Non-statistical components of research rigor in biology

Pamela Reinagel

Abstract

In recent years discussions of scientific reproducibility have proliferated, propelled by the narrative that current scientific practice is quite poor and in need of radical reform. This conclusion was originally based on examination of practices in psychology and other social sciences, but more recently the conclusions – and proposed reforms – have been extended to biology. These discussions tend to equate the scientific method with falsifying statistical null hypotheses, and therefore focus on whether that procedure is being correctly followed or interpreted. As a practicing biologist, I concur with the assessment that null hypothesis significance tests (NHST) are often performed and interpreted incorrectly in biology today, leading to widespread reporting of invalid p values. However, I think those same studies are often rigorous and their conclusions well-supported. I attribute this to reliance on other, mainly nonstatistical methods to avoid errors and strengthen inferences. If this is true, reforms aimed at enforcing the requirements of statistical significance testing are not needed, and could do substantial harm. Rather, the non-statistical methods important for research rigor need to be better articulated, defended, and communicated.

Introduction

Current rhetoric regarding the (un)reproducibility of research tends to equate rigor with formal statistical methodology, and in particular, tends to equate null hypothesis significance testing (NHST) with “the scientific method”. “Such is the grip of formal methods of statistical inference—that is, frequentist methods for generalizing from sample to population in numerative studies—in the drawing of scientific inferences that the two are routinely deemed equivalent in the social, management, and biomedical sciences.”(Hubbard, Haig, & Parsa, 2019). For example, the National Institute of Neurological Diseases and Stroke convened a panel that published “a core set of reporting standards for rigorous study design” as “universal” guidelines. Their recommendations emphasized several aspects of prospective study design – a requirement that is specific to NHST (Landis et al., 2012). A role for exploratory research is acknowledged, but only for generating hypotheses: “Potential discoveries arising from the exploratory phase of the research should be supported by follow-up, hypothesis-testing experiments.” (Landis et al., 2012).

Practices that undermine control of the false positive rate in an NHST context are widely dubbed “Questionable Research Practices” (QRPs) without qualification, even though these same practices may be harmless or beneficial to rigor in other inferential frameworks. Although authors often briefly acknowledge that these behaviors are not QRPs in the context of non-confirmatory research, the discussion (within the papers, in the literature, and in the community) tend to interpret findings about QRPs assuming an NHST context. For example, Lakens presumes a frequentist severe testing framework when he uses normative words like “admission” to refer to self-reported engagement in these research practices (Lakens, Mesquida, Rasti, & Ditroilo, 2024). This language ignores the possibility that the survey respondents may have been using other approaches that do not purport to control type I or type II error rates – even if they use terms like “test”, “hypothesis”, and “claim”.

Calls for the widespread adoption of preregistration (Nosek, Ebersole, DeHaven, & Mellor, 2018; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012) imply that most research is designed to test statistical null hypotheses, or should be. For example, Andrade states “In an ideal world, all research protocols should conform to these [clinical trial] norms and be made available in publicly accessible online registries before or by the time of study

commencement. This would help reduce the temptation to indulge in QRPs and make QRPs possible to identify."(Andrade, 2021). In summary, the literature acknowledges the existence of non-confirmatory research, minimizes its importance for establishing reliable or credible claims. This stance dismisses the possibility that research can establish valid, strongly supported conclusions by other means, including entirely non-statistical arguments (Hubbard et al., 2019).

As a counter-example to the prevailing assumption that NHST is required to establish secure scientific knowledge, many fields of experimental biology enjoyed extraordinary productivity in the 20th century (Morange & Cobb, 2020), in spite of a conspicuous absence of NHST. For example, there are no p values in seminal works of molecular biology, such as Hershey and Chase (Hershey & Chase, 1952), Meselson and Stahl (Meselson & Stahl, 1958), Nirenberg and Matthaei (Nirenberg & Matthaei, 1961), or Jacob and Monod (Jacob & Monod, 1961). Even the overtly statistical reasoning in Luria and Delbruck (Luria & Delbruck, 1943) refrained from using statistical null hypothesis tests or making claims of statistical significance. This raises an interesting question: what then accounts for the reliability of this research, if not formal statistical practice?

More recently, p values have reported by biologists perfunctorily, often in spite of extensive researcher flexibility in the generation of the data. While this is clearly misleading and inappropriate, it does not follow that the conclusions of those studies are poorly supported or have a high likelihood of being spurious and irreproducible. It only means that NHST is not a valid (or at least not a strong) basis for the conclusions of those studies. It is possible (and I think likely) that the authors and their readers reached their conclusions on the basis of other (typically non-statistical) reasoning also presented in the reports, which could be very rigorous.

In this paper I will explore non-statistical research practices that may have contributed to research rigor in experimental biology, drawing on decades of experience working in molecular biology (including biochemistry and classical genetics) and systems neuroscience (including neurophysiology and quantitative behavior). These fields share several characteristics that differentiate them from the fields in social science in which the "reproducibility crisis" was first identified (Simmons, Nelson, & Simonsohn, 2011). Here I will discuss five of these:

(1) The goal of the research is to develop mechanistic causal models that coherently account for a set of related biological observations

(2) The candidate mechanistic explanations are strongly informed by theory and highly constrained by existing knowledge, such as laws of physics and chemistry.

(3) The specific cases under study are of interest qua examples general phenomena, and therefore chosen for being advantageous for inference

(4) The research has an iterative or cumulative quality that entails implicit replications

(5) There is an intentional pursuit of multiple independent conceptual and experimental approaches (triangulation).

Whether statistical methods are necessary or sufficient for scientific rigor depends on the kinds of inferences being drawn, the kinds of uncertainty being faced, the material nature of the object of study, the extent of constraining knowledge and theory, the quality of tools of measurement and intervention, the predominance of alternative cross-checking methods, and the presence or absence of epistemic virtues in the research community. It is a worthy goal to assess and improve research rigor in biomedical science. To achieve this, we need a better account of the epistemic basis for the inductive inferences being made within each subdiscipline considered on its own terms (Devezer, Nardin, Baumgaertner, & Buzbas, 2019; Dupré, 1993; Norton, 2021; Peterson, 2021). In some biological research, statistical methodology is essential to inference, and improvements in statistical practice are sorely needed. But in other kinds of experimental biology, inductive inferences are well-warranted by a complex convergence of

many types of quantitative and qualitative evidence and reasoning, which may or may not include formal statistical methods, much less null hypothesis significance testing.

Clarifying these issues serves several purposes. First, if there are non-statistical methodologies that are valid and productive routes to securing knowledge, it is important that these be recognized and accepted, so that the past and potential contributions of such research are not devalued as an unintended consequence of current efforts to improve statistical practice. Second, to the extent that alternative methodologies and practices are implicit among researchers, making the methods and their justifications more explicit will facilitate teaching these methods, communicating them transparently, and assessing their use in a given case. Third, making non-statistical components of research rigor more explicit may help prevent them from being applied in circumstances where they are not appropriate.

My sources and scope

This work is based primarily on my first-hand experience working in and leading basic research biology laboratories. My claims about common practices, beliefs and attitudes are based on many kinds of interactions, notably: working in labs alongside others; weekly lab meetings in which lab members present their work-in-progress for feedback, and mentors providing guidance and conveying standards and expectations; graduate-level reading courses and journal clubs, in which trainees and researchers debate the strength of evidence for claims in recently published research papers; seminars, conference talks and poster sessions, where colleagues present and discuss their latest findings; thesis committee meetings in which PhD students present their plans and ongoing progress for feedback and guidance, and ultimately present their final results; the written arguments in publications through which authors lay out the reasons for their conclusions; and countless one-on-one conversations with colleagues. I enumerate these because I am not sure whether all these mechanisms of training or of critical discussion exist in all research fields. No one of these is of supreme importance; rather it is important that numerous practices of this kind are woven together into a tapestry of good scientific practice.

