Challenges and advances in methods for measuring phenotypic convergence

David M. Grossnickle[1]*, William H. Brightly[2], Lucas N. Weaver[3], Kathryn E. Stanchak[1], Rachel A. Roston[4], Spencer K. Pevsner[5], C. Tristan Stayton[6], P. David Polly[7], Chris J. Law[1,8]

[1]University of Washington, Department of Biology
[2]University of Sheffield, School of Biosciences
[3]University of Michigan, Department of Ecology and Evolutionary Biology and Museum of Paleontology
[4]University of Washington, School of Dentistry, Department of Oral Health Sciences
[5]University of Oxford, Department of Earth Sciences
[6]Bucknell University, Department of Biology
[7]Indiana University, Department of Earth and Atmospheric Sciences
[8]University of Texas, Austin, Department of Integrative Biology

*D. M. G., dmgrossn@uw.edu

ABSTRACT

Tests of phenotypic convergence can provide evidence of adaptive evolution, and the popularity of such studies has grown in recent years due to the development of novel, quantitative methods for identifying and measuring convergence. These methods include the commonly applied *C*1–*C*4 measures of Stayton (2015), which measure morphological distances between lineages, and Ornstein-Uhlenbeck (OU) evolutionary model-fitting analyses to test whether lineages convergently evolved toward adaptive peaks. We test the performance of *C*-measures and other convergence measures under various evolutionary scenarios. We reveal a

critical issue with *C*-measures: they often misidentify divergent lineages as convergent. We
28 address this issue by developing novel convergence measures (*Ct*1–*Ct*4-measures) that
measure distances between lineages at specific points in time, minimizing the possibility of
30 misidentifying divergent taxa as convergent. *Ct*-measures are most appropriate when putatively
convergent lineages are of the same or similar geologic ages (e.g., extant taxa), meaning that
32 the lineages' evolutionary histories include considerable overlap in time. Beyond *C*-measures,
we find that all convergence measures are influenced by the position of focal taxa in
34 morphospace, with morphological outliers often statistically more likely to be measured as
strongly convergent by chance. Further, we mimic scenarios in which researchers assess
36 convergence using OU models with *a priori* regime assignments (e.g., classifying taxa by
ecological traits), and we find that multiple-regime OU models with phenotypically divergent
38 lineages assigned to a shared selective regime often outperform simpler models. This highlights
that model support for these multiple-regime OU models should not be assumed to reflect
40 convergence among focal lineages of a shared regime. Our new *Ct*1–*Ct*4-measures provide
researchers with an improved comparative tool, but we emphasize that all available
42 convergence measures are imperfect, and researchers should recognize the limitations of these
methods and use multiple lines of evidence when inferring and measuring convergence.

44

46 **INTRODUCTION**

Phenotypic convergence among distantly related taxa is commonly associated with adaptive
48 evolution (e.g., Darwin 1859, Losos 2011), but it can also occur stochastically (Stayton 2008) or
as a byproduct of shared developmental constraints (Losos 2011, Speed and Arbuckle 2016).

50 Evidence for adaptive convergence is strengthened when the magnitude of convergence is greater than expected by chance, and also when the convergent phenotypes are shown to be
52 tied to similar ecological or functional roles. Thus, quantitative examinations of phenotypic convergence are important; they assist researchers in identifying adaptive phenotypes that are
54 driven by shared selective pressures and/or developmental constraints. Novel methods for identifying and measuring convergence have recently been developed (Mahler et al. 2013,
56 Arbuckle et al. 2014, Ingram and Mahler 2013, Stayton 2015A, Speed and Arbuckle 2017, Castiglione et al. 2019), and these methods are often accompanied by statistical tests for
58 comparing the measured convergence to that which is expected from a null model-fitting hypothesis or random data permutations. This has increased the accessibility of quantitative
60 tests for phenotypic convergence, leading to a flood of recent studies on that topic (e.g., Friedman et al. 2016, Zelditch et al. 2017, Baliga and Mehta 2018, Da Silva et al. 2018, Arbour
62 and Zanno 2020, Button and Zanno 2020, Grossnickle et al. 2020, Martinez et al. 2020, Serio et al. 2020, Rincon-Sandoval et al. 2020, Spear and Williams 2020, Baumgart et al. 2021, Huie et al.
64 2021, Rovinsky et al. 2021, Tamagnini et al. 2021, Alfieri et al. 2022, Bennion et al. 2022, Canale et al. 2022, Law 2022).

