The British Journal for the Philosophy of Science

Objectivity and Underdetermination in Statistical Model Selection

Journal:	*The British Journal for the Philosophy of Science*
Manuscript ID	BJPS-2020-271.R3
Manuscript Type:	Article
Keywords:	Akaike Information Criterion, error statistics, punctuated equilibrium, evolutionary biology

Objectivity and Underdetermination in Statistical Model Selection

Abstract:
The growing range of methods for statistical model selection is inspiring new debates about how to handle the potential for conflicting results. While many factors enter into choosing a model selection method, we focus on the implications of disagreements among scientists about whether, and in what sense, the true probability distribution is included in the candidate set of models. While this question can be addressed empirically, data often provide inconclusive results in practice. In such cases, we argue that differences in prior metaphysical beliefs can produce underdetermination of results, even for the same data and candidate models. As a result, data alone are sometimes insufficient to settle rational beliefs about nature.

1. Introduction

In the last several decades, classical hypothesis tests and especially p-values have come under heavy criticism from a number of directions (e.g., Wagenmakers 2007; Halsey 2019). Since landmark results by Hirotugu Akaike (Parzen et al. 1998), many scientists have adopted a new approach based on information-theoretic methods, which do not rely on error probabilities such as p-values; instead, these methods select models based on their relative ranking of goodness of fit to the data. In general, these methods invoke different conceptions of evidence and do not lead to the same answers. Efforts to settle on one option as generally best, however, have not produced a definitive consensus (Chakraborty and Ghosh 2011). As a result, scientists must determine which of the available model selection methods they should use relative to their question and background knowledge. Making this choice in an objective way has become a pressing methodological challenge for multiple disciplines, including ecology, economics, psychology, and evolutionary biology (Burnham and Anderson 2002; Wagenmakers and Farrell 2004; Spanos 2010; Aho et al. 2014).

An important factor in these discussions is whether, and in what sense, one believes the set of candidate models to contain the true distribution (Yang 2007; Anderson 2008; Clarke et al. 2013; Aho et al. 2014; Ding et al. 2018b). Pointedly, note that this is a relational property of the *candidate model set*, not any particular model selection *method*. The existing philosophical literature, by contrast, focuses on issues regarding particular methods, such as the Akaike Information Criterion (AIC), and whether its statistical behavior can be interpreted as consistent with a realist or instrumentalist worldview. Our topic here is methodologically prior—it concerns how the relation of the true distribution to the set of candidate models informs the applicability of these methods in the first place.

When the available data leave uncertainty about the relation of the candidate models to the true distribution, scientists in many cases attempt to settle the issue by invoking prior metaphysical beliefs about nature, by which we mean fairly abstract, general expectations about how their systems of study behave (e.g., Anderson 2008; Aho et al. 2014). This opens the door to disagreements about the choice of method based on conflicting views about whether, for example, nature is simple enough or too complex for any finite set of models to adequately represent. Because different model selection methods do not in general return identical results, scientists consequently may find themselves disagreeing about which hypothesis is best supported even for the same dataset and candidate models.

We will argue this conclusion represents a novel type of underdetermination problem, which we call the problem of unequal evidence, and poses an important challenge, as yet overlooked by philosophers, for the objectivity of model selection. We use objectivity here in the sense of being impartial, i.e., that the personal opinions or assumptions of scientists don't influence the practice of model selection in a way that predisposes the outcomes (Longino 1990). The pursuit of impartiality was an important historical motivation for introducing model selection in our case study below, but we do not take a position here on general discussions of the meanings and value of objectivity in science. In particular, we will focus on conflicts arising out of methodologies using Akaike's Information Criterion, the Bayesian Information Criterion (BIC), and severe testing (Mayo and Spanos 2006). We will not address Bayesian methodology in depth due to limitations of space. We start by introducing the problem of unequal evidence in general terms and then provide a motivating example from paleobiology. To achieve a rigorous demonstration of the actual workings at issue, we then show in more technical detail how the

unequal evidence problem is grounded in general features of contemporary model selection practice and theory.

2. The Problem of Unequal Evidence

In this section, we introduce the problem of unequal evidence and highlight its novel features in the context of recent practice-oriented work on underdetermination. Our account concerns a relationship between method and models plus data, and thus differs from the more common account of underdetermination as a relationship between hypotheses and evidence. We start by assuming a shared set of observations *and* models, where the models represent alternative hypotheses about the process causally responsible for generating the observations. We posit that these observations and models are shared by a group of scientists in a given research area who are all interested in determining which model is best supported by the evidence but who disagree about the whether the candidate models contain the true distribution as a result of prior beliefs about nature. We show how disagreement on this point leads to alternative choices of model selection methods, which generally do not return equivalent judgments about the evidence. The definition of the problem of unequal evidence is then that the data turn out to be insufficient to determine a single correct result because differences in prior assumptions about nature, even considering the same data and candidate models, lead to unequal judgments of evidence.

In practice, scientists have many reasons for disagreeing about which method is the correct one to use, but one issue in particular — whether the true process is represented among the candidate models — is both commonly cited (Burnham and Anderson 2002; Johnson and Omland 2004; Wagenmakers and Farrell 2004; Spanos 2010; Aho et al. 2014) and especially relevant to underdetermination. Key statistical theorems about the optimality of existing statistical model selection methods depend on assumptions about the relation of the entire set of candidate models to the true process, in a sense that we define more precisely in Section 5. These assumptions must be met for the methods to generate sound conclusions. While the adequacy of the candidate models for explaining the observed data is sometimes possible to determine empirically, e.g., through model misspecification testing, in many cases the evidence is inconclusive or the relevant tests are unavailable or simply intractable. The resulting uncertainty leaves an important gap in the empirical basis for determining which method or methods apply. Figure 1 illustrates the logical flow of how underdetermination arises in this scenario. In particular, we show in Section 6 that scientists invoke prior beliefs about nature to settle the adequacy of the candidate models on general grounds.

The problem of unequal evidence represents a novel form of underdetermination in several respects. While underdetermination has classically been viewed as a challenge to scientific realism, our interest will be in what this novel formulation can illuminate about how science works. Philosopher Derek Turner nicely characterizes this recent shift in motivation for studying underdetermination:

> "Philosophers of science have moved away from the general question whether scientific theories are always underdetermined by the evidence in order to focus on more localized epistemological problems… The philosophers who have taken this new approach are less interested in making *a priori* arguments about science in general, such as the argument for the Duhem-Quine thesis, and more interested in looking at the ways in which smaller-scale underdetermination problems make

> a difference to scientific practice… How often do these underdetermination problems crop up? When they do crop up, what is their source? How do (and how should) scientists deal with them?" (Turner 2011, 147-8).

We introduce the problem of unequal evidence in this practice-oriented spirit, but we note that the problem's origin in statistical methodology also gives it exceptionally broad relevance across scientific domains.