To cross-check my impression of prevailing views in molecular biology, however, I contacted some former mentors and asked them: “To what extent do you feel that statistics (specifically the framework of posing and falsifying null hypotheses with statistical tests and p-values) has been pivotal to the progress of your own research, or to biology in the past century in general?” Here are some of the answers I received (emphasis added):

> “I declined the invitation [to a statistics workshop] because [this] is something I know nothing about and because I dislike statistics. **Falsifying a series of null hypotheses using statistical tests is something I would never stoop to!** […] If you need statistics, you’ve designed a bad experiment. [For example] DNA replication is either semi-conservative or it isn’t.[1]”
>
> Matt Meselson, molecular biology

> “In molecular biology and genetics with model organisms (yeast, flies, mice), many experiments can be completed in less than a month, sometimes less than a week. Six months of work can complete many experiments which (at least) are

[1] Semi-conservative replication refers to the fact that when DNA replicates each new DNA helix contains one pre-existing strand and one newly-synthesized strand.The alternative would have been that after replication one molecule would contain both original strands and the other one contain two new strands. Meselson and Stahl designed an elegant experiment that made extremely specific, qualitatively different predictions in the two cases (Meselson & Stahl, 1958; Olby, 2002).

partial replications of each other. **Each individual experiment** may be focused on characterizing some aspect of the phenomenon, but it **also serves as a replication** of the early observations[2]. Individual experiments may be interpreted qualitatively from banding patterns in electropherograms, without any statistical analysis. In addition, some direct replications may be done, perhaps on a larger scale, to get a more precise estimate of a quantitative parameter. But a mean and standard error is the limit of statistical analysis.

I remember overhearing this advice at a meeting once decades ago, 'Don't worry about statistics; just make sure your experiment is repeatable.' That advice depended on the unspoken assumption that the experiment could be repeated, perhaps several times, in a reasonable amount of time. Similar but less helpful advice was, 'If your experiment needs statistics, you need a better experiment.'

As you well know, [however], many experiments need months or years of data collection before any analysis is possible. People doing those do not have the luxury of replication before publication."

Kenneth Manly, virology/mouse genetics

"It seems to me that the usefulness (or absence thereof) of statistics is entirely linked to the problem studied. If you identify a protein essential for a reaction, you put the protein in the test-tube, the reaction occurs, you do not put it in, the reaction does not occur or just very weakly, probably because other components of the reaction contain traces of the protein in question. **Journals ask these days for p-values and the like, so you can add what they ask for, but the fact is, it is totally useless**. In another case, you cannot obtain a clear answer unless you repeat the experiments many, many times. There statistics can guide you into believing your hypothesis is correct, or not."

Nouria Hernandez, gene regulation

"Already in the 1970's, I was using standard errors in my very first publications before I did my PhD work. It was to establish that what we saw with our eyes (i.e., what we knew was right) had statistical support. At the time, **we did not use statistics to show that something we could not see with our eyes was true**. Statistics, including null hypotheses, has a longstanding role in biological research. […] The tradition that for something to be 'significant' it has to have a p value of less than 0.05 and if it has a p value of less than 0.05 it is significant is (or at least was) among the most hideously abused concepts in science."

Winship Herr, gene regulation

"I suspect the answer to your general question will be strongly influenced by which period we are talking about. I may be wrong, but as I recall, classical molecular biology, with its amazing discoveries, required / involved little or no statistical analyses. That is perhaps because **each little finding made one or another testable prediction**[2], and often those predictions could be tested quickly. In a sense, statistical analyses involve abandoning rationality:[3] when a drug is tested, and you don't have a strong reason the think x or y, you rely on statistics to decide, or at least point in one direction or another. There are two

[2] This remark is in alignment with Guttinger's observations on the role of micro-replications.

[3] i.e., abandoning any attempt to reason theoretically, mechanistically, or on the basis of hard constraints governing the material components of the experiment (my interpretation)

related criteria for truth in a (molecular) biological result: coherence, and the results of the predicted experiments made possible by the result.

Nowadays the scene is quite different - fancy statistics (machine learning, etc.) are involved everywhere."

Mark Ptashne, gene regulation

Here I will reflect on characteristics I have noticed are common to both fields of biology in which I have worked (molecular biology and systems neuroscience), and possibly different from some of the social sciences that have reported a severe replicability crisis – characteristics that might account for differences in the robustness of findings. These comments do not apply to all kinds of biology research, however. For example, practices and attitudes seem to be different in subfields of biology where controlled experiments are generally impossible, such as epidemiology, evolution, or field ecology (Amrhein, Trafinnow, & Greenland, 2019; Fidler et al., 2017; Fraser, Parker, Nakagawa, Barnett, & Fidler, 2018; Popovic et al., 2024).

I am not sure to what broader class my observations might generalize – perhaps experimental basic research in model organisms. My description of practices is not meant to cover all researchers and all research products within those subdisciplines. Rather, they characterize my impressions of the best researchers in these fields. Naturally my observation set is biased. Nevertheless, within these fields I perceive certain attitudes, standards and practices to be typical, and will distinguish these from those I would consider idiosyncratic. For simplicity I will restrict myself to one concrete example from molecular biology, but the same points could equally well be illustrated with examples from systems neuroscience.

Characteristics of research practice

An example: gene regulation

To illustrate my points, I will use an example research question from the history of molecular biology. How is it that your liver cells look like liver cells and produce liver-specific proteins, while your brain cells – which have the identical genes – developed into brain cells and make brain proteins? This paradox was posed in the early 20th century in the context of embryonic development (Burian, 2005). But the question extends to single celled and even non-cellular organisms. For example, the single-celled yeast *S. cerevisiae* can make the enzymes to digest the sugar galactose – but it only does so if galactose is present, *and* the preferred sugar glucose is absent – two different phenotypes with the exact same genes. Similarly, the bacterium *E. coli* only makes the enzymes to digest the sugar lactose when lactose is present and glucose is absent.

Even bacteriophage – viruses that infect bacteria – can live two radically different lifestyles. A bacteriophage virus co-opts a bacterial host's cellular machinery to copy its genome, manufacture viral proteins, and assemble new viruses until the cell explodes ("lysis"). But some bacteriophage, such as λ, can also enter a cell and go dormant. Its genes are passively passed along to the bacterial progeny for many generations without making any viruses – even though all the genes for making viruses are present. These bacteria are called "lysogens" because at a later time, the dormant virus can come out of hiding, make viruses and lyse the cell (Bertani, 2004; Lwoff, 1953).

All of these are examples of a general phenomenon: gene regulation. Genes are turned on and off, or used vs. not used, in different contexts, and this occurs in an organized way that makes functional sense. After it was appreciated that genes were composed of DNA and gene expression required transcription, this paradox was posed at the molecular level of organization: what are the molecular mechanisms that cause genes to be transcribed or not transcribed in different contexts? This example will serve to concretize some of my general observations.

1. The inferential goal is to develop general mechanistic causal models

In much of biology research, the overarching goal of research is to develop mechanistic causal models that coherently account for a set of related biological observations (T. Baetu & Carl, 2015; T. M. Baetu, 2019; Bechtel, 2006; Machamer, Darden, & Craver, 2000). This is in contrast to research that seeks merely to describe or quantify statistical associations (without inferring causation), or to describe and quantify effects of manipulations (inferring causation, but without any mechanistic model), or to determine with some degree of certainty or estimate with some accuracy a specific association or effect of practical interest (as an end goal, rather than as a means to discovering a broader principle or theory).