66 Phenotypic convergence is often defined as lineages evolving to be more similar to one another than were their ancestors (Losos 2011, Stayton 2015A, Mahler et al. 2017), and we
68 follow that definition here. Thus, a signature of convergence is phylogenetic tips that are phenotypically more similar to one another than expected based on assumptions of random
70 change over time; the degree of this similarity of tips is often quantified by convergence measures (Speed and Arbuckle 2017). However, a confounding issue is that multiple types of

evolutionary trajectories can result in lineages that are more similar to one another than expected by chance but are not convergent (as defined above). This includes lineages that retain a shared ancestral phenotype (see discussion on 'conservatism' below) and lineages that have parallel evolutionary trajectories from a similar ancestral trait condition.

The *C*1–*C*4 measures (hereafter, '*C*-measures') developed by Stayton (2015A) have emerged as an especially popular means of quantifying morphological convergence (e.g., Friedman et al. 2016, Zelditch et al. 2017, Baliga and Mehta 2018, Da Silva et al. 2018, Arbour and Zanno 2020, Button and Zanno 2020, Grossnickle et al. 2020, Martinez et al. 2020, Rincon-Sandoval et al. 2020, Spear and Williams 2020, Baumgart et al. 2021, Huie et al. 2021, Rovinsky et al. 2021, Tamagnini et al. 2021, Alfieri et al. 2022, Bennion et al. 2022, Canale et al. 2022, Law 2022). *C*-measures are calculated using geometric distances in phenotypic space between focal lineages, relying on ancestral reconstructions for morphologies at ancestral nodes. The underlying feature of the *C*-measures is the comparison of two measurements: the maximum phenotypic distance between lineages at any points in their evolutionary histories (D_{max}) and the phenotypic distance between phylogenetic tips (D_{tip}). More specifically, D_{max} is the greatest distance between any two points along the lineages in phylomorphospace, with candidate distances including any points between the lineages' most recent common ancestor and the tips (Fig. 1A). *C*1 is the primary *C*-measure and calculated as $1 - (D_{tip}/D_{max})$, with the resulting value representing "the proportion of the maximum distance between two lineages that has been 'closed' by subsequent evolution" (Stayton 2015A). In our conceptual illustration (Fig. 1A), two focal lineages have convergently evolved such that their tips are 70% closer to each other than their D_{max}, resulting in a *C*1 score of 0.7.