Classically, philosophers have formulated underdetermination as a problematic relationship between a set of hypotheses and a set of evidence such that the evidence fails to decide among the hypotheses (Stanford 2017). Our formulation is novel in focusing on the relationship between a method of model selection and a set of models and data. The root problem in this formulation is that even a shared set of models and data can prove insufficient to determine the choice of method, leaving a scientist's empirical results sensitive to background views about, for instance, the relative simplicity or complexity of nature. Note that underdetermination in this case does not arise because different metaphysical beliefs lead to different construals of the same model, e.g., as causal or merely correlational, and hence to different judgments being made about the supporting evidence.[1] The problem of unequal evidence can arise even for identical data and interpretations of the candidate models.

The role of metaphysical beliefs is instead most analogous to using values already held by different scientists to break an empirical tie among theories. What must be appreciated is that it happens at the methodologically prior stage of choosing which method to use in quantifying evidence for the theories. Indeed, many arguments in the classical underdetermination debate presuppose that problems only arise when two or more theories are equally well supported by the evidence, e.g., debates over whether any two theories in the history of science have ever actually been equally well supported (Laudan and Leplin 1991) or whether equally good alternatives can always be constructed for any theory (Kukla 1996). These are precisely not the problem in the scenario we described, which arises as a result of finding unequal evidence for the models.

Another novel feature is that the problem of unequal evidence concerns what happens when a plurality of statistical methods are operating simultaneously in a community of scientists all working on similar kinds of problems. While philosophers have investigated underdetermination in light of multiple conceptions of evidence, including deductive falsification, Bayesian credence, and statistical likelihood, they have generally treated each conception in isolation (e.g., Dorling 1979; Kieseppä 2001; Spanos 2010). Some notable exceptions include Michael Dietrich and Robert Skipper Jr.'s use of debates over theories of molecular evolution to show how evolutionary biologists sought to "manipulate underdetermination" by arguing for differently weighted assessment strategies in order to both create and break ties between competing models (Dietrich and Skipper 2007). Samir Okasha has also argued that the distinction between theory and data can only be drawn contextually, which blocks any single all-purpose way of determining whether theories share identical sets of empirically testable consequences (Okasha 2002). Additionally, Elliott Sober has shown that the holistic character of falsification under deductive logic fails to hold for hypothesis testing using statistical likelihood (Sober 2004).

None of these papers, however, directly raise the possibility that prior metaphysical views may justify principled disagreements about how to quantify evidence even when scientists share the same data and candidate models. Malcolm Forster captures the general spirit of the challenge

[1] By metaphysical belief, recall that we mean an abstract, general expectation about how nature is. See also Section 7.

we are raising here: "There is no simple and no universal model of model selection, for the success of a selection method depends greatly on the circumstances, and to understand the complexities, we have to model the situation in which the model selection takes place. For philosophers of science, this is like making assumptions about the uniformity of nature in order understand how induction works. The problem is the same: How can we make assumptions that don't simply assume what we want to prove?" (Forster 2002, 87). The novelty of the unequal evidence problem in this respect is therefore to highlight the importance of an overlooked premise — whether the candidate models contain the true distribution — for determining what evidence the data provide. Given conflicting assumptions about the behavior of the natural system, conceptual critique is crucial to ensuring the objectivity of model selection as a shared social practice (Longino 1990, Chapter 4).

The problem of unequal evidence therefore hinges on the implications of disagreement about the natural system for the adequacy of the candidate model set and hence the applicability of alternative model selection methods. While the unequal evidence problem therefore shares some features of Stanford's (2006) unconceived alternatives argument, it does not rely on historical inductions about how science changes over time. Stanford argues that the historical record of scientific change shows that at any point in time there are hypotheses scientists have not yet conceived of but would agree are as equally good as or better than the ones they already have. As a result, Stanford concludes we should never be confident that our present set of hypotheses contains the true theory of nature. While he arrives at a global conclusion about the inadequacy of any candidate set, we focus on disagreement on precisely this issue among scientists in particular statistical analyses. The problem of unequal evidence therefore articulates a general but not *a priori* way in which potentially persistent underdetermination can arise in local scientific practices.

3. Unequal Evidence in Paleobiology: A Motivating Example

This section provides a case study in paleobiology illustrating the conditions needed for the problem of unequal evidence to arise in practice. For our purposes, paleobiology is a field that studies evolutionary changes on large time scales using fossil and other data sources (Currie 2019). The field has been an emerging focus for philosophers studying underdetermination in practice, with a number of recent studies examining data models (Bokulich 2018), fossil specimen preparation (Wylie 2019), and testing hypotheses about evolutionary events (Cleland 2001; Turner 2011). Our case study adds to this literature by examining paleobiologists' practices for classifying patterns of change in the traits of fossil lineages according to the qualitatively distinct "modes" of behavior they exhibit over geologic time series.

The challenge and importance of classifying these patterns has its roots in the debate over punctuated equilibrium versus neo-Darwinian gradualism in the 1970s (Eldredge and Gould 1972; Sepkoski 2012), and traces some of its key concepts back to George Gaylord Simpson's classic *Tempo and Mode in Evolution* (Simpson 1944). According to the punctuated equilibrium theory, most evolutionary change is concentrated during the phylogenetic branching of lineages in rapid bursts of speciation. Much longer episodes of relative morphological invariance, or stasis, follow speciation events. This evolutionary mode contrasts with incremental directional change within and between related lineages (gradualism), and sometimes with lineage patterns that cannot be distinguished from random trends.

But how to identify these patterns in an objective way and "test" their relative dominance in the fossil record? The relevant data consists of traits of species measured in sequences of fossil populations in successive sedimentary layers from a given locality or region (Gould and Eldredge 1977; Gingerich 1985). When measurements are available from multiple specimen populations from the same lineage over time, paleobiologists can construct a time series based on the population averages and variances at different time horizons. Example traits include the curvature of mollusk shells and the width of mammal molar teeth, and some example time series are shown in Figure 2. The data scientists have accumulated currently includes traits from several hundred fossil lineages, and these are broadly accepted in the field, ongoing methodological challenges with analyzing the underlying fossil record notwithstanding.[2]

Given a shared dataset, however, paleobiologists have applied multiple distinct statistical approaches to classifying individual time series (Lidgard and Hopkins 2015). Early statistical analyses applied null models to test if a trait exhibited a pattern of change over time that differed significantly from Brownian motion (also known as a random walk), but these suffered from low statistical power (Roopnarine 2001; Sheets and Mitchell 2001). In the early 2000s, Gene Hunt introduced a novel approach based on information-theoretic model selection using the AIC (Hunt 2006, 2007). His approach classifies a time series by selecting one among several models of evolutionary mode (stasis, gradualism, Brownian motion) that shows the highest goodness of fit, as measured by a version of the AIC. However, it does not conduct a hypothesis test in the sense of requiring the evidence for the best model to surpass a threshold limit on the probability of false positives (i.e., Type 1 errors). The BIC could also serve to pick out the true model in this context, although it hasn't been applied in practice yet.