In the field of gene regulation, for example, a "model" is not a mathematical abstraction like "samples drawn independently from a population with a normal distribution". A "model" is a drawing representing physical objects and their relevant physical properties, interactions and actions (Figure 1). Such a drawing has specific implications for what outcomes should be observed in a range of different experiments. For example, every place two blobs touch, it is implied that those objects are in physical contact. It should be possible to experimentally measure an interaction between those physical components using any of a number of different techniques; and preventing the interaction should have a predictable consequence for function. The epistemic role of such exploratory modeling has been discussed by Gelfert (Gelfert, 2016, 2019) and others (Knuuttila & Loettgers, 2021). Several hypothetical models may be suggested and experiments designed to differentiate among them (Platt, 1964).

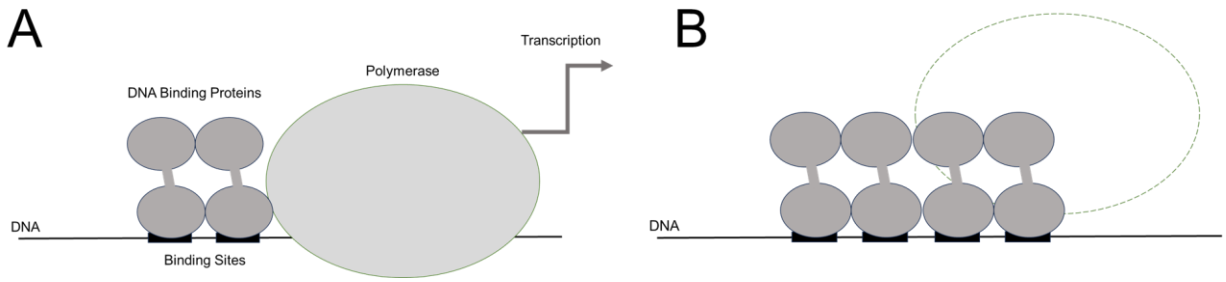

Figure 1. Mechanistic models of gene regulation. Shapes in these diagrams are intended to represent specific physical objects (such as protein or DNA). **A.** When two shapes are shown to touch, this indicates that these objects physically touch or bind to one another through specific attractive interactions between particular surfaces. Thus, this picture implies that the DNA binding proteins bind to one another, bind to the specific binding sites on DNA, and touch the polymerase enzyme, and the polymerase itself binds to DNA, leading to transcription. The fact that the DNA binding protein is drawn as a barbell instead of a single blob also has an implication: that the protein has two domains that are somewhat separable physically and functionally. **B.** Laws of physics are implied, such as that two objects cannot occupy the same location in space at the same time. This diagram implies that if DNA binding proteins bind to the sites on the right, this would occlude (prevent) polymerase from binding and thus prevent transcription.

While it is true that the data of each such experiment could be used to predict what would happen if the same experiment were to be repeated, that is not the end goal. The important inference is whether the outcome is consistent or inconsistent with a mechanistic model under consideration, in comparison with other models and other possible experimental outcomes. Of course, the experiments would not be informative if they were not repeatable in general. But the end goal is to accrue some degree of evidence in favor of or against one mechanistic model or another, rather than to establish an effect per se. Therefore, it is often more informative to test a large number of *different predictions* of the same model, than to invest the same experimental resources in a severe test of any one prediction[4].

[4] This point will be elaborated in the section on triangulation.

Each individual mechanistic model makes predictions that are very specific, and many of them are tested in great detail. But in basic research, these models are only a means to an end. The **specific models are interesting primarily qua examples of a general phenomenon**. The ultimate goal of having several such detailed examples is the hope that we will be able to recognize commonalities and eventually identify broad conceptual principles that apply to many particular cases. One reason for this hope is the expectation that many particular cases are related by descent from common evolutionary predecessors. Another reason is that many different examples will turn out to have common elements when viewed at the molecular level of detail, and those constituent elements will behave similarly wherever they are found.

For the problem illustrated, an example of such a general principle is “steric hindrance”: two molecules cannot be in the same volume of space at the same time. Therefore, one way a regulatory protein can prevent a gene from being read simply by physically getting in the way of the reading machine (Figure 1B). Firmly establishing this mechanism in a few particular cases was valuable, mainly because this resulted in formulation of a clear concept, and the invention of experimental designs that can demonstrate this mechanism and rule out alternatives. This concept and these experimental designs could then be applied to other cases.

A more subtle general principle is “cooperative binding”. That would require more unpacking than is possible here, but see (Ptashne, 2004) for an explicit discussion of how that principle, elucidated in the context of bacteriophage lambda’s lysis/lysogeny switch, was directly applicable to the molecular mechanisms by which a homologous protein achieves an analogous function in Drosophila embryonic development.

Although generalization is a goal, this generalization is at an abstract level. Every new example is expected to be unique – we don’t expect these models to predict the outcome of any particular experiment on any other gene in any other species. The outcome of successful research of the type I have described is the development of new frameworks, conceptual tools and experimental paradigms that allow us to rapidly work out the particulars of new examples (Schickore, 2016). It is only in subsequent applied research that estimating the sign, size, and reliability of some specific effect is of paramount concern.

2. The scientific hypotheses are highly constrained

Another characteristic of these mechanistic explanations is that they are highly constrained by existing knowledge, e.g., laws of physics and chemistry (Ross, 2023). In fields where models are highly abstract, nearly anything can be postulated because there are essentially no constraints (Meehl, 1978; Simmons et al., 2011). But the sort of research I am describing deals with a very limited space of possibilities. Models must be constructed out of known components (such as DNA and proteins) that have well-known compositions and physical properties. One cannot put two objects in the same place at the same time, or postulate an attractive force between two surfaces of like charge, or presume kinetics inconsistent with the limits of diffusion, and so forth. Extraordinary justification is needed to postulate a new type of entity or force. The discovery of the structure of DNA is a case in point: the challenge was not how to distinguish among multiple possible models, but rather in coming up with even one model that was consistent with all the constraints (Olby, 1994; Schindler, 2008).

There are other, softer constraints as well. For example, models are expected to make sense functionally (i.e., be reconcilable with natural selection), and models that are precedented by or analogous to other known biological mechanisms are considered more plausible. Thus, constructing a scientific hypothesis entails integrating vast amounts of background knowledge with direct observations and reasoning (Lennox, 2020).

These constraints on the space of admissible hypotheses have two implications for the reliability of statistical inferences. First, hypothesizing after results are known, or “HARKing”, might be less dangerous, or more valuable, than in research fields that lack strong constraints

on their hypotheses (Murphy & Aguinis, 2019). This is because so few hypotheses are tenable. Hypotheses that are constructed post-hoc to fit the data can be immediately excluded if they violate constraints, or can be immediately supported if they are consistent with many known constraints. Second, from a Bayesian perspective, the statistical hypotheses that are tested by experiments may have a higher prior probability (because they are already based on considerable background knowledge). Therefore, the posterior probability of a positive result being real will be higher than for a field that tests hypotheses with little or no supporting evidence and therefore lower prior probability[5].

3. Use of example cases chosen for inferential advantages

Biologists are often heard to say "if you need statistics, you've designed a bad experiment" (see quotes above). One unflattering interpretation is that such researchers have limited their scope of interest to only those phenomena that happen to be very large and reliable (leaving much of nature out of reach). But this is missing a key point. Because the goal of their research is to elucidate *general* principles, basic researchers have the luxury of seeking out *example cases* of the general phenomenon that *happen to be* particularly amenable to study.