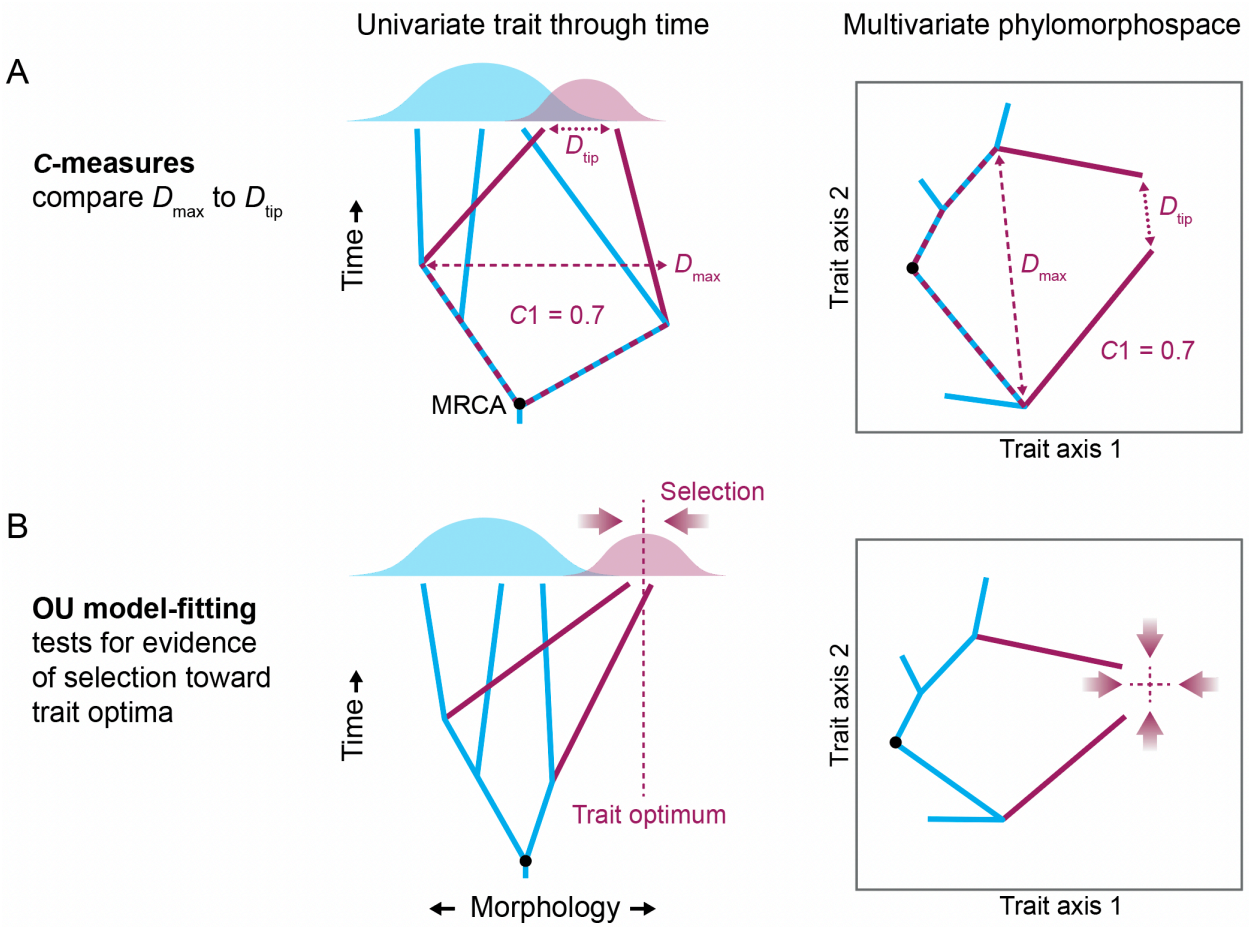

Figure 1. Conceptual illustrations of two methods for assessing phenotypic convergence of focal lineages (maroon): *A*, *C*1 of Stayton (2015A) and, *B*, multiple-regime Ornstein-Uhlenbeck (OU) model-fitting. The *C*1 score of 0.7 indicates that lineages have evolved toward each other to cover 70% of the maximum distance (D_{max}) between their lineages. D_{max} can be measured at any point along the evolutionary histories, including the dashed branches in *A*, and D_{tip} is the morphological distance between phylogenetic tips. Although time is a variable in the univariate illustration in *A*, the *C*-measures do not incorporate time. *B*, OU models include fitting a trait optimum parameter that is often interpreted as the location of an adaptive peak and an 'attraction' parameter that is commonly interpreted as the strength of selection. Abbreviation: MRCA, most recent common ancestor.

