doi:10.1111/evo.14213

How should functional relationships be evaluated using phylogenetic comparative methods? A case study using metabolic rate and body temperature

Josef C. Uyeda,[1,2] Nicholas Bone,[1] Sean McHugh,[1] Jonathan Rolland,[3,4,5] and Matthew W. Pennell[4,5]

[1]*Department of Biological Sciences, Virginia Tech, Blacksburg, Virginia 24061*

[2]*E-mail: juyeda@vt.edu*

[3]*Department of Computational Biology, University of Lausanne, Quartier Sorge, Lausanne 1015, Switzerland*

[4]*Biodiversity Research Centre, University of British Columbia, Vancouver, BC V6T 1Z4, Canada*

[5]*Department of Zoology, University of British Columbia, Vancouver, BC V6T 1Z4, Canada*

Received May 28, 2020

Accepted January 20, 2021

Phylogenetic comparative methods are often used to test functional relationships between traits. However, million-year macroevolutionary observational datasets cannot definitively prove causal links between traits—correlation does not equal causation and experimental manipulation over such timescales is impossible. Although this caveat is widely understood, it is less appreciated that different phylogenetic approaches imply different causal assumptions about the functional relationships of traits. To make meaningful inferences, it is critical that our statistical methods make biologically reasonable assumptions. Here we illustrate the importance of causal reasoning in comparative biology by examining a recent study by Avaria-Llautureo et al (2019). that tested for the evolutionary coupling of metabolic rate and body temperature across endotherms and found that these traits were unlinked through evolutionary time and that body temperatures were, on average, higher in the early Cenozoic than they are today. We argue that the causal assumptions embedded into their models made it impossible for them to test the relevant functional and evolutionary hypotheses. We reanalyze their data using more biologically appropriate models and find support for the exact opposite conclusions, corroborating previous evidence from physiology and paleontology. We highlight the vital need for causal thinking, even when experiments are impossible.

KEY WORDS: **Macroevolution, models/simulations, phylogenetics, physiology.**

The evolutionary causes and consequences of endothermy have fascinated biologists for decades. But this is naturally a difficult thing to study: the evolutionary events we are interested in occurred over the course of hundreds of millions of years. At this scale, direct tests of causal hypotheses are impossible. Nonetheless by piecing together multiple lines of evidence, including physiological experiments, mathematical theory, and the paleontological record, we have gained a rich understanding of the relationships between metabolic rates, body mass, and temperature and how they might have evolved (Grigg et al. 2004; Clarke et al. 2010; Lovegrove 2017), even as mysteries remain.

Macroevolutionary comparative analyses complement these approaches; by leveraging interspecific data in a phylogenetic framework, we can potentially gain new insights into how physiological processes (White et al. 2009; Uyeda et al. 2017; White et al. 2019) and scaling relationships, more generally, evolve (Hansen and Bartoszek 2012; Pélabon et al. 2014; Voje et al. 2014). And although developments in this area are tremendously exciting, it is critical to keep in mind that inferences drawn from phylogenetic statistical methods, just like many other statistical methods, have embedded, and often implicit, causal assumptions (Uyeda et al. 2018)—and one must always keep in mind the

© 2021 The Authors. *Evolution* © 2021 The Society for the Study of Evolution.

motivating theory for how variables are functionally related to underlying causal processes and to assess whether we are indeed measuring the right thing (Houle et al. 2011). No matter how elegant the method, if it is not appropriate for the theoretical context, it is not going to tell you anything meaningful and even worse, will likely be misleading. We argue that inferences drawn in all comparative analyses should be carefully evaluated in light of their relationship to hypothesized causal relationships, and skeptically evaluated in light of the nature of our necessarily simplistic models.

Our motivating case study is a recent study by Avaria-Llautureo et al. (2019), which presents evidence from phylogenetic models that provides a new perspective on the major features of endothermic evolution, with the provocative conclusion that metabolic rate and body temperature are decoupled—a conclusion largely at odds with previous studies (Clarke et al. 2010). To support this conclusion, Avaria-Llautureo et al. (2019) use statistical phylogenetic comparative methods inspired by similar tests for convergence in selective pressures as demonstrated by convergent evolutionary rates, which are commonly used in molecular evolution (Chikina et al. 2016; Sackton et al. 2019). Although superficially a reasonable analogy, we argue in this note that the novel conclusions of Avaria-Llautureo et al. (2019) are consequences of a failure to use models that reflect the processes that are important in the evolution of endothermic traits—and phenotypic traits more generally. Using process-based models and simulating different causal scenarios, we demonstrate how easily estimates from complex statistical models can become dissociated from the underlying biological hypotheses. Although our assumptions about cause and effect and/or biological constraints may be incorrect (e.g., if past processes are fundamentally different from extant processes), we argue it is far superior to make these assumptions explicit and part of the process of making inference from comparative data.

Coupling of States or Rates?

It is well established that the basal metabolic rate (*BMR*) of an organism is a function of its body mass M, its internal body temperature T_b, and the temperature in which it lives (ambient temperature; T_a) (Clarke et al. 2010, and references therein). These relationships form the basis of several ecological theories (e.g., the *Metabolic Theory of Ecology*; Brown et al. 2004) and, more broadly, are used for ensuring that *BMR* is measured in comparable way across individuals, species, and studies. That is to say, *BMR*, M, and T_b are known to be functionally coupled. For this reason, the conclusion in Avaria-Llautureo et al. (2019) that there is no evidence of evolutionary or functional coupling between *BMR* and T_b may appear surprising.