Simplicity of examples

For example, early work on gene regulation focused on genes in bacteriophage and bacteria, specifically because these organisms are far simpler than, say, a human cell. They have far fewer physical components (for example they don't have chromatin, or a cell nucleus). They can also be grown very rapidly in small spaces, which makes it easier to do large N experiments and to replicate experiments many times.

Of course, this entails a risk that the chosen example case might not prove to generalize to cases of greater practical interest. This is mitigated by the fact that individual researchers select different example cases, and make different judgement calls on how broadly they define the phenomenon of interest, and thus how simplified a case counts as an example. For the problem of gene regulation, it did turn out that principles learned in bacteriophage and bacteria generalized remarkably directly to multicellular eukaryotes (Ptashne, 2004). Nevertheless, many additional aspects of gene regulation in eukaryotes depend on additional mechanisms that were not present in prokaryotes[6].

There could also be a risk of overgeneralization: some might fallaciously assume that all facts established for simpler example cases will necessarily and literally be true of other more complex cases, as the quip "What is true for *E. coli* is true for the elephant" (Monod & Jacob, 1961) is sometimes interpreted[7]. However, I think this is a straw man. In my experience, biologists understand that elephants are not *E. coli*. They think, however, that the common ancestor of *E. coli* and elephants already possessed highly evolved molecular mechanisms for many basic cellular and sub-cellular biological functions (inheritance, gene expression, metabolism, etc.), from which both *E. coli*'s and elephants' current mechanisms were historically

[5] Note that an effect being real does not guarantee it will be reproduced in a given replication attempt; that depends also on how the experiment is designed, and on chance. Nevertheless, for any reasonably good experimental design, real results are far more likely to be replicable than fluke findings.

[6] The choice of bacteriophage as the simplest example of some biological phenomena historically paid off (Cold Spring Harbor Laboratory of Quantitative Biology., Cairns, & Delbrück, 1966; Salmond & Fineran, 2015). But, importantly, simplest examples at other levels of complexity have been studied in parallel (e.g., *Escherichia coli* as a simple prokaryote; *Saccharomyces cerevisiae* as a simple, single-celled eukaryote; *Drosophila melanogaster* and later *Caenorhabditis elegans* as simple multicellular animals, and so forth).

[7] I think it very unlikely Jacob and Monod advocated the extreme view some attribute to this statement, and suspect they were using poetic license to emphasize how much more had been found to generalize than might have been expected. But ascertaining how this aphorism was understood at that time is beyond the scope of this essay.

derived. The most basic and fundamental functions of cells are *a priori* likely to (and in fact often do) persist in largely the same form in both lineages – as Jacob and Monod were in a position to know at that time (e.g., DNA is the molecule carrying genetic information, the same amino acids are used to comprise proteins). Nevertheless, generalization is always a conjecture until tested, as divergence is always possible; and more complex systems very often layer additional mechanisms on top of intact pre-existing simpler ones. I believe this is generally recognized.

Large effect sizes

Researchers can also choose example cases in which the effect sizes of interest are so large that the fluctuations due to noise or individual variation are too small to influence the outcomes of experiments. In the case of gene regulation, researchers focused first on genes in model organisms that are turned "on" or "off" in an almost deterministic or binary way (hundred- or thousand-fold effects), rather than the many other genes whose level of expression is modulated by environmental conditions by only a few percent. This strategic choice enabled many of the ground-breaking experiments to be performed with entirely qualitative assays: either a bacterium lyses, or survives bacteriophage infection; either a bacterial colony turns blue, or it doesn't; either a yeast grows on a specific growth medium, or it doesn't; either a radioactive band on a gel is present, or it is absent (see quotes above). Such experiments therefore did not require quantitative measurements, much less statistical methodology, to produce reliable results.

I don't think this strategy entails a strong assumption that the more subtle examples will work by the same mechanisms. Many molecular-level mechanisms can exist to a greater or lesser degree, for example binding affinity falls on a continuum. Therefore, mechanisms observed in cases with large effect sizes, when present to a lesser degree, could account for smaller effects in other examples. But this is understood to be a gamble, not a guarantee. Nevertheless, even if the more difficult-to-study cases work by entirely distinct mechanisms (as some do), the expectation is that solving the simpler cases first and identifying where they do generalize, will put us in a stronger position to work out those additional mechanisms.

Some argue that while formal statistical methods may not be *necessary* when effect sizes are very large, they are *possible*, so what is the excuse for not using them? I have two answers to this. First, using a formal statistical method just so you can state that some blatantly clear difference is "significant $p<10^{-12}$" does not make the inference any more rigorous, because the *limiting factor* for the reliability of the inference is not stochastic uncertainty, but some other kind of uncertainty (e.g., your reagent tube might be mislabeled; or the band you are seeing on the gel might not be caused by the mechanism you are assuming)[8]. If a result has a 1% or 10% chance of being wrong for some non-statistical reason, it does not matter whether the chance of a false positive due to sampling is 10^{-3} or 10^{-12}. Second, all rigorous methods come at a cost, and it is proper to weigh the cost against the benefit. Null hypothesis significance testing places considerable constraints on experimental practice. If there is no benefit, there is no justification for bearing the cost.

Often it is evident that stochastic variability is negligible, as routinely verified by running some or all of the test conditions in duplicate in every experiment, and visually confirming that results of repeated conditions are indistinguishable. When variability is observed and cannot be controlled by more careful procedures, researchers may be more inclined to look for another example case in which the effect of interest is larger or more consistent (as discussed above), or look for a different experimental approach that will test their mechanistic model more unequivocally (see Triangulation, below), rather than rely on statistical methodology to discriminate signals that are small compared to the noise.

A related practice that deserves comment is the strategic ignoring of borderline cases, exceptions, and avoidable complexities. I consider this one of the more idiosyncratic rather than

[8] I will elaborate more on these sources of uncertainty in the section on triangulation.

universal practices. Some researchers strongly lean towards simplifying approximations, while others lean strongly towards exhaustive detail. For example, in the early work on regulation of eukaryotic genes, some researchers side-skirted the complexity of chromatin – a structure in which DNA is wound around proteins in a complex pattern inside the nucleus of the cell. Their justification for ignoring this was that there were at least some examples of gene regulation systems that could be faithfully reconstituted in vitro using purified transcription factors, regulatory proteins and DNA, without any chromatin. Therefore, at least some of the mechanisms of gene regulation could be understood independently of chromatin. For some researchers, it seemed most efficient to study these simpler cases first. Other researchers focused on understanding the role of chromatin in gene regulation, which also exists.

Similarly, sometimes a provisional mechanistic model will explain most but not all properties of an example case; or a generalization will hold for most but not all the examples of a phenomenon. Some researchers feel it is most strategic to focus first on finding a simple mechanistic model that explains most properties in most cases, worrying about exceptions later; whereas others feel that all known examples and properties should constrain one's working model at all times.

Over time, the findings of both camps tend to be synthesized into a unified mechanistic model. For example, contemporary models of Eukaryotic gene regulation include mechanisms found in prokaryotes (such as: sequence specific DNA binding proteins, cooperative binding, mutually exclusive binding due to steric hindrance, protein interactions between activators and transcriptional machinery, and functional control of regulatory proteins by metabolites) working in concert with additional mechanisms specific to Eukaryotes (such as long-range "enhancer" control elements and chromatin-dependent mechanisms) (Watson, 2008). Thus, this difference in scientific style does not lead to divergent conclusions, just different paths of investigation. It is hard to know whether one or the other approach is more efficient, or whether diversification (having both) is optimal, as we can never know how long it would have taken to arrive at any particular state of knowledge by another historical path.