One reason for the popularity of *C*-measures is that they can distinguish between convergence and conservatism, which both result in distantly-related phylogenetic tips with similar phenotypes. The key difference between convergence and conservatism centers on the ancestral morphologies of the lineages. Whereas convergence involves ancestors that were less morphologically similar to each other than their descendant tips are to one another (Losos 2011, Stayton 2015A, Mahler et al. 2017), conservatism is the lack of substantial phenotypic divergence from ancestral morphologies relative to what is expected from random processes (Losos 2008, Moen et al. 2013, McLean et al. 2018). The 'blue' lineages in Figure 1B could be considered an example of conservatism; they have not evolved far from the ancestral morphology. *C*-measures account for ancestral patterns via the D_{max} measurement (Fig. 1A). Alternative distance-based methods for testing for convergence (e.g., Wheatsheaf index, Arbuckle et al. 2014, Arbuckle and Minter 2015; θ, Castiglione et al. 2019) cannot adequately differentiate between convergence and conservatism (or parallelism) because phenotypic distances between ancestral morphologies are not considered or, in the case of θ, only partially integrated (Castiglione et al. 2019).

In addition to distance-based measures, researchers often use evolutionary model-fitting analyses to assess convergence, relying on strong fits of multiple-regime Ornstein-Uhlenbeck (OU) models (Hansen 1997, Butler and King 2004) to morphological data for evidence of convergence (e.g., Mahler et al. 2013, Ingram and Mahler 2013, Friedman et al. 2016, Mahler et al. 2017, Baliga and Mehta 2018, Grossnickle et al. 2020, Martinez et al. 2020). An OU process involves 'attraction' toward an 'attractor' or trait optimum (commonly interpreted as the location of an adaptive peak in morphospace), and this attraction and any

resulting convergence is often assumed to be due to selective pressures toward adaptive peaks (Fig. 1B). Convergence is sometimes assumed when the best-supported model indicates that two or more focal lineages have independently evolved toward the same trait optimum. Implementations of these multiple-regime OU models either classify focal taxa into regimes *a priori* (e.g., based on shared ecological or behavioral traits) or examine whether focal taxa are independently identified (i.e., without *a priori* classifications) as shifting toward shared adaptive peaks. The latter method can be implemented using functions in the *SURFACE R* package (Ingram and Mahler 2013; but see Adams and Collyer 2018 for a critique), which was developed specifically for identifying convergence. Other *R* packages have been designed to identify peak shifts without *a priori* classifications (e.g., *Bayou*, Uyeda and Harmon 2014; *l1ou*, Khabbazian et al. 2016; *phyloEM*, Bastide et al. 2018), although they are not specifically designed to identify convergence. However, an OU process does not require that independent lineages shifting toward a peak come from more dissimilar ancestors and thus lineages independently evolving to a shared peak (as evidenced by support for OU models) may not meet the criteria of convergence used in this paper. In addition, OU models may be prone to overfitting to data, and model support should thus be interpreted with caution (Ho and Ané 2014, Cooper et al. 2016, Adams and Collyer 2018; but see Grabowski et al. 2023). Further, in some scenarios it may be challenging to use OU model-fitting analyses to differentiate convergence from conservatism; focal taxa may share a given morphology with their common ancestor, but lacking information about ancestral states, it may be difficult to determine whether this is because the taxa retained the ancestral morphology (conservatism) or have converged on this ancestral morphology from diverse ancestors which did not exhibit that morphology.

Nonetheless, a benefit of OU model-fitting analyses is that the magnitude of the attraction parameter allows an estimate of selective strength toward adaptive peaks, therefore providing valuable information about the evolutionary process that may be driving convergence.