Hopkins and Lidgard (2012) adapted Hunt's methods by requiring the best model to be substantially better supported than the next best model according to a rule of thumb threshold. This is similar in spirit to controlling the false positive rate of a hypothesis test but in practice does not map directly onto Type 1 or 2 errors (Cullan et al. 2020). More recently, Kjetil Voje has further developed Hunt's approach by incorporating a set of misspecification tests to help determine whether the best-fitting model's statistical assumptions are also consistent with the observed data (Voje 2018). Finally, Cullan et al. (2020) have introduced a novel method for hypothesis testing that enables control of the Type 1 error rate but does not require null models.

Behind this proliferation of approaches lie deeper theoretical questions about how one should analyze the statistical evidence provided by fossil trait time series. Hunt's adoption of the AIC led a quiet shift away from paleobiology's prior emphasis on using error probabilities in classical hypothesis testing to quantify the evidence for a model. As we'll see in more detail below, the AIC has no fixed interpretation in terms of error probabilities across contexts of application (Ding et al. 2018b). Hopkins and Lidgard, by contrast, relied on Akaike weights, which some have argued are interpretable as the probability of a model being true (Burnham and Anderson 2002). Voje's approach, by contrast, combines a goodness-of-fit criterion (the AIC) with null hypothesis tests for model adequacy. Cullan et al. (2020) have reintroduced a focus on error probabilities as a way to quantify evidence for the best model; their procedure generalizes beyond the classical basis of hypothesis tests in likelihood ratios.

Which method and associated conception of evidence is correct? As we'll argue more generally below, it depends on how one construes the relationship of the probability models to the real data-generating processes. While everyone recognizes even the most complex models

[2] Data models are very important in this context but outside the scope of our discussion.

used by Hunt et al. are vastly simplified compared to the full evolutionary processes, Voje takes the stasis model, for example, to accurately represent some of the trait time series at a coarse-grained level. In this case, it is sensible to ask whether some coarse-grain mathematical function of the true probability distribution is contained in one of the candidate models. Alternatively, Hunt et al. (2015) treated their suite of models as providing first-order approximations to time trends in the observed fossil traits. One way to interpret this stance is that the candidate models come "close" to but do not strictly include even a coarse-grained representation of the true process. It is reasonable, then, to ask which model contains the distribution that best approximates the observed data. Even so, there are alternative ways to construe what it means to best approximate the data, and these correspond to different ways of quantifying evidence.

In sum, while the introduction of statistical methods has arguably advanced scientists' ability to impartially detect trends in fossil traits, a plurality of methods are in current use which quantify evidence in conflicting ways and make different assumptions about the adequacy of the candidate model set relative to the true distribution. Considerable empirical uncertainty remains about this latter question, leaving the door open to scientists to choose one method for their analyses based on prior beliefs about the simplicity versus complexity of macroevolutionary processes relative to the available models. The choice of method does lead to conflicting results for analyses of particular trait data (Cullan et al. 2020), and its impact on broader conclusions about the importance of punctuated equilibrium are unknown. As a result, the case provides an example of how the problem of unequal evidence can arise in scientific practice.

4. Evidence in Statistical Model Selection

We now present a more general analysis of how the problem of unequal evidence can arise with current model selection methods. We start by introducing in more detail some common ways scientists quantify evidence using the AIC, BIC, Akaike weights, and error statistical severity of a test. Our discussion is inspired by common practices of scientists (e.g., Wagenmakers and Farrell 2004; Hunt 2007). Even so, our intention is more modest than expansive; we do not aim to address the full range of possible interpretations of these methods in light of philosophical positions on the nature of probability and interpretation of statistical models. What matters for our argument is that the problem of unequal evidence can arise among at least some positions on these issues as relevant to scientific practice. This section shows that alternative ways of quantifying the support for the same candidate models and data can differ (most vividly) to the degree that they select entirely different models as best. Crucially, the issue is not that models have different numerical scores under alternative methods (as these are not especially meaningful in isolation), but that differences in AIC or BIC scores, for example, can lead to substantially different judgments about evidence.

The basic setup of statistical model selection includes a set of parametric probability models M_1, M_2, … M_k whose parameters index non-redundant spaces of probability distributions. The general purpose of model selection in this context would be to identify which of the models contains the distribution that performs best relative to the goal of the statistical analysis, which may be predictive accuracy on new data or correctly identifying the process that generated the observed data. For example, we could have three regression models:

$$Y = \mu + \beta_1 X + \epsilon,$$
$$Y = \mu + \beta_1 X + \beta_2 X^2 + \epsilon,$$
$$Y = \mu + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon,$$

where Y is the response variable, X is a single covariate, and μ, $\beta_1 \ldots \beta_3$ are parameters. These models overlap but are not redundant because they include different sets of probability distributions (in a model-theoretic sense). Being clear up front about the goal of selecting a model is important: the most complex model might be true, but for small sample sizes a simpler model may still provide better predictive accuracy due to smaller errors in its estimated parameters.

Given data D, the first step is to estimate parameter values for each model, for example using maximum likelihood. This identifies a single probability distribution from each model that will stand for that model during the selection process. Maximum likelihood estimation also provides us with a measure of how well the model accounts for the data, i.e., the likelihood of the data given the model. In a generic case where each model M_i has parameters θ_i, it's common to write this likelihood as $L(D,\theta_i)$. The likelihood is central to most information criteria, such as the AIC and BIC, as well as hypothesis testing (e.g., likelihood ratio tests).

Akaike's work on the AIC initiated a deep transformation in statistics toward information theory as a foundation for statistical inference. In particular, the AIC is based on the Kullback-Leibler (KL) divergence as a way of quantifying the differences between two probability distributions (Burnham and Anderson 2002):

$$D_{KL}(f,g) = \int_{-\infty}^{\infty} f(x) * \log\left(\frac{f(x)}{g(x)}\right) dx$$

The KL divergence quantifies the degree to which continuous distribution *g* can be substituted for *f*. For example, we can think of *f* to be the true distribution and *g* to be a candidate distribution from one of our models. The logarithm term in the integral compares these two functions over all values of *x*, and returns a non-zero value when *f* and *g* are unequal. To get the overall divergence, any differences are weighted against the probability density of *f* at that value of *x* and summed. The KL divergence is zero if and only if the two continuous distributions are equal, and otherwise takes on positive values.

The AIC provides a statistical estimate of the KL divergence between the true and proposed distributions. It is defined as follows:

$$AIC_i = -2\ln\left(\hat{L}_i\right) + 2k_i.$$

Here k_i stands for the number of model parameters in model M_i, and $\hat{L}_i$ is the estimated likelihood of M_i based on maximum likelihood estimates $\hat{\theta}_i$ for the parameters θ_i. A lower AIC score means a better match between the model and the true distribution. The first term of the equation is a function of the model likelihood, and the second term accounts for the fact that models with more parameters tend to fit data better. This latter term is commonly called a "complexity penalty" in the literature. In the derivation of the AIC, it arises as a first-order correction that ensures the AIC is an unbiased estimator of the KL divergence when the true data generating distribution is inside or close to one of the models in the model set. When the true distribution is far from any of the models, the Takeuchi Information Criterion (TIC) provides a generalized approach using the same principles as the AIC, but it is unfortunately hard to calculate in practice (Burnham and Anderson 2002). For cases with small sample sizes relative to the number of model parameters (e.g., N=20 and k=5), there is a corrected version of the AIC, called the AICc, which tweaks the complexity penalty above (Burnham and Anderson 2002).