4. The iterative, cumulative nature of research

A very common feature of biology research is that results or effects are not isolated but build upon one another cumulatively (Platt, 1964). For example, to test the model shown in Figure 1A, a researcher might test the specific assumption that this protein binds to this DNA binding site. An experiment to test this *in vitro* could be to take purified DNA containing the binding site and subject it to a chemical or enzyme that degrades DNA, either in the absence or presence of the purified protein. If the protein binds specifically to the DNA sequence as shown, it is reasoned that this should protect this specific location on the DNA from breakage, compared to others.

If this result were observed, the researcher might further refine this by testing if the effect is concentration-dependent, because according to principles of chemistry and physics, higher protein concentration should result in the DNA site being bound more of the time, and thus provide more protection. That second experiment would include the no-protein and high-concentration protein conditions of the first experiment as negative and positive controls, along with intermediate concentrations that test an additional prediction. If the control conditions reproduce the original result, this one experiment both replicates the original finding and generates new information.

If the originally observed effect were not seen in the controls, however, the researcher would be triggered to investigate. Among the potential explanations considered would be that the new experiment was performed incorrectly, or that the original effect had been due to a fluke or error. Investigation might entail identifying any aspect that differed between the experiments (such as the batch of purified protein or DNA used), and systematically testing if this difference

accounted for the different results; or generating a completely fresh set of reagents to repeat the experiment from scratch.

A single published study can include a chain of experiments with a dozen such steps, each contingent on the previous conclusion. These sequential experiments build on each other, progressively refining the empirical claims (e.g., “the protein binds” is refined to “the binding constant is x”). These experiments are linked together in the elaboration of the model’s causal claims. Moreover, many of the elemental factual claims end up being replicated incidentally many times (“micro-replications”), both within a study and across studies in the same and other laboratories (Guttinger, 2018, 2019; Lewandowsky & Oberauer, 2020; Nichols, Oli, Kendall, & Boomer, 2021).

Because of this factor, factual claims about effects that are stated in papers are often supported by far more observations than those shown in the graph or used in the reported statistical calculations. The displayed data are intended to be taken as representative examples, rather than full accounts of the evidential warrant of the claim. Some of the additional evidence may be available to the reader – for example, the control conditions shown in other experiments of the same paper or in other papers. Some of this evidence may not be open to external inspection (not mentioned, or “personal observations”), however[9].

Within biology it is considered the prerogative (and some might say, the duty) of a reader to consider only the evidence that is shown, when deciding his or her own confidence in an empirical claim[10]. On the other hand, the readers of a primary research manuscript are mainly other researchers with related expertise, and paradigmatic example cases are often studied in many laboratories. Therefore, it is often the case that the reader can corroborate (or cast doubt on) the claim based on their own independent observations.

Thus, each researcher in a field holds their own estimate of the credibility of every claim, and that estimate is based on a combination of the data presented in published papers and their own direct observations. Each researcher may also take into account the plausibility of the claim in the face of other factual or theoretical knowledge they possess. Because different researchers in a field possess different sets of direct observations, factual and theoretical knowledge, they often disagree on the credibility of a claimed effect[11]. These disagreements stimulate effort towards resolution, but are not considered a sign of a problem.

A consequence of this situation is that outside observers (such as metascience researchers) will not gain clear insight into the epistemic situation in the field if they presume that claims made in a paper are predicated entirely on the data presented in a single figure or even in that paper; or if they presume that a claim of a “significant effect” in one paper would be interpreted as an established fact within the field.

It might even be the case that claims in some literatures will turn out to be *more replicable* than one might expect if they had been based only on the data shown in the paper (because researchers selectively report the claims for which they have an abundance of unreported replications, or those corroborated by other supporting theory or evidence).

---

[9] Where a claim is based on more data than shown in a paper, making this explicit would clarify the epistemic structure of arguments to a wider audience of readers. Often it is not stated, however. One reason may be that this is presumed to be understood by the intended readers. More recently, many journals no longer permit “data not shown” or “personal observation” statements. It is worth considering whether such rules increase rigor or merely impede transparency.

[10] An important component of PhD training programs in Biology has always been teaching students how to critically evaluate the literature.

[11] These individual judgements of the credibility of claims are made available in the literature in the background and the discussion sections of primary research articles, where authors discuss their contextualized evaluations of previously reported results. These sections of a research manuscript are therefore more relevant to the epistemic support of the paper’s claims than may be commonly assumed.

Claims in the literature are understood by researchers to range from well-supported to weakly-supported, in a way that can't be inferred from the power or p value or confidence interval of the isolated report, but that might be inferred from the aggregated remarks in background sections, discussion sections, and review articles within that field[12].

For the sake of transparency, it would be desirable for this "dark data" to be available to everyone, but this would entail a substantial bookkeeping effort, especially if one wanted to combine the evidence of the entire community. As data stewardship becomes more automated and open data becomes more standard, it may be the case in the future that the cumulative evidential weight of micro-replications can routinely be collated to obtain a meta-analysis that includes incidental observations that are not otherwise reported.

5. Triangulation

One research practice that I was explicitly taught and that I observe to be widely practiced in experimental biology is the intentional pursuit of multiple independent conceptual and experimental approaches to confirm a conclusion (Kuorikoski & Marchionni, 2016; Munafo & Davey Smith, 2018). Every experimental method makes some assumptions and requires some indirect interpretation. Even when these assumptions are well founded and interpretations defensible, we tend to be wary of the possibility of being fooled if the assumptions or interpretations are wrong. Therefore, we are rarely content to accept a scientific hypothesis or model until it is confirmed by independent means. This is of course not unique to Biology, but is a factor contributing to research rigor in fields where it is employed.

I think of this as a "non-statistical" method because it is often employed in the absence of any numerical statistical tests, and its contribution to inference is not related to quantifying, controlling or reducing uncertainties relating to random variation or sampling. I suppose one could conceptualize this as a kind of statistical practice, inasmuch as technical assumptions about one's materials or measurement instruments could be considered part of the hypothesis "tested" in an NHST (Greenland, Rafi, Matthews, & Higgs, 2022). If the same mechanistic prediction has been borne out using multiple methods that entail different assumptions, the likelihood becomes low that those assumptions are the aspect of the null hypothesis falsified by the statistical test. One can imagine placing priors on these material caveats and calculating this likelihood numerically. But this line of thinking would be quite foreign to most biologists.

To give a concrete example, above I described testing part of the model in Figure 1A – that the DNA binding protein binds specifically to the indicated sites of the DNA – using a "protection assay". This test often produces large effect sizes and highly reproducible results. But the inference that the protein binds to the DNA site is still indirect and dependent on our theory of how the assay works. It is possible that a protein which in fact binds specifically to that DNA site will reproducibly fail to protect it from degradation; or a protein could reproducibly prevent degradation at the DNA site by some mechanism other than by binding to it.

Therefore, having observed such an effect, rather than repeat this assay multiple times, it would be considered more informative to follow up with a completely different test of DNA binding. For example, one can also test whether a purified regulatory protein binds to DNA using a gel shift assay (if a protein sticks to the DNA, it is reasoned, this slows the progress of radiolabeled DNA through a matrix). One can include various controls to show that the change in the radioactive pattern is consistent with the proposed mechanism, such as showing that other DNA molecules are not shifted by the protein, other proteins don't shift this DNA, and the amount of DNA that gets shifted depends on the concentration of the protein.