Here, we test the performances of *C*-measures and other means of assessing convergence by applying those methods to simulated data in which a subset of lineages are modeled as truly convergent or truly divergent. Our results highlight a critical concern with the *C*-measures: in many circumstances, they may lead to misclassification of divergent lineages as convergent. Further, we find that the magnitude of convergence measured by distance-based metrics is influenced by the position of focal lineages in morphospace. *C*-measures and θ (i.e., the angle between phenotypic vectors; Castiglione et al. 2019) measure greater convergence in morphological outliers, whereas the Wheatsheaf index (Arbuckle et al. 2014, Arbuckle and Minter 2015) measures greater convergence among lineages that retain their shared ancestral morphology. We also find that in specific scenarios in which OU model regimes are defined *a priori* by traits independent of morphology (e.g., shared ecological or behavioral traits), OU models treating divergent lineages as having a shared optimum (via a shared regime assignment) outperform simpler models; such *a priori* approaches are prevalent, highlighting that researchers should use caution when using model-fitting analyses to assess convergence. Finally, we present an improved method for calculating *C*-measures that minimizes the possibility of erroneously measuring divergent lineages as convergent.

METHODS

Evolutionary simulations

We generated a series of simulated trait datasets to ascertain how frequently convergence measures correctly identify *convergent* lineages and misclassify *divergent* lineages as convergent. Simulated datasets are intended to reflect typical empirical datasets, but we include some extreme morphological outliers to help identify trends. Thus, we simulated traits for 201 species included in a recent phylogenetic treatment of extant mammals, resulting in taxon sampling and phylogenetic structure comparable to many recent comparative studies (Chen and Wilson 2015, Grossnickle et al. 2020, Weaver and Grossnickle 2020, Pevsner et al. 2022). We obtained 1000 randomly chosen phylogenetic trees from the posterior distribution of Upham et al.'s (2019) 'completed trees' analysis. We then used *TreeAnnotator* (Drummond et al. 2012) to generate a maximum clade credibility tree, which was pruned to the species in our sample. The sample includes 13 gliding-mammal species representing five independent evolutionary origins of gliding behavior. We treated the gliders as the focal lineages (*sensu* Grossnickle et al. 2020); they were the subject of manipulation in our simulations (as such, we refer to those simulated glider data as 'gliders'). The five 'glider' clades are spread across the mammalian phylogeny and have varying evolutionary origin ages, making them ideal for representing typical empirical datasets.

Each simulated dataset included six morphological traits. In the non-glider (i.e., 'base') parts of the tree, all six traits were simulated to evolve via Brownian motion (BM). In 'glider' clades, between three to six of these traits were simulated to evolve via an OU process (except for drift-based divergence simulations in which all traits were BM-evolved; see descriptions

196 below), with optima varying to produce convergent or divergent trajectories. The BM component of the simulation ('base simulation') was performed using the

198 *SimulateContinuousTraitsOnTree* function in the *Phylogenetics for Mathematica* package (Polly 2019). The ancestral value for each trait was arbitrarily set to 0.0 and the step rate, σ^2, was set

200 at 1.0 per million years. (Note that the phylogenetic tree was the same for each simulation, but the BM trait values were re-simulated with every 'base simulation.') In 'gliders,' traits evolving

202 via an OU process toward trait optima (see below) were simulated using the *LineageEvolution* function in *Phylogenetics for Mathematica* (ancestral trait values were those generated via BM

204 at the base of each clade). Phylogenetic branches of 'gliders' were those tipped by one of the 13 'glider' species, plus the subtending branches below clades whose tips were all 'gliders.'

206 We generated four types of simulations: 1) convergence of 'gliders' via selection (OU-evolved traits), 2) divergence of 'gliders' via drift (BM-evolved traits), 3) 'constrained

208 divergence' of 'gliders' via selection (OU-evolved traits, limited to positive values), and 4) 'unconstrained divergence' of focal taxa via selection (OU-evolved traits; no limits on traits).