The key difference between the BIC and AIC is in their complexity penalty. While the AIC's penalty stays the same for any sample size *n*, the BIC's penalty grows logarithmically with *n*:

$$BIC_i = -2\ln\left(\hat{L}_i\right) + k_i \ln(n).$$

This greater penalty on complexity has the effect of tending to pick simpler models than the AIC because it amplifies the difference between models with different numbers of parameters. This effect has the virtue of making the BIC statistically consistent: when two models both contain the true distribution but differ in complexity, the BIC will pick the simpler model asymptotically. In contrast, the AIC will pick the more complex model with non-zero probability even as the sample size increases to infinity.

Despite their differences in how they penalize model complexity, the AIC and BIC both share the same underlying formula for statistical evidence: calculating the difference of scores between the best and second-best models. In other words, if the true data generating distribution is f(x) and we have two competing models $g_1(x)$ and $g_2(x)$, we can measure the evidence supporting g_1 versus g_2 by taking the difference between their AIC or BIC scores, denoted ΔAIC and ΔBIC. This is a generalization of the classical likelihood ratio (which is equivalent to the difference of log-likelihoods), and reduces mathematically to it in certain cases (Royall 1997). There is no general principled threshold at which one model is definitively better supported, but Burnham and Anderson (2002) suggested a common rule of thumb that assigns weak evidence when ΔAIC > 2 and strong evidence when ΔAIC > 10.

Two other conceptions of evidence are also relevant for our purposes: Akaike weights and severe testing. Both aim to say something more directly interpretable about the probability of making a correct or incorrect choice of model. Akaike weights are based on the ΔAIC values for each model compared to the best-scoring model, and normalize how well any model performs relative to the candidate set as a whole. For a more detailed introduction and definition, see (Burnham and Anderson 2002). Several scientists, including Akaike himself, have argued that these weights can be interpreted as a model's probability of being correct given the data, i.e. that it has the smallest KL divergence to the true distribution among the candidates (Akaike 1981; Burnham and Anderson 2002; Wagenmakers and Farrell 2004). Akaike proposed the weights serve as an approximation to a posterior model probability in a Bayesian context, but they have also been applied much more widely. In any case, treating a model's Akaike weight as the probability the model is correct provides a very different way than ΔAIC to understand the evidence for candidate models. If we select a model based on its weight exceeding some fixed threshold, e.g., 0.95, then we have implicitly relativized this threshold to the set of candidate models we are considering. However, this dependence on the candidate set drops out if we compare the ratio of Akaike weights for the best and second-best models, as Hopkins and Lidgard (2012) did in the example from paleobiology above.

Error statistics provides a third way of conceptualizing evidence for a model, in this case in terms of severe testing: "Data $\boldsymbol{x_0}$ in test T provide good evidence for inferring [hypothesis] H (just) to the extent that H passes severely with $\boldsymbol{x_0}$, i.e., to the extent that H would (very probably) not have survived the test so well were H false" (Mayo and Spanos 2006, 328). Instead of using the AIC to quantify the relative "closeness" of models to the true distribution, severity is based on the probability of making an incorrect choice between models during the selection procedure. In the simple setting of Neyman-Pearson testing, there are two types of error to consider: a false positive (type 1), where the null model is incorrectly rejected in favor of the alternative; and a false negative (type 2), where the alternative model is true but the null fails to be rejected. A severe test should minimally ensure that the probability of a false positive is low, but it ideally should also have a low probability of false negatives.

Severe testing, Akaike weights, and ΔAIC are not equivalent. For example, Spanos (2010) gives a simple example of selecting between two alternative regression models. In this setting, the decision procedure of selecting the model with the lowest AIC score can be translated into an F-test where we can compute the corresponding error probabilities analytically. The Type 1 error probability turns out to be 0.18, considerably higher than conventional standard of 0.05 or smaller. The bigger problem, though, is that the same procedure (e.g., choose the model with the lowest AIC) will translate into potentially very different Type 1 and 2 errors if we change the model set, estimated model parameters, and sample size. While it's possible to construct a local mapping between model selection decision thresholds relative to a particular target system, associated set of data, and candidate models (e.g., Bandyopadhyay and Boik 1999), no general mathematical description of this mapping exists across all such contexts. Moreover, the availability of such a general mapping would not eliminate actual differences in practice, as scientists using alternative procedures may not be able to agree on a single appropriate threshold. For example, Hunt et al. (2015) used a ratio threshold for Akaike weights that does not translate into conventionally acceptable alpha levels for error statistics (Cullan et al. 2020). What matters is that scientists' disagreement about the appropriate effective threshold (even given a local mapping between AIC, BIC, etc.) can lead to conflicting conclusions about the best model.

5. Containing the true distribution

The applicability of the model selection methods we've described depends in part on a crucial fact about the set of candidate models: among all the distributions these models contain, is one of them the true one? Although this question is easy to state informally, giving a precise answer involves some considerable complications. In this section, we distinguish three senses in which a model "includes" or "contains" a distribution:

1. A model contains multiple probability distributions in an abstract, mathematical sense
2. A model contains the true distribution
3. An infinite sequence of models contains a sequence of distributions converging on the true distribution

The second and third senses in particular will prove crucial to articulating how prior beliefs about the metaphysics of a phenomenon can inform the choice of a model selection method.

Meaning 1 (abstract mathematical): a parametric probability model typically contains infinitely many probability distributions, indexed by unique combinations of parameter values. The distribution and model in this context are both abstract mathematical objects and are related by membership in the standard set theoretic sense. As such, they lack a semantic interpretation as referring to data-generating processes in the world.

Meaning 2 (finite model): This sense of a model "containing" a probability distribution requires us to link the model as a mathematical object to a data-generating process in the world. We will assume here that we have access to an arbitrarily large dataset drawn from independent and identically distributed (i.i.d.) random variables with continuous values. More concretely, we could exactly replicate an experimental setup as many times as we want to generate a dataset where each observation is not influenced by any others (independence), comes from the same probability distribution (identical distribution), and is represented by a real number (continuous-valued).