However, here we look closer at how functional coupling can be defined and how it relates to the evolutionary questions being asked. In Avaria-Llautureo et al. (2019), they open their paper discussing this expected covariation of the values of *BMR*, T_b, and T_a, but then propose defining "coupling" to mean the correlation of evolutionary *rates* rather than *states*. This shift is subtle but important. Indeed, consistent with previous analyses of the same data, Avaria-Llautureo et al. (2019) do find evidence of a positive association between T_b and *BMR* (Clarke et al. 2010), and a negative association between T_b and T_a (Clarke et al. 2010). We use this study as a motivating example to ask two questions that are important to consider when choosing from the variety of comparative methods available to test associations between traits. First, are the coupling of rates and the coupling of states simply alternative ways of measuring the same thing, or do they measure different things altogether? And second, if they are different, how does one choose a method that meaningfully corresponds to the theoretical context under study?

We illustrate our points in two ways. We first simulate trait data under three alternative and biologically plausible causal scenarios (depicted in Fig. 1 and described next) on the phylogeny of mammals used by Avaria-Llautureo et al. (2019) and examine the resulting covariance structure in both the states and the rates. We then use the simulated scenarios as reference points for interpreting our reanalysis of the original empirical data of Avaria-Llautureo et al. (2019).

- *Scenario 1* is a multivariate Brownian motion (mvBM; Felsenstein 1985) model. Note that this model generates data with the same distribution as a phylogenetic generalized least squares (PGLS) model, though mvBM and PGLS depict different causal scenarios that are not identifiable in extant-only data (Blomberg et al. 2012). For simplicity, we simulate *BMR* and T_b under mvBM where the two traits are correlated ($\rho = 0.8$) but where there is no variation in rates across different lineages (meaning there is no variation to detect a branch-wise correlation in rates).
- In *Scenario 2*, we simulate uncorrelated states of T_b and *BMR* by changing the correlation in the evolutionary rate matrix to $\rho = 0$. However, we introduce correlations in branch-specific rates by drawing shared branch scalars for both traits from a Γ distribution (as is typical when capturing rate variation in phylogenetic studies) with shape parameter set to 1.1 and a scale parameter set to 1. This results in collinear evolutionary rates for both traits, but no correlation in states. In other words, even though the traits are not correlated in any way, branches on which there is rapid evolution in one trait are also branches in which there is rapid evolution in the other trait. Perhaps surprisingly, we consider the most straightforward process consistent with this type of "rate-coupling" to be one in which there is

