reliably be applied on every data set. But various tests are available to assess how informative an operational analysis has been. For instance, data envelopment analysis, used to find efficiencies in multiple inputs and outputs, evaluates extreme points as efficient. To apply this type of operational analysis, we have to exclude the outliers from the data set. A sensitivity analysis can be conducted for this purpose (Ben-Gal 2005). Another case is the limitation of statistical methods, in particular, types I and II errors. Mayo and Spanos (2006) argue that when hypotheses that pass severe tests are used, these errors are minimized. It

is important to note that if a specific OA proves inadequate for a given data set, a different one might apply, providing informative results.

Put otherwise, the approach we suggest does not disqualify other approaches but sets a standard against which they can be developed. For instance, the emerging use of simulations of scientific networks that strive towards the empirical calibration of the models can introduce a level of inductive coherence in the analysis, improving its justification and transparency. Moreover, a possible convergence of different inductive analyses on the reliability of a specific research pursuit might argue in favour of a justified application of the operational analysis of a scientific network in which the pursuit is embedded.

Acknowledgements This work was presented at the conference "Formal Methods of Scientific Inquiry" held at the Ruhr-University, Bochum in 2017. We are greatful to the participants of the conference, audience at the Center for Formal Epistemology at the Carnegie Mellon University, Kevin T. Kelly, Oliver Schulte, Konstantine (Casey) Genin, anonymous referees and guest editors of the special issue for a number of comments and constructive criticisms. This work was supported by grant #179041 of the Ministry of Education, Science, and Technological Development of the Republic of Serbia.

References

Alexander, J. M., Himmelreich, J., & Thompson, C. (2015). Epistemic landscapes, optimal search, and the division of cognitive labor. *Philosophy of Science, 82*(3), 424–453.

Allen, L., Brand, A., Scott, J., Altman, M., & Hlava, M. (2014). Credit where credit is due. *Nature, 508*(7496), 312–313.

Baltag, A., Gierasimczuk, N., Smets, S. (2015). On the solvability of inductive problems: A study in epistemic topology. In R. Ramanumam (Ed.), *Proceedings of the 15th conference on theoretical aspects of rationality and knowledge* (pp. 65–74), TARK 2015.

Ben-Gal, I. (2005). Outlier detection. In O. Maimon & L. Rockach (Eds.), *Data mining and knowledge discovery handbook: A complete guide for practitioners and researchers* (pp. 131–146). Dordrecht/Berlin: Kluwer/Springer.

Bonaccorsi, A., & Daraio, C. (2005). Exploring size and agglomeration effects on public research productivity. *Scientometrics, 63*(1), 87–120.

Borg, A. M., Frey, D., Šešelja, D., & Straßer, C. (2017). An Argumentative agent-based model of scientific inquiry. In S. Benferhat, K. Tabia, & C. Straßer (Eds.), *Advances in artificial intelligence: From theory to practice. IEA/AIE 2017. Lecture notes in computer science*, Vol. 10350 (pp. 507–510). Cham: Springer.

Bornmann, L. (2017). Measuring impact in research evaluations: A thorough discussion of methods for, effects of, and problems with impact measurements. *Higher Education, 73*(5), 775–787.

Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45–80.

Brainard, J., & You, J. (2018). What a massive database of retracted papers reveals about science publishing's 'death penalty'. *Science*. https://doi.org/10.1126/science.aav8384.

Braun, T. (2010). How to improve the use of metrics. *Nature, 465,* 870–872.

Campanario, J. M. (1993). Consolation for the scientist: Sometimes it is hard to publish papers that are later highly-cited. *Social Studies of Science, 23*(2), 342–362.

Carillo, M. R., Papagni, E., & Sapio, A. (2013). Do collaborations enhance the high-quality output of scientific institutions? Evidence from the Italian Research Assessment Exercise. *The Journal of Socio-Economics, 47,* 25–36.

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research, 3*(Nov), 507–554.

Contopoulos-Ioannidis, D. G., Alexiou, G. A., Gouvias, T. C., & Ioannidis, J. P. A. (2008). Life cycle of translational research for medical interventions. *Science, 321*(5894), 1298–1299.

Corley, E. A., Boardman, P. C., & Bozeman, B. (2006). Design and the management of multi-institutional research collaborations: Theoretical implications from two case studies. *Research Policy, 35*(7), 975–993.

Darriulat, P. (2004). The discovery of *W* & *Z*, a personal recollection. *European Physical Journal C, 34*(1), 33–40.