In this context, we can define an empirical cumulative distribution using the observed data and examine how it behaves as the sample goes to infinity:

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

Given a sample size of n observations, denoted $X_1 \ldots X_n$, $\widehat{F}_n(x)$ counts up the number of observations with values less than or equal to x and assigns them each a probability of $\frac{1}{n}$. The function $I(X_i \leq x)$ inside the summation is an indicator variable that returns the value 1 when $X_i \leq x$. The function $\widehat{F}_n(x)$ therefore estimates how fast or slow the observed values stack up as a function of x. Given our assumption that the random variables are i.i.d. and continuous-valued, we know that $\widehat{F}_n(x)$ converges to the true cumulative distribution function $F(x)$ as n goes to infinity.

We can now define the true distribution to be the one on which the empirical cumulative distribution converges (in probability) as the sample size goes to infinity. To say that a finite set of models includes the true distribution is then to assert that some distribution contained by one of the models has an identical cumulative distribution to the one on which the empirical cumulative distribution function converges. One way to operationalize the notion of identical distributions is the KL divergence we introduced in Section 3. The model-specified distribution is then identical to the true distribution if the KL divergence between their cumulative distributions is zero. Qualitatively, this means they are perfectly interchangeable in terms of their predictive information content.

Meaning 3 (infinite models): In this case, we consider what it means for a countably infinite set of models to include a probability distribution. We will be interested in cases where we have a countably infinite sequence of models, such as regression polynomials or time series models, but the true distribution is not included in any finite subset of this sequence. Nonetheless, there can be a sense in which the infinite sequence of models comes "arbitrarily close" to the true distribution, much as a sequence of rational numbers can come arbitrarily close to an irrational real number.

Let f be the true probability distribution and M_i be a countably infinite sequence of models for i from 1 to positive infinity. Then we can say that f is contained in the infinite model set M_i if and only if for any epsilon ϵ of KL divergence, there exists a distribution g in M_i such that $KL(f,g) \leq \epsilon$. We can also define what it means for a distribution to be contained in a finite subset of the M_i. Consider the limit of the empirical cumulative distribution as defined above. This limiting distribution F is contained in a finite subset of the M_i if and only if there exists a positive integer N such that $KL(F,G_N) = 0$ for some cumulative distribution G_N of a probability distribution in model M_N. In the case where no finite subset is adequate, the minimum KL divergence for all distributions in M_1 to M_N may get arbitrarily small as N increases to positive infinity, but it never equals zero.

6. Views on the candidate model set

In this section, we articulate four distinct views a scientist may take on the adequacy of the candidate model set with respect to including the true distribution. We limit our discussion to the typical scenario where a scientist aims to evaluate a finite set of candidate parametric models. We then show how these positions figure in justifying alternative understandings of the purpose of model selection and views on the correct choice of method.

Building on the results of the previous section, we can distinguish four positions regarding the adequacy of a candidate model set as follows:

1. The true distribution is contained in the finite set of candidate models.
2. The true distribution is not contained in the finite candidate set, but there exists another finite model set that does include the true distribution.
3. The true distribution is not contained in any finite model set, but it is contained in a specifiable infinite model set.
4. The true distribution is not contained in any finite or infinite model set.

In trying to determine which view is correct for some situation, data can provide relevant empirical evidence for one option over the others. Misspecification testing, for instance, provides a basis for rejecting parametric models whose modeling assumptions are not well supported by the data. In the paleobiology example, Voje (2018) defined null hypothesis tests based on whether the data is consistent with fixed means, independent error residuals, and constant variance. If the test statistic is improbable according to the null model, this supports the claim the true distribution is not contained in the null model.

For information-theoretic methods, a new tool called the parametricness index is able to quantify the extent to which the data single out one model as best (Liu and Yang 2011; Ding et al. 2018a). Given a countably infinite sequence of models, the parametricness index looks at how fast the best model improves with increasing sample size relative to models with one fewer parameter. Liu and Yang (2011) refer to a scenario as non-parametric if the best model continues to grow in complexity indefinitely as the sample size grows. In this case, they show that the parametricness index converges to one in probability. Alternatively, they refer to a scenario as parametric if the best model stabilizes even as the sample size grows to infinity. In this case, they show the parametricness index increases asymptotically to positive infinity, since the performance gap between the best and next-best model of smaller size continues to expand as the sample size increases. However, Liu and Yang treat cases as "effectively parametric" where one model is a stable best choice even though the true distribution is outside the candidate model. The parametricness index therefore doesn't reliably discriminate between all four positions we've articulated.

The data may also fall short of supporting any definitive answer. Common scientific standards of evidence in this regard are conventional thresholds applied to a continuum. A misspecification test, for example, might return a p-value larger than the chosen significance threshold, e.g., a result of $p = 0.06$ when the chosen threshold was 0.05. Alternatively, the parametricness index might show intermediate levels of fluctuation in the models chosen by the BIC.[3] These indeterminate situations leave room for individual judgment. Is the evidence for model adequacy strong enough to proceed with the goal of selecting the true model, or is the appropriate response to respecify the model set? Are the data sufficient to support "effective" convergence on the true model, or is it better to assume no model even approximately contains the true distribution? This gray zone permits researchers' views on the relation of the candidate model set to the true process to influence how they quantify evidence.

7. How prior metaphysical beliefs enter in

[3] The magnitude of the index also depends on two parameters, λ_n and d, whose values depend on expert judgment about how "close" the parametric and nonparametric scenarios being compared are (Liu and Yang 2011, 2084).

What we need to show now is how prior metaphysical beliefs can reasonably inform a scientist's choice of method. In particular, we are interested in views about whether nature is ultimately simple or complex as expressed in terms of what can be represented using a finite or infinite set of probability models. This section presents four contrasting positions on this point and shows how they map onto alternative methodological approaches using examples from the paleobiology case. The first two positions are grounded on the shared belief that nature is simple, but they differ more narrowly on the adequacy of the candidate model set at hand. The second two are both based on nature being complex but differ on the possibility of approximating this complexity.

1. *Strong adequacy:* The candidate finite model set contains the true distribution, or some coarse-grained function thereof (Position #1). In the paleobiology case, this approach treats the models as representing accurate coarse-grained descriptions of the actual probabilistic structure of the evolutionary processes driving the observed trait changes. Selecting the true finite model is possible. Misspecification testing is relevant and a high parametricness index is necessary given sufficient data. The true model is also theoretically optimal for predictive accuracy and simulating data, assuming sufficient data are available.
2. *Modest adequacy:* The candidate finite model set doesn't contain the true distribution (or a coarse-grained function of it), but it is possible to find another finite model that does (Position #2). In the paleobiology case, this approach treats the candidate models as adequate for classifying and simulating data as first-order approximations of the actual evolutionary processes at work. Selecting an approximately true finite model is possible if the differences between the model and true distribution aren't relevant for further use in testing hypotheses. Misspecification testing is useful to help expand the candidate model set as needed in attempting to include the true model. Assuming one model is uniquely closest to the true distribution in KL divergence, a high parametricness index is necessary given sufficient data. The model minimizing KL divergence will be asymptotically optimal for minimizing predictive error, but this model may or may not be best for simulating data. Multiple interpretations of the "correct" model are therefore possible, given that the true distribution is not included in the candidate set.
3. *Modest instrumentalism:* No finite model set exists that contains the true distribution, but it is possible to approximate the true distribution as the limiting case of an infinite set of candidate models (Position #3). In the paleobiology case, this approach treats the models as adequate for classifying and simulating first-order trends in the time series. Identifying a unique, approximately true finite model is not possible, since the best-fitting model will continue to change as the sample size goes to infinity. While misspecification testing will not uncover a true finite model, it is still relevant for showing that actual divergence between the best model and the true distribution is small enough for practical purposes. The parametricness index is expected to converge to 1, since the best-fitting model will not stabilize with more data. As with the modest adequacy view, multiple interpretations of "correct" are possible and will be optimal for different statistical aims.
4. *Strong instrumentalism:* No finite or infinite model set contains the true distribution (Position #4). In the paleobiology case, this approach treats the models as discriminating among different qualitative types of patterns in the time series but rejects their use for parametric simulation. Selecting a unique, approximately true finite model may or may not be possible, depending on the whether the infinite model set progressively approaches