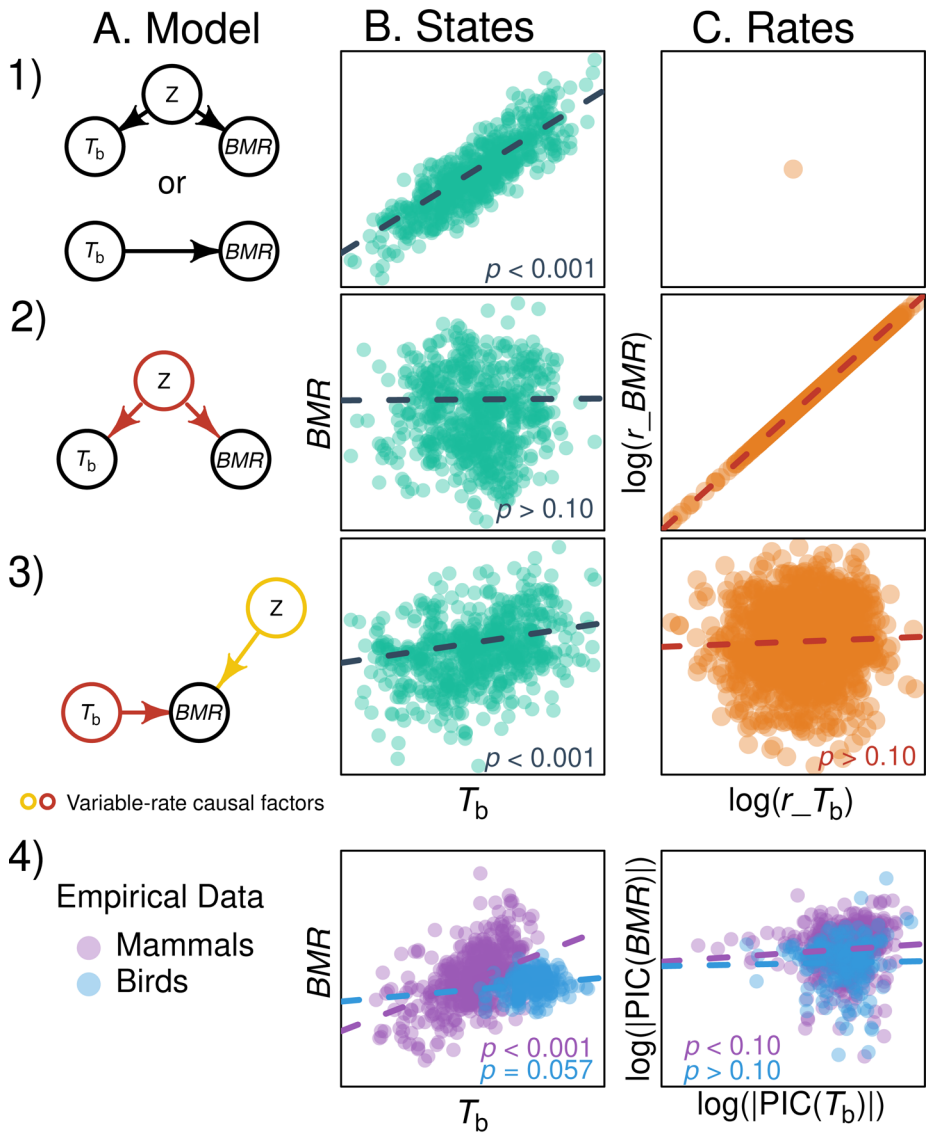

Figure 1. Simulated (Rows 1–3) and Empirical (Row 4) relationships between T_b and *BMR* on the mammalian phylogeny used in Avaria-Llautureo et al. (2019). (Row 1; *Scenario 1*): If T_b is direct cause of *BMR* or if both evolve under a common cause, *Z*, in a correlated fashion (1A), then we observe a correlation in state (1B; dotted line = PGLS regression) despite an absence of rate variation (1C) (cf. Avaria-Llautureo et al. 2019, Fig. 1a). (Row 2; *Scenario 2*): Multivariate Brownian motion (2A) where T_b and *BMR* are uncorrelated in state (2B) but share the same source of rate variation (2C). Shared rate variation indicated by colored nodes/arrows. This is the scenario that Avaria-Llautureo et al. (2019) define as evolutionary "coupling" (cf. Avaria-Llautureo et al. 2019, Fig. 1b), despite the fact that it can occur absent a functional relationship between traits (e.g., evolution by genetic drift with *Z* representing time-varying effective population size). (Row 3; *Scenario 3*) T_b is a direct cause of *BMR* and has rate variation, but *Z* is also a cause of *BMR* with independent rate variation (3A). Rates in *Z* are 20× the rates in T_b. This results in highly significantly correlated states (3B), but nonsignificantly correlated rates (3C) (cf. Avaria-Llautureo et al. 2019, Fig. 1C). Plotted rate scalars for simulated data are true parameter values, giving a best case scenario. Using estimated rate scalars would drastically decrease power to detect significant rate correlations even further; particularly for methods that "shrink" rate variation (like the variable-rate regression model—VRRM). However, state correlations will remain robustly estimated. (Row 4): Empirical data for birds and mammals used in Avaria-Llautureo et al. (2019) using phylogenetically mass-corrected log *BMR*. Rates are estimated as the log of magnitude of the phylogenetic independent contrasts. *P*-values from full, best-fitting phylogenetic regression models. The empirical data are consistent with the causal model in (3A), showing strong evidence of "evolutionary coupling" between T_b and *BMR*. Details and complete R script are available on Dryad; https://doi.org/10.5061/dryad.z612jm6bj.

a third factor, Z, affecting both traits (Row 2, Fig. 1). Should such a scenario be evidence of functionally relevant "coupling" as argued by Avaria-Llautureo et al. (2019)? We will return to this shortly.

- Finally, in *Scenario 3*, we first simulate T_b with branch scalars from $\Gamma(\text{shape} = 10, \text{rate} = 10)$ and an overall Brownian motion rate of $\sigma^2_{T_b} = 1$. We then simulate another trait Z (which is not measured at the tips of the phylogeny) with branch scalars from the same Γ-distribution, but a higher rate of evolution where $\sigma^2_Z = 20$. For each tip i, we then generate data on *BMR* by adding the effects of Z and T_b, such that

$$BMR_i = \beta_Z Z_i + \beta_{T_b} T_{b.i} + \epsilon_i,$$

where ϵ_i represents additional residual variation where $\epsilon \sim \mathcal{N}(0, \sigma^2 = 1)$. Because Z is unmeasured at the tips, Z is subsumed in the residual variation in the regression between *BMR* and T_b. Here Z could represent body size or another related variable. Because the range of variation in Z is greater than T_b, we would expect that rate correlations between *BMR* and T_b would be difficult to detect even if there is a true underlying causal relationship, as is the case here (Row 3, Fig. 1).

These simulated scenarios guide our thinking on the meaning and relationships between rate and state associations. With these in mind, we reanalyzed the empirical datasets of Avaria-Llautureo et al. (2019) in two different ways. First, we fit a standard phylogenetic regression model (Martins and Hansen 1997) to estimate a correlation between states (Row 4, Fig. 1). We found that, as expected, T_b and $\ln(BMR)$ are positively correlated in both birds and mammals using models that include $\ln(M)$ (and in the case of mammals, $\ln(M)^2$, because previous studies have found curvature in this relationship; Kolokotrones et al. 2010). We found the best model for the residual variation by comparing Brownian motion, Pagel's λ (Pagel 1999), and Ornstein–Uhlenbeck (OU; Hansen 1997) models with fixed or random roots fit using the R package `phylolm` (Ho and Ané 2014). In mammals, we found very strong coefficients between T_b and $\ln(BMR)$ ($\beta_{T_b} = 10\%$ increase in *BMR* per degree Celsius, $P < 0.001$) and significant negative relationship between T_a and T_b ($\beta_{T_a} = -0.03^\circ$C per $^\circ$C of T_a, $P < 0.01$). In birds, the former is marginally significant ($P = 0.057$) and also positive ($\beta_{Tb} = 2\%$ increase in *BMR* per degree Celsius), and the latter is nonsignificant ($P = 0.45$). We note that very little variation exists for avian T_b and that proper analysis of these correlations would likely require careful accounting of measurement error (Hansen and Bartoszek 2012). These results are consistent with previous analyses of these relationships (Clarke et al. 2010).