Dissertori, G., Knowles, I. G., & Schmelling, M. (2003). *Quantum chromodynamics: High energy experimetns and theory*. Oxford: Clarendon Press.
Genin, K., & Kelly, K. T. (2015). Theory choice, theory change, and inductive truth-conduciveness. In R. Ramanumam (Ed.), *Proceedings of the 15th conference on theoretical aspects of rationality and knowledge* (pp. 111–119), TARK 2015.
Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine, 8*(341), 341ps12.
Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, *89*(22), 10915–10919.
Kelly, K. T. (2004). Justification as truth-finding efficiency: How Ockham's razor works. *Minds and Machines, 14*(4), 485–505.
Kelly, K. T. (2007). A new solution to the puzzle of simplicity. *Philosophy of Science, 74*(5), 561–573.
Kelly, K. T., Genin, K., & Lin, H. (2016). Realism, rhetoric, and reliability. *Synthese, 193*(4), 1191–1223.
Kelly, K. T., Schulte, O., & Juhl, C. (1997). Learning theory and the philosophy of science. *Philosophy of Science, 64*(2), 245–267.
Kitcher, P. (1990). The division of cognitive labor. *The Journal of Philosophy, 87*(1), 5–22.
Kocabas, S. (1991). Conflict resolution as discovery in particle physics. *Machine Learning, 6*(3), 277–309.
Koonin, E. (2016). Horizontal gene transfer: essentiality and evolvability in prokaryotes, and roles in evolutionary transitions. *F1000Research, 5,* 1805.
MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for Information Science, 40*(5), 342–349.
Martin, B. R., & Irvine, J. (1984a). CERN: Past performance and future prospects: I. CERN's position in world high-energy physics. *Research Policy, 13*(4), 183–210.
Martin, B. R., & Irvine, J. (1984b). CERN: past performance and future prospects: III. CERN and the future of world high-energy physics. *Research Policy, 13*(4), 311–342.
Maruyama, K., Shimizu, H., & Nirei, M. (2015). Management of science, serendipity, and research performance: Evidence from scientists' survey in the US and Japan. *Research Policy, 44*(4), 862–873.
Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *The British Journal for the Philosophy of Science, 57*(2), 323–357.
Peltonen, T. (2016). *Organization theory: Critical and philosophical engagements*. Bingley, UK: Emerald Group Publishing.
Perović, S., Radovanović, S., Sikimić, V., & Berber, A. (2016). Optimal research team composition: Data envelopment analysis of Fermilab experiments. *Scientometrics, 108*(1), 83–111.
Prusiner, S. (1982). Novel proteinaceous infectious particles cause scrapie. *Science, 216*(4542), 136–144.
Pusztai, L., Hatzis, C., & Andre, F. (2013). Reproducibility of research and preclinical validation: Problems and solutions. *Nature Reviews Clinical Oncology, 10,* 720–724.
Rosenstock, S., O'Connor, C., & Bruner, J. (2017). In epistemic networks, is less really more? *Philosophy of Science*, *84*(2), 234–252.
Schulte, O. (2000). Inferring conservation laws in particle physics: A case study in the problem of induction. *The British Journal for the Philosophy of Science, 51*(4), 771–806.
Schulte, O. (2018). Causal learning with Occam's razor. *Studia Logica*. https://doi.org/10.1007/s11225-018-9829-1.
Schulte, O., & Drew, M. S. (2010). Discovery of conservation laws via matrix search. In O. Schulte & M. S. Drew (Eds.), *Discovery science. DS 2010. Lecture notes in computer science*, Vol. 6332 (pp. 236–250). Berlin/Heidelberg: Springer.
Soto, C. (2011). Prion hypothesis: The end of the controversy? *Trends in Biochemical Sciences, 36*(3), 151–158.
Thagard, P., Holyoak, K. J., Nelson, G., & Gochfeld, D. (1990). Analog retrieval by constraint satisfaction. *Artificial Intelligence, 46*(3), 259–310.
Ting, Samuel C. C. (1977). The discovery of the J particle: A personal recollection. *Reviews of Modern Physics, 49*(2), 235–249.
Valdés-Pérez, R. E., & Żytkow, J. M. (1996). A new theorem in particle physics enabled by machine discovery. *Artificial Intelligence, 82*(1–2), 331–339.
van der Wal, R., Fischer, A., Marquiss, M., Redpath, S., & Wanless, S. (2009). Is bigger necessarily better for environmental research? *Scientometrics, 78*(2), 317–322.
Van Noorden, R. (2014). Transparency promised for vilified impact factor. *Nature News, 29*, 2014.
Voinnet, O., Rivas, S., Mestre, P., & Baulcombe, D. (2003). Retracted: An enhanced transient expression system in plants based on suppression of gene silencing by the p19 protein of tomato bushy stunt virus. *The Plant Journal, 33*(5), 949–956.

Warner, J. (2000). A critical review of the application of citation studies to the Research Assessment Exercises. *Journal of Information Science, 26*(6), 453–459.
Weisberg, M., & Muldoon, R. (2009). Epistemic landscapes and the division of cognitive labor. *Philosophy of Science, 76*(2), 225–252.
Yang, Z., & Rannala, B. (2012). Molecular phylogenetics: Principles and practice. *Nature Reviews Genetics, 13,* 303–314.
Yang, B., Wang, Y., & Qian, P. Y. (2016). Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics, 17*(1), Article number 135.
Zollman, K. J. (2010). The epistemic benefit of transient diversity. *Erkenntnis, 72*(1), 17–35.
Zur Hausen, H. (2009). The search for infectious causes of human cancers: Where and why. *Virology, 392*(1), 1–10.