the true distribution as model complexity increases. For the same reason, the best-fitting model may or may not continue to change as the sample size goes to infinity. Misspecification testing is not relevant, and the behavior of the parametricness index will depend on the relation of the true distribution to the infinite model set. As with modest adequacy and instrumentalism, many interpretations of "correct" are possible and will be optimal for different statistical aims.

Note that all of these approaches can agree on the general existence of statistical patterns on spatial and temporal scales relevant to human interests. However, they disagree about whether and in what sense these patterns constitute strong evidence for a model's truth. Both the modest adequacy and instrumentalist approaches, for example, are consistent with the best model providing an arbitrarily close fit to the true distribution as the sample size increases, but they differ on whether any finite parametric model we construct could ever be true.

A quote from statistician David Cox nicely motivates both the strong and modest adequacy views: "For substantive purposes it is usually desirable that the model can be used fairly directly to simulate data. The essential idea is that if the investigator cannot use the model directly to simulate artificial data, how can 'Nature' have used anything like that method to generate real data?" (Cox 1990, 172). Parametric simulation is a common use for probability models,[4] but its validity is doubtful unless we believe the model at least approximates the true distribution in relevant respects. For example, just as a finite Taylor series can approximate an infinite dimensional curve, so a finite regression or time series model may approximate an infinite dimensional distribution well enough for the purpose at hand. The strong and modest adequacy approaches in this regard both share the assumption that some finite model can be stated that contains the true distribution. They differ, though, about whether the true distribution is included in the model set under consideration by the researcher, and hence whether the correct model in the candidate set exactly represents the data generating process (strong view) or only approximately (modest view).

We can find a contrasting instrumentalist perspective in the work of ecologist David Anderson: "As a crutch, we can think of full reality as infinite dimensional; however, full reality is unlikely to be parameterized. Parameters are a construct useful in many science contexts, but many parts of full reality are not even parameterized. A 'good' model successfully separates information from 'noise' or noninformation in the data, but never fully represents truth" (Anderson 2008, 101). Aiming to identify the true finite model isn't possible on his view, but there may still be a valid sense in which the correct model in terms of predictive accuracy is also an approximation to the true distribution. Ultimately, however, if "full reality is unlikely to be parameterized," then a model could only be correct in the sense of being the best from the candidate set at predicting new observations drawn from the same process.

The strong and modest instrumentalist approaches agree in this regard on rejecting any one finite model as an adequate representation of the true distribution. They correspond to a view of models as "mere" tools for prediction with no necessary commitment to their representational accuracy. We can understand the strong approach as rejecting the representational adequacy of parametric models wholesale — reality is too complex for what can be expressed in even an infinite set of parametric models. The modest version instead endorses the position that the true distribution exceeds any finite parametric model, though it remains within our grasp as a limit expressible by an infinite model set.

[4] Indeed, it figures centrally in the method of Cullan et al. (2020) for the paleobiology case.

In conclusion, when empirical uncertainty remains about which of these methodological approaches applies in a given situation, a researcher's prior beliefs about the simplicity or complexity of nature relative to the available models can provide a relevant reason for choosing one model selection method over another based on its applicability according to the four positions we described.

8. The problem is unlikely to go away soon

No single model selection method serves equally well for all four views we identified on the adequacy of the candidate model set. In this section, we show why this context-sensitivity of methods — and hence the general scope of the problem of unequal evidence — is likely to persist in scientific practice. The central issue here is that the adequacy of the candidate models for a particular natural system also has implications for which statistical goals are sensible, and the model selection methods we've discussed also differ in their optimality for these alternative aims. Key mathematical results from statistical theory block the hope of finding one universal concept of evidence suitable for all purposes, regardless of candidate model adequacy. While science can always surprise us, it therefore appears there are good reasons to expect that a multiplicity of methods will persist whose application depends, at least in part, on one's assessment of the adequacy the candidate model set.

In particular, some general mathematical theorems prove necessary tradeoffs between the AIC and BIC for the goals of identifying the true model and maximizing predictive accuracy. A common way to quantify predictive accuracy is the Mean Squared Error (MSE) between the true parameter θ and estimated parameter θ_b produced in the model selection process,

$$MSE(\theta,\theta_b) = E_\theta(\theta_b - \theta)^2 = Var_\theta(\theta_b) + Bias^2.$$

The subscript θ for the expectation and variance terms indicates they are computed using the true distribution.

We will need to distinguish two ways of evaluating how a procedure's predictive risk behaves with growing sample size (Yang 2007). First, we can measure how fast the procedure's MSE decays for a particular true distribution and compare that rate to the performance of the best possible procedure relative to that true distribution. Yang (2007) calls this the "point-wise" behavior of the procedure because it is relativized to a single true distribution (i.e. one point in the space of possible distributions). For example, in a simple scenario where we have to choose between two regression models, $Y = \mu + \beta_1 X + \epsilon$ and $Y = \mu + \epsilon$, the space of possible distributions is indexed by (μ,β_1). The MSEs of the AIC and BIC are a function of the value of β_1 in the true distribution, since for values of β_1 close to zero the two criteria will show different tendencies to pick the larger or simpler model. When the true distribution is in the finite model set (View 1 from Section 6), the BIC is point-wise optimal in that for any particular true distribution, the BIC's MSE will decay as fast as the best procedure in that scenario. However, when the true distribution is not in the finite candidate set (Views 2–4), the AIC is point-wise optimal while the BIC is not.

A second way to evaluate predictive risk is to consider how slowly a procedure's MSE decays in the worst possible case, i.e. when the distribution with the largest possible MSE for that method happens to be true. We then compare this worst-case performance to the results of a procedure chosen to have the best worst-case result across all possible distributions (commonly known as a minimax procedure or minimax estimator). Yang (2007) calls this the "minimax" behavior of a procedure because we evaluate it relative to the whole space of possible

distributions and set of procedures. The AIC is minimax optimal when the true distribution is not included in any finite model set (Views 3–4), but the BIC is not minimax optimal in any scenario.