Second, we estimate the evolutionary correlation in rates. We note that rate correlations are considerably more difficult to estimate than state correlations because we cannot actually measure these in any organism and must infer them from model parameters; furthermore, even perfectly estimated rates carry only information about the magnitude—not the direction—of change. In their analyses, Avaria-Llautureo et al. (2019) estimated branch-specific rates using a variable-rate regression model (VRRM) (Venditti et al. 2011) fit using Bayesian reversible jump Markov Chain Monte Carlo (Green 1995). The model they fit is quite complex; as such, to visualize the essence of their results without getting mired in the details (e.g., choice of priors, assessment of convergence, etc.) we plot the signal for rate correlations in branch-specific rates using the logarithm of the absolute value of independent contrasts computed at each node (Garland et al. 1992). Although an imperfect measure, the expected value of each contrast will increase with increasing evolutionary rate and thus we can test whether these node-wise estimates of rates are correlated between different sets of traits. For $\ln(BMR)$, we again took the residuals of the best-fitting phylogenetic regression with log body mass (and for mammals, $\ln(M)^2$) to calculate these contrasts. To verify that correlated rate scalars can be identified using this approach, we simulated scenarios of rates with shared shift locations on the mammal phylogeny with correlations in rates ranging from 0 to 1 and evaluated whether significantly positive regressions were obtained (see Supporting Information on Dryad, https://doi.org/10.5061/dryad.z612jm6bj).

For mammals, there is weak but marginally significant evidence that $\ln(BMR)$ and T_b have correlated evolutionary rates ($\rho = 0.10$, $P = 0.03 - 0.08$, depending on how contrasts of 0 are treated; Fig. 1.4C). Our simulations indicate that this value of the slope is consistent with the conclusion of a weak to moderate correlation between rates (Supporting Information). For birds, there is no apparent relationship ($\rho = 0.04$, $P = 0.58$, Fig. 1.4C). Both these results are qualitatively in-line with the findings of Avaria-Llautureo et al. (2019) and we do not dispute their results in this regard. Instead, we wish to highlight the importance of the interpretation of rate correlations, rather than to dispute the methods themselves. Rather, we argue that the presence or absence of associations of rates can occur with or without functional coupling, and care should be taken in their interpretation. For metabolic rate and body temperature, we consider that the combination of strong state coupling and weakly detectable rate-coupling using contrasts, and the VRRM in Avaria-Llautureo et al. (2019), to be consistent with our *Scenario 3*. We therefore read the evidence to suggest that T_b is functionally coupled with *BMR*, but that metabolic rate is also strongly affected by other traits with additional, independent rate variation.

Furthermore, our *Scenario 2* clearly illustrates that shared rate variation can exist even if there is no functional or evolutionary association between the traits (cf. Avaria-Llautureo et al.'s (2019) definition of "coupling"; their Fig. 1B, our