[12] Prediction markets for replication studies offer another interesting window into researchers' ability to assess the epistemic status of specific claims in their fields (Dreber et al., 2015; Gordon, Viganola, Dreber, Johannesson, & Pfeiffer, 2021; Munafo et al., 2015).

Either one of these assays can give a false result, altering the observable measurements (radioactive bands) by some mechanism other than the intended one. When this happens, the observable outcome of the experiment is *perfectly reproducible*, but the interpretation is wrong. But the causes of rare artifacts in each method are different. Therefore, if the same effect is found by both methods, the interpretation that the effect is real is more plausible than the interpretation that two separate rare exceptions occurred.

But even those two methods could have come common artifacts. Therefore, such a postulated interaction would be tested in other ways. For example, one might also do a co-immunoprecipitation assay (using an antibody to precipitate the protein) or a DNA pulldown assay (using a high affinity tag like biotin to trap the DNA molecule) to test whether physically extracting one molecule from solution brings the other one with it. One might use a crosslinking assay (chemicals or UV light is used to covalently cross-link molecules including DNA and proteins that are in close proximity). Equivalents of the DNA protection assay and crosslinking assay can also be performed *in vivo*. Imaging methods like fluorescent resonance electron transfer (FRET) or surface plasmon resonance (SPR) may be used to optically detect contact or close proximity of molecules.

At a wider level of abstraction, models like the ones in Figure 1 were based on convergent evidence from extremely different types of research: genetics (studying the phenotypes of mutations), biochemistry (purified proteins or in vitro reactions), biophysics (x-ray crystallography or other ways of determining atomic-level structure), phylogeny (observing what regions of proteins or DNA control elements are conserved between genes and across species), molecular biology (constructing artificial systems that ought to work if our mechanistic model is accurate) and technological applications (using the model to design tools that actually work in biotechnology or as research tools). Often many of these approaches were pursued simultaneously in the same laboratory. In my experience, this is one of the strongest and most common explicit epistemic strategies used by biologists for addressing uncertainties and avoiding errors.

To the extent that triangulation is a crucial epistemic practice in biology, discussions debating the merits of this approach (Heesen, Bright, & Zucker, 2019; Stegenga, 2009; Stegenga & Menon, 2017) may be of interest to practicing biologists.

Statistical malpractice

Some research fields, such as social psychology, are accused of generating mostly unreproducible results as a consequence of mis-use of statistical methodology. The “replication crisis” in psychology has led to widespread calls for reforms, including improving training in statistical theory and methodology, and normalizing pre-registration as a way to prevent “researcher degrees of freedom”, “questionable research practices”, and “p-hacking” that are believed to be the root of unreproducible results in that literature (Benjamin et al., 2018; Gosselin, 2020; J. P. Ioannidis, 2005; J. P. A. Ioannidis, 2016).

Above I have argued that certain fields of biology have historically not depended on or even appealed to formal statistical methods. I suggested that the reason these biological fields nevertheless made so much progress is partly that statistical methods were unnecessary (the phenomena were such that stochastic variation was negligible), and partly because alternative non-statistical inferential methods were effective at mitigating whatever uncertainties there were, statistical or non-statistical.

However, in recent decades there has been increasing pressure for biologists in these fields to couch their research results in null hypothesis testing narratives, and to report NHST statistics (p values) in support of their empirical claims. Although there has been a substantial increase in the reporting of p values and other formal statistics, I suspect these are usually purely performative (Gigerenzer, 2004; P. Stark & A. Saltelli, 2018). Properly implemented

statistical null hypothesis testing (with rigid pre-specification of statistical null hypotheses, sampling plans, data analysis procedures, and interpretation) is still quite rare in my experience.

I think most biologists are not aware that prospective study design is implied or required by NHST tests, and are still doing flexible, data-driven exploratory research, making decisions about sampling and data analysis on the basis of interacting with the data, quite often hypothesizing after results are known, and then only as a bureaucratic exercise crunching their numerical data through an NHST test procedure to obtain an invalid p value. Even if this is being done innocently out of ignorance, it is definitely statistical malpractice, and it should be stopped (P. B. Stark & A. Saltelli, 2018). But in many cases I think the research is sound and the conclusions are well supported; the remedy is to refrain from reporting p values or otherwise implying an NHST was conducted (Gigerenzer, 2018). Absent such claims, the research practices *per se* are not "QRPs".

If it is true that gross misuse of statistical methods has not led to irreproducibility in biology in the same way it has in some other sciences, this would be interesting and worth understanding, even though I'm not advocating that the practice be continued. I have not here established that this is the case, however. That would entail an extensive program of research to empirically study the professed attitudes and actual practices of researchers in various subdisciplines, and to objectively determine the statistical and other epistemic implications of those attitudes and practices (Peterson, 2015; Samuel & Konig-Ries, 2021). I think this would be well worth undertaking. It is important, however, for practicing biologists to become more integrally involved in such research, to help correct the current mismatch between professed, perceived, expected, and actual epistemic norms.

Closing thoughts

In the process of asking how experimental biology could make so much progress without using formal statistical procedures, several possible explanations surfaced. Many different characteristics of the research and many different research practices appear to separately have the potential to improve rigor including reproducibility, often without any formal statistical methods; and this is not exhaustive. I have focused on later stages of study when some mechanistic models have been formulated, but at earlier stages there can also be systematic and rigorous programs of observation and testing to determine the nature of phenomena prior to formulating any theory or testing any hypothesis (Burian, 2013; Steinle, 1997).

Perhaps the question should have been, considering how many factors can be leveraged to ensure research rigor in addition to, or without, formal statistical methodology, how is it that reproducibility projects find so many irreproducible results? One possible answer is that the crisis narrative is overblown or misinterpreted, even in those cases (Amrhein et al., 2019; Fanelli, 2018; Leonelli, 2018; Munafò, Chambers, Collins, Fortunato, & Macleod, 2022; Redish, Kummerfeld, Morris, & Love, 2018). Although many analyses presume that questionable research practices or statistical malpractice are the main cause of irreproducibility, others have questioned this assumption (Fanelli, 2018; Head, Holman, Lanfear, Kahn, & Jennions, 2015; Reinagel, 2023; Rubin, 2022; Ulrich & Miller, 2020).

If the inferential goal of research is the successive refinement of mechanistic models rather than establishing claims of associations or effects, this has implications for metascience. Assessments of the rigor of research or solidity of knowledge in such fields should address this at the level of the mechanistic models, as these are the substantive "claims" of the research program. The aspects of a working model that are considered to be established, reliable facts within a field should be held to a different standard from those considered controversial or tentative. The scope of generalization or universality of a claim should not be presumed, and may be quite delimited. This level of abstraction is best conveyed by review articles and

textbooks, but could be gleaned from primary research articles by taking into account the narrative reasoning in the introduction and discussion sections as well as results and figures.

Research and educational reforms centered on encouraging improved NHST practices are likely warranted for any field of research where NHST is used, particularly if it is the primary inferential methodology, if intrinsic variability is high, if the effects under study are small, if there is a lack of constraining theory or mechanisms, or if crosschecks such as micro-replications and triangulation are absent or impossible.

Some fields in biology are not truly engaging in statistical null hypothesis significance testing, however. Interventions aimed at enforcing adherence to NSHT procedures in those fields could cause harm by disrupting non-statistical research practices and inferential methods that are in fact the basis of existing research rigor. In such cases it would be better to advocate for more transparency about the *de facto* research practices and methods – including coming clean about the fact that statistical null hypothesis significance testing is not being used and p values are not the basis of inferences.