The lingering question is whether it's possible to design a single criterion that delivers both point-wise and minimax optimality regardless of which view about the candidate models is correct. This was shown to be impossible mathematically for deterministic criteria (i.e. whose complexity penalties are not random variables), but statisticians more recently have been pursuing an adaptive approach that blends the AIC and BIC stochastically (see Ding et al. 2018a for references). Yang and collaborators, for example, have developed a new approach called the Bridge Criterion that uses the BIC when it selects a single, stable model across resampling of the data and the AIC otherwise (Ding et al. 2018a). However, the Bridge Criterion only delivers point-wise optimality and not minimax optimality.[5]

Despite the merits of combining the AIC and BIC adaptively based on the data, a universal best criterion delivering all forms of optimality appears to be ruled out on mathematical grounds. If one adopts either instrumentalist view of the candidate models and prioritizes minimax optimality for prediction, then the AIC is still preferable to the BIC and Bridge Criterion. There's also no definitive rule for choosing between the BIC or AIC in scenarios corresponding to Positions 2 or 3 when one is interested in picking an approximately true model, especially given that an adequate approximation is relative to purpose (e.g., precisely what features of the data generating process one wants to identify). Finally, if one prioritizes controlling error probabilities, then none of the AIC, BIC, or Bridge Criterion deliver reliable performance.

8. Conclusion

The problem of unequal evidence represents a novel form of underdetermination with broad applicability to statistical model selection practices across the sciences. In the paleobiology case we considered, Gould and Eldgredge originally presented punctuated equilibrium as a gestalt change in what paleobiologists should see as signal versus noise in gaps in fossil data (Sepkoski 2012). Introducing advanced statistical modeling was supposed to bring new objectivity to the classification of fossil trends, since individual scientists frequently saw different processes at work behind the same data. This goal is undercut if model selection results turn out to depend on scientists' prior beliefs about the natural system of interest. Our results open a new direction for investigating how complexity poses epistemic problems for science without necessarily invoking unique historical events, as has been a major focus of prior work in philosophy of the historical sciences (e.g., Tucker 1998, Currie 2018).

Indeed, the model selection methods we've discussed are not limited to the historical sciences, and they only represent a small sample of the variety currently in use. This suggests the relevance of underdetermination for scientific practice is severely restricted if we believe it only matters when multiple theories are equally well supported by evidence. Scientists across many fields, including economics, ecology, psychology, and statistics, are developing their own typologies informing model selection practice based on the relation of candidate models to reality (Wagenmakers and Farrell 2004; Yang 2007; Clarke et al.. 2013; Aho et al. 2014; Ding et al. 2018b). This proliferation of typologies has the potential to magnify the problem, and an

[5] Asymptotic behaviors are not definitive for real-world cases, moreover. The performance ranking of information criteria can shift dramatically as the true parameters vary (Yang 2007).

important direction would be to analyze how and when statistical methods can be made equivalent across the metaphysical positions we articulated above.

While the problem of unequal evidence represents a novel and important form of underdetermination in practice, it does not support a strong new argument for global anti-realism. Note that the problem of unequal evidence arises precisely because a range of positive and negative views about the adequacy of the candidate models are defensible in a given context. This is hard to reconcile, for example, with Stanford's (2006) globally anti-realist stance based the existence of unconceived alternatives. Nonetheless, the problem of unequal evidence represents an actual rather than hypothetical challenge for scientific practice, and both its generality and persistence are grounded in mathematically deep results from statistical theory.

References

Aho, Ken, DeWayne Derryberry, and Teri Peterson. 2014. "Model Selection for Ecologists: the Worldviews of AIC and BIC." *Ecology* 95 (3): 631–36.

Akaike, Hirotugu. 1981. "Likelihood of a Model and Information Criteria." *Journal of Econometrics* 16 (1): 3–14.

Anderson, David. 2008. *Model Based Inference in the Life Sciences*. New York: Springer.

Bandyopadhyay, Prasanta, and Robert Boik. 1999. "The Curve Fitting Problem: A Bayesian Rejoinder." *Philosophy of Science* 66: S390–402.

Bokulich, Alisa. 2018. "Using Models to Correct Data." *Synthese* 53 (6): 1211–22.

Burnham, Kenneth, and David Anderson. 2002. *Model Selection and Inference*. New York: Springer-Verlag.

Chakrabarti and Ghosh. 2011. "AIC, BIC and Recent Advances in Model Selection." In Bandyopadhyay and Forster eds. Handbook of Philosophy of Statistics.

Clarke, Bertrand, Jennifer Clarke, and Chi Wai Yu. 2013. "Statistical Problem Classes and Their Links to Information Theory." *Econometric Reviews* 33 (1-4): 337–71.

Cleland, Carol. 2001. "Historical Science, Experimental Science, and the Scientific Method." *Geology* 29 (11): 987–90.

Cox, D. 1990. "Role of Models in Statistical Analysis." *Statistical Science* 5 (2): 169–74.

Cullan, Michael, Scott Lidgard, and Beckett Sterner. 2019. "Controlling the Error Probabilities of Model Selection Information Criteria Using Bootstrapping." *Journal of Applied Statistics* 7: 1–17.

Currie, Adrian. 2018. *Rock, Bone, and Ruin*. Cambridge: MIT Press.

Currie, Adrian. 2019. "Paleobiology and Philosophy." *Biology & Philosophy* 34 (2): 31.

Dietrich, Michael, and Robert Skipper Jr. 2007. "Manipulating Underdetermination in Scientific Controversy." *Perspectives on Science* 15 (3): 295–326.

Ding, Jie, Vahid Tarokh, and Yuhong Yang. 2018a. "Bridging AIC and BIC." *IEEE Transactions on Information Theory* 64 (6): 4024–43.

Ding, Jie, Vahid Tarokh, and Yuhong Yang. 2018b. "Model Selection Techniques." *IEEE Signal Processing Magazine* 35 (6): 16–34.

Dorling, Jon. 1979. "Bayesian Personalism, the Methodology of Scientific Research Programmes, and Duhem's Problem." *Studies in History and Philosophy of Science* 10 (3): 177–87.

Eldredge, Niles, and Stephen Jay Gould. 1972. "Punctuated Equilibria." In *Models in Paleobiology*, edited by T.J.M. Schopf, 82–115. San Francisco.

Forster, Malcolm. 2002. "The New Science of Simplicity." In *Simplicity, Inference and Modelling*. Cambridge.

Gingerich, Philip. 1985. "Species in the Fossil Record." *Paleobiology* 11 (1): 27–41.

Gould, Stephen Jay, and Niles Eldredge. 1977. "Punctuated Equilibria: The Tempo and Mode of Evolution Reconsidered." *Paleobiology* 3 (2): 115–51.

Halsey, Lewis G. 2019. "The Reign of the P-Value Is over" *Biology Letters* 15 (5): 20190174.