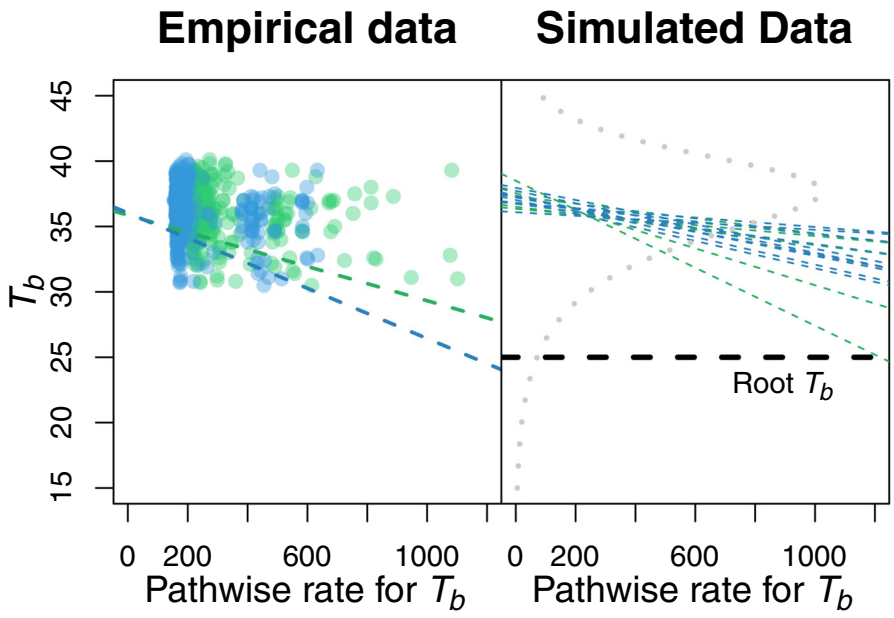

Figure 2. Relationship between pathwise rates estimated from (A) empirical data (cf. Avaria-Llautureo et al. 2019, Fig. 4) and (B) 10 simulated datasets with increasing T_b over time. This negative slope was used by Avaria-Llautureo et al. (2019) to support their conclusion that T_b and T_a had decreased over the evolutionary history of mammals and birds. Analyses of data with the variable-rate regression model (VRRM) in *bayestraits* (green) and in *bayou* (blue) are shown with dotted lines indicating the phylogenetic regression. When the true simulated process includes constraints (instead of purely unconstrained BM) under realistic macroevolutionary landscapes, we recover negative relationships between pathwise rates and trait values, despite an overall increasing trend (from the root value of 25°C to an optimum of 38°C). This macroevolutionary landscape was represented by the Sharpe–Schoolfield thermal-performance curve (gray dotted line), which has a harder bound at high temperatures than at low temperatures. Pathwise rates should never be interpreted as reconstructing ancestral states when the trait in question shows evidence of strongly constrained evolution, as constrained evolution will erase phylogenetic signal and measured rates will simply either reflect asymmetries in the macroevolutionary landscape (as depicted here) and/or asymmetries in diversification/taxonomic sampling (e.g., trait-dependent diversification, likely the case for T_a). Data were simulated with the R package BBMV Boucher et al. (2017). Supporting Information and complete scripts and data are available on Dryad, https://doi.org/10.5061/dryad.z612jm6bj.

Table 1. Phylogenetic signal estimated using an OU model of adaptive (i.e., constrained) evolution.

Clade	Trait	Phylogenetic half-life[a]	ΔAIC over BM
Mammals			
	T_b	0.26	46.3
	BMR[b]	0.16	86.0
	T_a	0.10	141.2
Birds			
	T_b	0.01	132.4
	BMR[b]	0.01	382.4
	T_a	0.01	168.8

[a]Half-life ($\frac{\ln(2)}{\alpha}$) measures the time it takes for a lineage to evolve halfway to the OU optimum in units of tree height. Values of 0 indicate constrained evolution with no phylogenetic signal; while values >1 indicate BM-like evolution and strong phylogenetic signal. All traits strongly reject the BM model ($\Delta AIC > 4$).

[b]Mass-corrected log *BMR*.

Fig. 1.2A–C). For example, consider if both traits were evolving with rates jointly set by a third factor (Z). This could occur if rates of change in traits evolving via genetic drift are all jointly set by effective population size ($Z = N_e$; Walsh and Lynch 2018), even when traits are functionally uncoupled. Thus, we observe that strong and easily detectable causal relationships between traits can be found by state associations (as we observed in the empirical data), even when coupling of rates are too weak to be detectable (Fig. 1). Furthermore, tests of rate-coupling can produce positive results even when traits are not interdependent—which seems to us the most relevant questions to the evolutionary and ecological theories that motivate interest in these traits—but instead whether an additional causal factor might be driving shared rate variation among the two traits.

It is worth noting that a number of researchers investigating molecular evolution have interpreted "convergent rates" in molecular sequence evolution as possible evidence of shared functional significance (Hu et al. 2019; Smith et al. 2020). Rather than assessing convergent states (e.g., Li et al. (2010)), such

studies ask whether shared selective pressures cause joint shifts in substitution rates across genes–which corresponds closely to our *Scenario 2*. Although not the focus of the present manuscript, we think it is worth thinking carefully about the underlying logic of such tests for convergence in molecular evolution. Importantly, studies examining genomic characters have the advantage that additional data is often available to support claims of evolutionary coupling, such as the enrichment for meaningful functional annotations against a large sample of background rates and *a priori* hypotheses of specific causal factors (Chikina et al. 2016; Sackton et al. 2019). Although these additional data make evaluating shared function via coupled rates a more meaningful inference, we would strongly caution against interpreting a lack of evidence for joint shifts in substitution rates as evidence of an absence of functional relationship between molecular sequences—consistent with the arguments we have made here.

Given the existing evidence and known mechanism for how T_b and *BMR* are coupled, combined with the ease of testing state relationships with continuously varying traits (as opposed to the complexities of evaluating convergent states in molecular sequences), the evidence and test for functional coupling here is much more direct. In the context of metabolism, we believe that subsequent evaluation of *Scenario 2* in this case, at least without reference to a background suite of traits or a specific causal factor, is largely meaningless. We acknowledge that we may not have fully understood the causal model envisioned by Avaria-Llautureo et al. (2019). For this reason, we think it is critical that practitioners of phylogenetic comparative methods present graphical causal models of their hypotheses to make clear how the inferences drawn relate to the evidence presented.

Coupled Trends or Model Inadequacy?

Thinking clearly about the traits under study also requires that we consider the causal processes that affect their evolution. This choice of process model can critically affect how we interpret the data, and helps explain another somewhat surprising feature of metabolic evolution found by Avaria-Llautureo et al. (2019): they find that T_b has *decreased* consistently throughout the Cenozoic. Of course, the vast amount of evidence from both physiology and paleontology strongly supports the conclusions that endotherms originated from ectotherms (Grigg et al. 2004; Lovegrove 2017)—this statement is not controversial and it would be extraordinary if this was found to be incorrect. Nevertheless, a plausible alternative is that Avaria-Llautureo et al. (2019) identified a signal of endotherms rapidly increasing to high average body temperatures and subsequently decreasing over the Cenozoic.