This shift could usefully focus attention on articulating, validating and communicating the real nature of, and basis of, claims in those fields. This would in turn facilitate a more critical assessment of whether the methods in use are adequate to the inferential requirements of the research in general or in any given case. This will be increasingly important as the field of biology is rapidly changing, and many previously successful research workflows or inferential heuristics may prove to be inapplicable to new contexts.

Bibliography

Amrhein, V., Trafinnow, D., & Greenland, S. (2019). Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication. *American Statistician, 73*, 262-270. doi:10.1080/00031305.2018.1543137

Andrade, C. (2021). HARKing, Cherry-Picking, P-Hacking, Fishing Expeditions, and Data Dredging and Mining as Questionable Research Practices. *J Clin Psychiatry, 82*(1). doi:10.4088/JCP.20f13804

Baetu, T., & Carl, F. C. (2015). In search of mechanisms: discoveries across the life sciences. *History and Philosophy of the Life Sciences, 36(3)*(3), 459-461. Retrieved from <Go to ISI>://CCC:000358146300016

Baetu, T. M. (2019). Mechanisms in Molecular Biology. *Mechanisms in Molecular Biology*. Retrieved from <Go to ISI>://CCC:000506218100003

Bechtel, W. (2006). *Discovering cell mechanisms : the creation of modern cell biology*. New York: Cambridge University Press.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., . . . Johnson, V. E. (2018). Redefine statistical significance. *Nat Hum Behav, 2*(1), 6-10. doi:10.1038/s41562-017-0189-z

Bertani, G. (2004). Lysogeny at mid-twentieth century: P1, P2, and other experimental, systems. *Journal of Bacteriology, 186*(3), 595-600. doi:10.1128/Jb.186.3.595-600.2004

Burian, R. M. (2005). *The epistemology of development, evolution, and genetics : selected essays*. Cambridge, England ; New York: Cambridge University Press.

Burian, R. M. (2013). Exploratory Experimentation. In W. Dubitzky, O. Wolkenhauer, K.-H. Cho, & H. Yokota (Eds.), *Encyclopedia of Systems Biology* (pp. 720-723). New York, NY: Springer New York.

Cold Spring Harbor Laboratory of Quantitative Biology., Cairns, J., & Delbrück, M. (1966). *Phage and the origins of molecular biology; [essays]*. Cold Springs Harbor, N.Y.

Devezer, B., Nardin, L. G., Baumgaertner, B., & Buzbas, E. O. (2019). Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *Plos One, 14(5)*(5), -. Retrieved from <Go to ISI>://CCC:000467949100024

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y. L., . . . Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences of the United States of America, 112(50)*(50), 15343-15347. Retrieved from <Go to ISI>://CCC:000366404200048

Dupré, J. (1993). *The disorder of things : metaphysical foundations of the disunity of science*. Cambridge, Mass.: Harvard University Press.

Fanelli, D. (2018). Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proc Natl Acad Sci U S A, 115*(11), 2628-2631. doi:10.1073/pnas.1708272114

Fidler, F., Chee, Y. E., Wintle, B. C., Burgman, M. A., Mccarthy, M. A., & Gordon, A. (2017). Metaresearch for Evaluating Reproducibility in Ecology and Evolution. *Bioscience, 67*(3), 282-289. doi:10.1093/biosci/biw159

Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. *PLoS One, 13*(7). doi:ARTN e0200303
10.1371/journal.pone.0200303

Gelfert, A. (2016). *How to Do Science with Models : A Philosophical PrimerSpringerBriefs in Philosophy,* (1st ed., pp. 1 online resource (X, 135 pages 138 illustrations, 137 illustrations in color). doi:10.1007/978-3-319-27954-1

Gelfert, A. (2019). *Probing Possibilities: Toy Models, Minimal Models, and Exploratory Models*, Cham.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*(5), 587-586-586. doi:10.1016/j.socec.2004.09.033

Gigerenzer, G. (2018). Statistical Rituals: The Replication Delusion and How We Got There. *Advances in Methods and Practices in Psychological Science, 1(2)*(2), 198-218. Retrieved from <Go to ISI>://CCC:000746358200005

Gordon, M., Viganola, D., Dreber, A., Johannesson, M., & Pfeiffer, T. (2021). Predicting replicability-Analysis of survey and prediction market data from large-scale forecasting projects. *Plos One, 16(4)*(4), -. Retrieved from <Go to ISI>://CCC:000640604500074

Gosselin, R. D. (2020). Statistical Analysis Must Improve to Address the Reproducibility Crisis: The ACcess to Transparent Statistics (ACTS) Call to Action. *Bioessays, 42*(1), e1900189. doi:10.1002/bies.201900189

Greenland, S., Rafi, Z., Matthews, R., & Higgs, M. (2022). To Aid Scientific Inference, Emphasize Unconditional Compatibility Descriptions of Statistics. *arXiv*. doi:https://doi.org/10.48550/arXiv.1909.08583

Guttinger, S. (2018). Replications Everywhere Why the replication crisis might be less severe than it seems at first. *Bioessays, 40(7)*(7), -. doi:10.1002/bies.201800055

Guttinger, S. (2019). A New Account of Replication in the Experimental Life Sciences. *Philosophy of Science, 86(3)*(3), 453-471. doi:10.1086/703555
Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biol, 13*(3), e1002106. doi:10.1371/journal.pbio.1002106
Heesen, R., Bright, L. K., & Zucker, A. (2019). Vindicating methodological triangulation. *Synthese, 196*(8), 3067-3081. doi:10.1007/s11229-016-1294-7
Hershey, A. D., & Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol, 36*(1), 39-56. doi:10.1085/jgp.36.1.39
Hubbard, R., Haig, B. D., & Parsa, R. A. (2019). The Limited Role of Formal Statistical Inference in Scientific Inference. *American Statistician, 73*, 91-98. doi:10.1080/00031305.2018.1464947
Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine, 2*(8), e124. doi:10.1371/journal.pmed.0020124
Ioannidis, J. P. A. (2016). Why Most Clinical Research Is Not Useful. *PLoS medicine, 13*(6), e1002049. Retrieved from <Go to ISI>://MEDLINE:27328301
Jacob, F., & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol, 3*, 318-356. doi:10.1016/s0022-2836(61)80072-7
Knuuttila, T., & Loettgers, A. (2021). Biological Control Variously Materialized: Modeling, Experimentation and Exploration in Multiple Media. *Perspectives on Science, 29*(4), 468-492. doi:10.1162/posc_a_00379
Kuorikoski, J., & Marchionni, C. (2016). Evidential Diversity and the Triangulation of Phenomena. *Philosophy of Science, 83*(2), 227-247. doi:Doi 10.1086/684960
Lakens, D., Mesquida, C., Rasti, S., & Ditroilo, M. (2024). The benefits of preregistration and Registered Reports *Evidence-Based Toxicology, 2*(1). doi:https://doi.org/10.1080/2833373X.2024.2376046
Landis, S. C., Amara, S. G., Asadullah, K., Austin, C. P., Blumenstein, R., Bradley, E. W., . . . Silberberg, S. D. (2012). A call for transparent reporting to optimize the predictive value of preclinical research. *Nature, 490*(7419), 187-191. doi:10.1038/nature11556
Lennox, J. G. (2020). *Aristotle on inquiry : erotetic frameworks and domain specific norms*(pp. 1 online resource).
Leonelli, S. (2018). Re-Thinking Reproducibility as a Criterion for Research Quality In *Including a Symposium on Mary Morgan: Curiosity, Imagination, and Surprise*.
Lewandowsky, S., & Oberauer, K. (2020). Low replicability can support robust and efficient science. *Nat Commun, 11*(1), 358. doi:10.1038/s41467-019-14203-0
Luria, S. E., & Delbruck, M. (1943). Mutations of Bacteria from Virus Sensitivity to Virus Resistance. *Genetics, 28*(6), 491-511. doi:10.1093/genetics/28.6.491
Lwoff, A. (1953). Lysogeny. *Bacteriological Reviews, 17*(4), 269-337. doi:Doi 10.1128/Mmbr.17.4.269-337.1953
Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1-25.