Hopkins, Melanie, and Scott Lidgard. 2012. "Evolutionary Mode Routinely Varies Among Morphological Traits Within Fossil Species Lineages." *Proceedings of the National Academy of Sciences* 109 (50): 20520–25.

Hunt, Gene. 2006. "Fitting and Comparing Models of Phyletic Evolution." *Paleobiology* 32 (4): 578–601.

Hunt, Gene. 2007. "The Relative Importance of Directional Change, Random Walks, and Stasis in the Evolution of Fossil Lineages." *Proceedings of the National Academy of Sciences* 104 (47): 18404–8.
Hunt, Gene, Melanie Hopkins, and Scott Lidgard. 2015. "Simple Versus Complex Models of Trait Evolution and Stasis as a Response to Environmental Change." *Proceedings of the National Academy of Sciences* 112 (16): 4885–90.
Kieseppä, I.A. 2001. "Statistical Model Selection Criteria and Bayesianism." *Philosophy of Science* 68 (3): S141–52.
Kukla, Andre. 1996. "Does Every Theory Have Empirically Equivalent Rivals?" *Erkenntnis* 44 (2): 137–66.
Laudan, Larry, and Jarrett Leplin. 1991. "Empirical Equivalence and Underdetermination." *The Journal of Philosophy* 88 (9): 449.
Lidgard, Scott, and Melanie Hopkins. 2015. "Stasis." In *Oxford Bibliographies in Evolutionary Biology*, edited by Jonathan Losos, 1–31.
Liu, Wei, and Yuhong Yang. 2011. "Parametric or Nonparametric? A Parametricness Index for Model Selection." *Annals of Statistics* 39 (4): 2074–2102.
Longino, Helen. 1990. *Science as Social Knowledge*. Princeton: Princeton University Press.
Mayo, Deborah, and Aris Spanos. 2006. "Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction." *British Journal for the Philosophy of Science* 57 (2): 323–57.
Okasha, Samir. 2002. "Underdetermination, Holism and the Theory/Data Distinction." *The Philosophical Quarterly* 52 (208): 303–19.
Roopnarine, Peter. 2001. "The Description and Classification of Evolutionary Mode." *Paleobiology* 27 (3): 446–65.
Royall, Richard. 1997. *Statistical Evidence*. New York: CRC Press.
Sepkoski, David. 2012. *Rereading the Fossil Record*. Chicago: University of Chicago Press.
Sheets, David, and Charles Mitchell. 2001. "Uncorrelated Change Produces the Apparent Dependence of Evolutionary Rate on Interval." *Paleobiology* 27 (3): 429–45.
Simpson, George Gaylord. 1944. *Tempo and Mode in Evolution*. New York: Columbia University Press.
Sober, Elliott. 2004. "Likelihood, Model Selection, and the Duhem-Quine Problem." *The Journal of Philosophy* 101 (5): 221–41.
Spanos, Aris. 2010. "Akaike-Type Criteria and the Reliability of Inference: Model Selection Versus Statistical Model Specification." *Journal of Econometrics* 158 (2): 204–20.
Sprenger, Jan. 2012. "The Role of Bayesian Philosophy Within Bayesian Model Selection." *European Journal for Philosophy of Science* 3 (1): 101–14.
Stanford, Kyle. 2006. *Exceeding Our Grasp*. Oxford: Oxford University Press.
Stanford, Kyle. 2017. "Underdetermination of Scientific Theory." *The Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/scientific-underdetermination/.
Turner, Derek. 2011. *Paleontology: A Philosophical Introduction*. Cambridge: Cambridge University Press.
Tucker, Aviezer. 1998. "Unique Events: The Underdetermination of Explanation." *Erkenntnis* 48 (1): 61–83.
Voje, Kjetil. 2018. "Assessing Adequacy of Models of Phyletic Evolution in the Fossil Record." *Methods in Ecology and Evolution* 106 (Suppl. 2): 19699.

Wagenmakers, Eric-Jan. 2007. "A Practical Solution to the Pervasive Problems of P Values." *Psychonomic Bulletin & Review* 14 (5): 779–804.
Wagenmakers, Eric-Jan, and Simon Farrell. 2004. "AIC Model Selection Using Akaike Weights." *Psychonomic Bulletin & Review* 11 (1): 192–96.
Wylie, Caitlin. 2019. "Overcoming the Underdetermination of Specimens." *Biology & Philosophy* 34 (2): 24.
Yang, Yuhong. 2007. "Prediction/Estimation with Simple Linear Models." *Econometric Theory* 23 (1): 1–36.

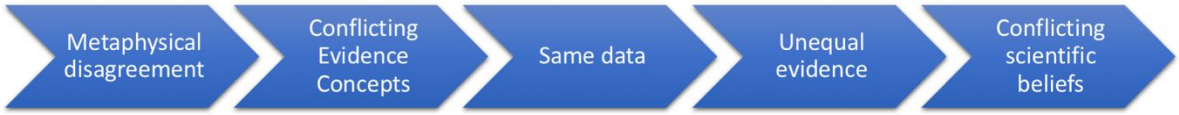

Figure 1: Underdetermination arises in cases where prior metaphysical differences lead scientists to apply alternative conceptions of evidence and arrive at conflicting conclusions even for the same data and candidate models.

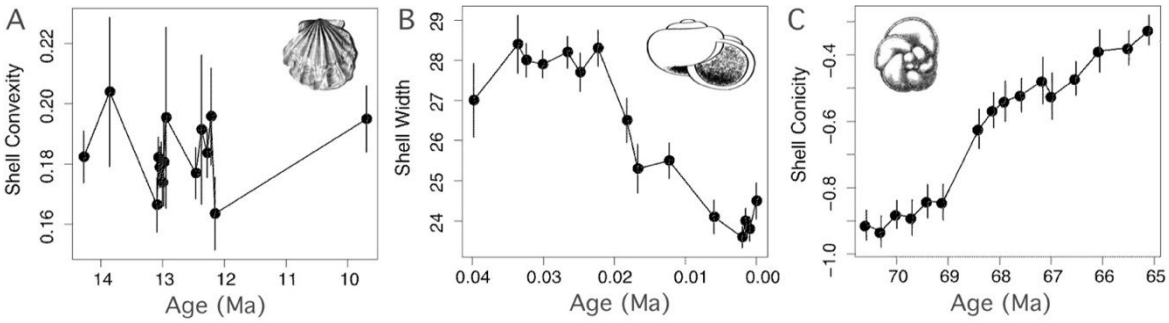

Figure 2: Sequences of measured traits from fossil populations sampled in successive sedimentary strata. The mean and variance of sampled traits provide an indication of the phenotypic step size, used in fitting the models. (A) Stasis: shell convexity (a shape measure) in the bivalve *Chesapecten*. (B) Random walk: shell width in the *Mandarina*, a land snail. Gradualism: shell shape in the foraminifera *Contusotruncana*. Figure modified from (Hunt 2007).