To evaluate these alternatives, we first recognize that all models are simplifications of reality, and necessarily make simplifying assumptions. Despite this caveat, we believe that flawed models can often be used to make robust macroevolutionary inferences. We argue that the key to progress despite imperfect tools is to consider carefully how such traits are *expected* to evolve given our knowledge of their function, dynamics, and constraints—knowledge that comes from outside of macroevolutionary statistical models. Here, we reexamine the conclusions of Avaria-Llautureo et al. (2019) to demonstrate our view of this process, and how our expectations can easily explain the observed patterns with entirely expected violations of standard macroevolutionary models.

As stated above, Avaria-Llautureo et al. (2019) used a VRRM based on a BM model of evolution, where traits evolve in an undirected and unbounded manner as they would under genetic drift or randomly varying selection (Hansen and Martins 1996). This seems entirely inappropriate for traits such as *BMR* that are biophysically constrained to stay close to an optimum value (West et al. 2002; Glazier 2005, 2010; Clarke et al. 2010). Although far from perfect, the "adaptation" (or OU) model proposed by Hansen (1997) better captures this biological reality. Analyzing the data of Avaria-Llautureo et al. (2019) bears this out: OU is a much better fit than BM for T_b, T_a, and mass-specific *BMR* in both mammals and birds (Table 1). Birds are especially strongly constrained with little variation in these traits and virtually no phylogenetic signal; this is consistent with both the basic physiology involved and with previous phylogenetic analyses (Uyeda et al. 2017). We note that the models we used for this comparison are much simpler than the ones used by Avaria-Llautureo et al. (2019); comparing variable OU processes (Uyeda et al. 2017) with variable BM processes is beyond the scope of this article, but we argue that the simple case captures the essence of the problem.

So how could the choice of Brownian models over constraint models affect the apparent trend in T_b? Let us assume for the moment that there is no rate variation in the data and that the evolutionary dynamics of T_b and T_a do indeed resemble that of an OU process. One feature of evolutionary processes that follow an OU model is that because the amount of divergence is constrained, the overall rate one would measure if one fits a BM model would depend strongly on the length of the branch itself (the denominator of any estimate of rate); therefore if one uses a Brownian model to fit data generated under an OU-like process, rates will be estimated to be higher on short terminal branches, whereas deeper and longer branches will have lower estimates of rate (for mathematical explanation, see Uyeda et al. 2015). If one computes, as Avaria-Llautureo et al. (2019) did, the total pathwise rates (summing the rates on all the branches from the root to a given tip), one will detect highest pathwise rates in any trait that

is correlated with short terminal branches. Avaria-Llautureo et al. (2019) interpret correlations between pathwise rates and the tip state to be evidence of a secular evolutionary trend. But given the likely dependence on the rate estimates for T_b and T_a with branch lengths, any association between these traits and diversification will result in longer pathwise rates. For example, there is abundant evidence that terminal branches are shorter in temperate regions than in the tropics for these taxa owing to higher species turnover in temperate regions (Weir and Schluter 2007; Schluter 2016; Schluter and Pennell 2017). Thus, we think a far more likely explanation is that this trend is an artifact of fitting Brownian models to constrained data.

Indeed, we are able to recover the exact same pattern of decreasing T_b over time as Avaria-Llautureo et al. (2019) find in their Figure 4, even when we simulate a true trend of *increasing* T_b over time. For example, starting from a low ancestral body temperature of (25°C), it is straightforward to obtain negative relationships between pathwise rates and body temperature under biologically realistic macroevolutionary landscapes (Fig. 2). Specifically, we used a constrained macroevolutionary landscape that reflects standard enzymatic and temperature preference curves for organisms in which high temperatures impose a stronger constraint on organismal performance than low temperatures (i.e., the Sharpe–Schoolfield equation, Schoolfield et al. 1981, see Supporting Information for details of simulations). In other words, this model is similar to an OU model in that divergence is constrained within high-performance regions of the macroevolutionary landscape, but allows asymmetry that is expected given well-known physiological performance curves for T_b (Huey and Kingsolver 1989). We simulated data on this landscape with the R package *BBMV* (Boucher et al. 2017). These asymmetries are expected to generate a negative relationship between T_b and the pathwise rates of T_b *even if there is a true trend towards higher body temperatures towards the present*. This also could explain why Avaria-Llautureo et al. (2019) conclude that T_a and T_b follow the same trajectory toward decreasing temperature over time (cf. Fig. 1.2A–C), despite evidence they are negatively correlated with each other when we consider state correlations (Clarke et al. 2010)—which likely results from the well-known phenomenon of countergradient selection (Schultz et al. 1996; Fangue et al. 2009).

Although estimation of ancestral states under a VRRM may be possible if the trait approaches unconstrained BM-like evolution, we caution interpretation of ancestral states even in the best circumstances—and especially with evidence of constrained evolution and when trends are expected. The inability to estimate the ancestral states for constrained traits is a special case of the "Darwinian uncertainty principle" (Gascuel and Steel 2019), which describes the trade-off in estimating ancestral states versus the rates of the evolutionary process. Just as in the case of molecular sequence data, quantitative traits with high rates that explore a constrained macroevolutionary landscapes will progressively erase evolutionary history, eroding any confidence we should have in already uncertain ancestral state estimates (Boucher et al. 2017).