Meehl, P. E. (1978). Theoretical Risks and Tabular Asterisks - Karl, Ronald, and Slow Progress of Soft Psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806-834. doi:Doi 10.1037/0022-006x.46.4.806

Meselson, M., & Stahl, F. W. (1958). The Replication of DNA in Escherichia Coli. *Proc Natl Acad Sci U S A, 44*(7), 671-682. doi:10.1073/pnas.44.7.671

Monod, J., & Jacob, F. (1961). Teleonomic mechanisms in cellular metabolism, growth, and differentiation. *Cold Spring Harbor Symposia on Quantitative Biology, 26*, 389-401. Retrieved from <Go to ISI>://MEDLINE:14475415

Morange, M., & Cobb, M. (2020). *The black box of biology : a history of the molecular revolution*. Cambridge, Massachusetts ; London, England: Harvard University Press.

Munafò, M. R., Chambers, C., Collins, A., Fortunato, L., & Macleod, M. (2022). The reproducibility debate is an opportunity, not a crisis. *Bmc Research Notes, 15(1)*(1), -. Retrieved from <Go to ISI>://CCC:000753868000012

Munafo, M. R., & Davey Smith, G. (2018). Robust research needs many lines of evidence. *Nature, 553*(7689), 399-401. doi:10.1038/d41586-018-01023-3

Munafo, M. R., Pfeiffer, T., Altmejd, A., Heikensten, E., Almenberg, J., Bird, A., . . . Dreber, A. (2015). Using prediction markets to forecast research evaluations. *Royal Society Open Science, 2(10)*(10), -. Retrieved from <Go to ISI>://CCC:000377967100003

Murphy, K. R., & Aguinis, H. (2019). HARKing: How Badly Can Cherry-Picking and Question Trolling Produce Bias in Published Results? *Journal of Business and Psychology, 34*(1), 1-17. doi:10.1007/s10869-017-9524-7

Nichols, J. D., Oli, M. K., Kendall, W. L., & Boomer, G. S. (2021). Opinion: A better approach for dealing with reproducibility and replicability in science. *Proc Natl Acad Sci U S A, 118*(7). doi:10.1073/pnas.2100769118

Nirenberg, M. W., & Matthaei, J. H. (1961). The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *Proc Natl Acad Sci U S A, 47*(10), 1588-1602. doi:10.1073/pnas.47.10.1588

Norton, J. D. (2021). *The material theory of induction*. Calgary, Alberta, Canada: University of Calgary Press.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proc Natl Acad Sci U S A, 115*(11), 2600-2606. doi:10.1073/pnas.1708274114

Olby, R. (1994). *The Path to the Double Helix: The Discovery of DNA.* . New York Dover Publications, Inc.

Olby, R. (2002). Meselson, Stahl, and the replication of DNA: A history of "the most beautiful experiment in biology". *Nature, 417*(6885), 121-122. doi:DOI 10.1038/417121a

Peterson, D. (2015). All That Is Solid: Bench-Building at the Frontiers of Two Experimental Sciences. *American Sociological Review, 80(6)*(6), 1201-1225. Retrieved from <Go to ISI>://CCC:000365730200005

Peterson, D. (2021). The replication crisis won't be solved with broad brushstrokes. *Nature, 594*(7862), 151. doi:10.1038/d41586-021-01509-7

Platt, J. R. (1964). Strong Inference - Certain Systematic Methods of Scientific Thinking May Produce Much More Rapid Progress Than Others. *Science, 146(364)*(364). Retrieved from <Go to ISI>://CCC:A19643064C00289

Popovic, G., Mason, T. J., Drobniak, S. M., Marques, T. A., Potts, J., Joo, R., . . . Pottier, P. (2024). Four principles for improved statistical ecology. *Methods in Ecology and Evolution, 15*(2), 266-281. doi:10.1111/2041-210x.14270

Ptashne, M. (2004). *A genetic switch : phage lambda revisited* (3rd ed.). Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press.

Redish, A. D., Kummerfeld, E., Morris, R. L., & Love, A. C. (2018). Opinion: Reproducibility failures are essential to scientific inquiry. *Proc Natl Acad Sci U S A, 115*(20), 5042-5046. doi:10.1073/pnas.1806370115

Reinagel, P. (2023). Is N-Hacking Ever OK? The consequences of collecting more data in pursuit of statistical significance. *PLoS biology, 21*(11), e3002345. doi:10.1371/journal.pbio.3002345

Ross, L. N. (2023). The explanatory nature of constraints: Law-based, mathematical, and causal. *Synthese, 202(2)*(2), -. Retrieved from <Go to ISI>://CCC:001048566000003

Rubin, M. (2022). The Costs of HARKing. *The British Journal for the Philosophy of Science, 73*(2), 535-560. doi:10.1093/bjps/axz050

Salmond, G. P. C., & Fineran, P. C. (2015). A century of the phage: past, present and future. *Nature Reviews Microbiology, 13*(12), 777-786. doi:10.1038/nrmicro3564

Samuel, S., & Konig-Ries, B. (2021). Understanding experiments and research practices for reproducibility: an exploratory study. *PeerJ, 9*, e11140. doi:10.7717/peerj.11140

Schickore, J. (2016). "Exploratory experimentation" as a probe into the relation between historiography and philosophy of science. *Stud Hist Philos Sci, 55*, 20-26. doi:10.1016/j.shpsa.2015.08.007

Schindler, S. (2008). Model, Theory, and Evidence in the Discovery of the DNA Structure. *British Journal for the Philosophy of Science, 59*(4), 619-658. Retrieved from <Go to ISI>://CCC:000261300600002

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366. doi:10.1177/0956797611417632

Stark, P., & Saltelli, A. (2018). Cargo-Cult Statistics and Scientific Crisis. *Significance, 15*(4), 40-42. doi:https://doi.org/10.1111/j.1740-9713.2018.01174.x

Stark, P. B., & Saltelli, A. (2018). Cargo-cult statistics and scientific crisis. *Significance, 15*(4), 40-43.

Stegenga, J. (2009). Robustness, Discordance, and Relevance. *Philosophy of Science, 76*(5), 650-661. doi:Doi 10.1086/605819

Stegenga, J., & Menon, T. (2017). Robustness and Independent Evidence. *Philosophy of Science, 84*(3), 414-435. doi:10.1086/692141

Steinle, F. (1997). Entering new fields: Exploratory uses of experimentation. *Philosophy of Science, 64(4)*(4), S65-S74. Retrieved from <Go to ISI>://CCC:000072709000007

Ulrich, R., & Miller, J. (2020). Questionable research practices may have little effect on replicability. *eLife, 9*. doi:10.7554/eLife.58237

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. *Perspect Psychol Sci, 7*(6), 632-638. doi:10.1177/1745691612463078

Watson, J. D. (2008). *Molecular biology of the gene* (6th ed.). San Francisco; Cold Spring Harbor, N.Y.: Pearson/Benjamin Cummings ; Cold Spring Harbor Laboratory Press.