Conclusions

Phylogenetic comparative datasets cannot be generated from experimentation, and as such causal claims are viewed with appropriate skepticism. However, it is vital to not throw away our causal and process-based evolutionary thinking when applying the increasingly complex and sophisticated statistical toolkit of comparative methods. Instead, our recommendation is to make explicit the causal reasoning that justifies the use of a particular macroevolutionary model—even when such models are likely gross simplifications of reality. Furthermore, we recommend critically evaluating the conclusions of that model in light of likely model violations from the known biology of the trait, even when evaluating these assumptions may be impossible. In our specific example, we examined the relationships between metabolic rate and temperature. Although these relationships have been studied by physiologists for decades, we certainly think there is potential for novel phylogenetic comparative methods to provide new insights into the problem (White et al. 2009; Uyeda et al. 2017; White et al. 2019). However, the causal assumptions of inferences drawn from statistical methods, which are often implicit, must be consistent with the fundamental biology underlying the data (Uyeda et al. 2018). This is especially true because complex phylogenetic comparative models can have many-to-one mappings of interpretations to patterns (Louca and Pennell 2020). Realizing the potentially large extent of such issues in macroevolution, we believe that it is vital to carefully consider how plausible biological causal processes *would* map onto macroevolutionary patterns, even when definitively inferring causation from observational data alone is impossible. This step we think is too often ignored in macroevolutionary models, and by no means unique to the question and study examined here. We hope that our exploration has provided a useful model for how we can make not just our methods and data analyses reproducible, but also our reasoning and inferences.

We demonstrate in this note that we are at odds with the interpretation found in Avaria-Llautureo et al. (2019) who concluded that (1) *BMR* and T_b are evolutionarily decoupled; (2) this decoupling is likely related to positive correlations between cooling ambient global temperatures and decreasing T_b; and (3) that T_b and T_a have decreased since the origin of endothermy. We present evidence that these three major claims rely on, in our opinion, a flawed mapping of statistical methods to relevant causal processes and fitting models that are fundamentally

inconsistent with the biological context. By applying (admittedly, also flawed) models that capture key components of the underlying process and its constraints, we conclude that (1) T_b and *BMR* are strongly coupled in evolutionary state and possibly weakly in evolutionary rate; (2) T_b is negatively correlated to T_a in mammals; and (3) that evidence for a decrease in T_b over the course of endotherm evolutionary history could likely be a spurious result of model inadequacy and the constraints imposed by physiology. More broadly, we want to encourage researchers to recognize the limitations of phylogenetic comparative models and think critically about how to model well-studied biological processes. That is to say, we must choose our statistical tools based on biology rather than let our view of biology be shaped by our choice in statistical tools.

CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

AUTHOR CONTRIBUTIONS

JCU and MWP conceived of the manuscript. JCU, NB, and SM performed analyses. JCU, NB, SM, JR, and MWP wrote the manuscript.

ACKNOWLEDGMENTS

We thank Luke Harmon, Barbara Neto-Bradley, Martha Muñoz, Graham Slater, Joel McGlothlin, and two anonymous reviewers for providing insight and feedback. MWP was supported by an NSERC Discovery Grant. JCU was support by NSF-DEB-1942717.

DATA ARCHIVING

Supporting Information and complete scripts and data are available on Dryad, https://doi.org/10.5061/dryad.z612jm6bj.

LITERATURE CITED

Avaria-Llautureo, J., C. E. Hernández, E. Rodríguez-Serrano, and C. Venditti. 2019. The decoupled nature of basal metabolic rate and body temperature in endotherm evolution. Nature 572:651–654.

Blomberg, S. P., J. G. Lefevre, J. A. Wells, and M. Waterhouse. 2012. Independent contrasts and PGLS regression estimators are equivalent. Syst. Biol. 61:382–391.

Boucher, F. C., V. Démery, E. Conti, L. J. Harmon, and J. Uyeda. 2017. A general model for estimating macroevolutionary landscapes. Syst. Biol. 67:304–319.

Brown, J. H., J. F. Gillooly, A. P. Allen, V. M. Savage, and G. B. West. 2004. Toward a metabolic theory of ecology. Ecology 85:1771–1789.

Chikina, M., J. D. Robinson, and N. L. Clark. 2016. Hundreds of genes experienced convergent shifts in selective pressure in marine mammals. Mol. Biol. Evol. 33:2182–2192.

Clarke, A., P. Rothery, and N. J. Isaac. 2010. Scaling of basal metabolic rate with body mass and temperature in mammals. J. Anim. Ecol. 79:610–619.

Fangue, N. A., J. E. Podrabsky, L. I. Crawshaw, and P. M. Schulte. 2009. Countergradient variation in temperature preference in populations of killifish *Fundulus heteroclitus*. Physiol. Biochem. Zool. 82:776–786.

Felsenstein, J. 1985. Phylogenies and the comparative method. Am. Nat. 125:1–15.

Garland, T., P. H. Harvey, and A. R. Ives. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. Syst. Biol. 41:18–32.

Gascuel, O., and M. Steel. 2019. A Darwinian uncertainty principle. Syst. Biol. 69:521–529.

Glazier, D. S. 2005. Beyond the "3/4-power law": variation in the intra-and interspecific scaling of metabolic rate in animals. Biol. Rev. 80:611–662.

Glazier, D.S. 2010. A unifying explanation for diverse metabolic scaling in animals and plants. Biol. Rev. 85:111–138.

Green, P. J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82:711–732.

Grigg, G. C., L. A. Beard, and M. L. Augee. 2004. The evolution of endothermy and its diversity in mammals and birds. Physiol. Biochem. Zool. 77:982–997.

Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351.

Hansen, T. F., and K. Bartoszek. 2012. Interpreting the evolutionary regression: the interplay between observational and biological errors in phylogenetic comparative studies. Syst. Biol. 61:413–425.

Hansen, T. F., and E. P. Martins. 1996. Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. Evolution 50:1404–1417.

Ho, L. S. T., and C. Ané. 2014. Intrinsic inference difficulties for trait evolution with Ornstein–Uhlenbeck models. Meth. Ecol. Evol. 5:1133–1146.

Houle, D., C. Pélabon, G. P. Wagner, and T. F. Hansen. 2011. Measurement and meaning in biology. Q. Rev. Biol. 86:3–34.

Hu, Z., T. B. Sackton, S. V. Edwards, and J. S. Liu. 2019. Bayesian detection of convergent rate changes of conserved noncoding elements on phylogenetic trees. Mol. Biol. Evol. 36:1086–1100.

Huey, R. B., and J. G. Kingsolver. 1989. Evolution of thermal sensitivity of ectotherm performance. Trends Ecol. Evol. 4:131–135.

Kolokotrones, T., V. Savage, E. J. Deeds, and W. Fontana. 2010. Curvature in metabolic scaling. Nature 464:753–756.

Li, Y., Z. Liu, P. Shi, and J. Zhang. 2010. The hearing gene prestin unites echolocating bats and whales. Curr. Biol. 20:R55–R56.

Louca, S., and M. W. Pennell. 2020. Extant timetrees are consistent with a myriad of diversification histories. Nature 508:502–505.

Lovegrove, B.G. 2017. A phenology of the evolution of endothermy in birds and mammals. Biol. Rev. 92:1213–1240.

Martins, E. P., and T. F. Hansen. 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am. Nat. 149:646–667.

Pagel, M. 1999. Inferring the historical patterns of biological evolution. Nature 401:877–884.

Pélabon, C., C. Firmat, G. H. Bolstad, K. L. Voje, D. Houle, J. Cassara, A. L. Rouzic, and T. F. Hansen. 2014. Evolution of morphological allometry. Ann. N. Y. Acad. Sci. 1320:58–75.

Sackton, T. B., P. Grayson, A. Cloutier, Z. Hu, J. S. Liu, N. E. Wheeler, P. P. Gardner, J. A. Clarke, A. J. Baker, M. Clamp, et al. 2019. Convergent regulatory evolution and loss of flight in paleognathous birds. Science 364:74–78.

Schluter, D. 2016. Speciation, ecological opportunity, and latitude: (American Society of Naturalists Address). Am. Nat. 187:1–18.

Schluter, D., and M. W. Pennell. 2017. Speciation gradients and the distribution of biodiversity. Nature 546:48–55.

Schoolfield, R. M., P. Sharpe, and C. E. Magnuson. 1981. Non-linear regression of biological temperature-dependent rate models based on absolute reaction-rate theory. J. Theor. Biol. 88:719–731.

Schultz, E., K. Reynolds, and D. Conover. 1996. Countergradient variation in growth among newly hatched *Fundulus heteroclitus*: geographic differences revealed by common-environment experiments. Funct. Ecol. 10:366–374.

Smith, S. D., M. W. Pennell, C. W. Dunn, and S. V. Edwards. 2020. Phylogenetics is the new genetics (for most of biodiversity). Trends Ecol. Evol. 35:415–425.

Uyeda, J. C., D. S. Caetano, and M. W. Pennell. 2015. Comparative analysis of principal components can be misleading. Syst. Biol. 64:677–689.

Uyeda, J. C., M. W. Pennell, E. T. Miller, R. Maia, and C. R. McClain. 2017. The evolution of energetic scaling across the vertebrate tree of life. Am. Nat. 190:185–199.

Uyeda, J. C., R. Zenil-Ferguson, and M. W. Pennell. 2018. Rethinking phylogenetic comparative methods. Syst. Biol. 67:1091–1109.

Venditti, C., A. Meade, and M. Pagel. 2011. Multiple routes to mammalian diversity. Nature 479:393–396.

Voje, K. L., T. F. Hansen, C. K. Egset, G. H. Bolstad, and C. Pelabon. 2014. Allometric constraints and the evolution of allometry. Evolution 68:866–885.

Walsh, B., and M. Lynch. 2018. Evolution and selection of quantitative traits. Oxford Univ. Press, Oxford.

Weir, J. T., and D. Schluter. 2007. The latitudinal gradient in recent speciation and extinction rates of birds and mammals. Science 315:1574–1576.

West, G. B., W. H. Woodruff, and J. H. Brown. 2002. Allometric scaling of metabolic rate from molecules and mitochondria to cells and mammals. Proc. Natl. Acad. Sci. 99:2473–2478.

White, C. R., T. M. Blackburn, and R. S. Seymour. 2009. Phylogenetically informed analysis of the allometry of mammalian basal metabolic rate supports neither geometric nor quarter-power scaling. Evolution 63:2658–2667.

White, C. R., D. J. Marshall, L. A. Alton, P. A. Arnold, J. E. Beaman, C. L. Bywater, C. Condon, T. S. Crispin, A. Janetzki, E. Pirtle, et al. 2019. The origin and maintenance of metabolic allometry in animals. Nat. Ecol. Evol. 3:598–603.

Associate Editor: A. McAdam
Handling Editor: A. McAdam