210 The following text describes these four types of simulations, and Figure S1 provides a graphical summary.

212 *Convergence simulations.* For the convergence simulations, we systematically altered 1) the number of traits (of the six total traits) that were subjected to convergent selection (three

214 through six) and 2) their optimum values (or 'targets') (Fig. S1). Convergent traits were evolved via selection (i.e., OU-evolved) toward the same trait optimum (target). The convergent traits

216 were subjected to directional OU-selection for their full branch lengths, which allowed most 'glider' tips (besides those with the shortest branches) to arrive at the adaptive peak. Traits not

subject to convergent evolution continued to evolve by BM. Note that the OU-parameters for all simulations were the same, and the ‘strength’ of convergence was represented by the number of traits subject to convergence rather than the strength of selection per se. Six of six traits being convergent thus represents very strong convergence, whereas three of six traits being convergent represents weaker convergence. Further, we alter the number of convergent traits because it is likely that in empirical datasets some traits are convergent whereas other traits are not convergent (e.g., see the discussion on long bone lengths versus long bone widths in Grossnickle et al. 2020).

By changing the values of trait optima (‘targets’) we systematically controlled the morphological distance of ‘gliders’ from their ancestral condition, allowing us to test how morphospace position influences the results of applied convergence measures (Fig. S1). We used a series of nine trait optima positioned at successively greater distances from the ancestral point in morphospace, starting at zero (convergence toward the ancestral trait values) and successively increasing by 10 to a distance of 80 trait units from the ancestral value. For example, in a simulation with four convergent traits and an optimum of 30, the first four traits all evolved toward an optimum trait value of 30 and the two remaining traits evolved by BM. The range of tip values in the ‘base’ portion of the tree (BM-evolved ‘non-gliders’) has a radius of about 15 trait units, so the first two optima in this iteration (zero and 10) lie within the morphospace occupied by 'non-glider' taxa and the last seven lie increasingly outside the range of morphology of 'non-gliders.' The farthest tested optima (i.e., ~60–80) result in the ‘gliders’ being very far from ‘non-gliders’ in morphospace and thus unlikely to reflect most empirical scenarios, but we include them here to help infer trends in how morphospace position

240 influences convergence measures. Each simulation with all of its iterations was repeated with 15 unique 'base simulations,' and we report results for the means and standard errors of these
242 15 replicates.

Divergence simulations. To simulate divergence via drift (i.e., BM-evolved 'gliders'), we
244 simulated six traits using the *fastBM* function of the *phytools* package (Revell 2012) for *R* (*R* Core Team 2020). Ancestral trait values were set at zero, and, to mimic natural variation, the
246 rate parameter (σ^2) was sampled from a log-normal distribution with log-mean and standard deviation 0 and 0.75, respectively. This was repeated to produce 15 replicate datasets. In
248 contrast to all other simulations, gliders in these datasets did not possess any OU-evolved traits.

250 For the selection-based (OU-evolved) methods for generating divergence, we simulated divergence as selection of the individual 'glider' lineages each toward a different trait optimum.
252 We used a procedure that is as similar as possible to that used in the convergence simulations. Between three and six traits were selected toward the clade-specific trait optimum with a
254 series of target distances ranging from 30 trait units from the ancestral morphology (which extends the lineages past the periphery of the base BM tree and thus ensures that the targets
256 are divergent) to 100 units in steps of 10. This choice, however, means that the divergence simulations are limited to cases in which the lineages are morphological outliers, whereas the
258 drift-based divergence simulations include non-outlying lineages. A different target was randomly selected for each 'glider' clade by choosing a random trait value for each of the traits
260 under selection with the condition that their sum of squares equal the squared target distance

(i.e., that the target lies at a distance of 30, 40, etc. units from the ancestral trait values). The
262 selected lineages are allowed to fully reach their trait optima.

For our primary selection-based divergence simulations, we limited the OU-evolved
264 'glider' traits to positive values. Choosing only positive trait values ensures that the trait optima ('targets') are divergent yet lie within the same quadrant (or hyper-region) of multidimensional
266 space. This constriction of adaptive peaks to one region of morphospace helps to better mimic empirical datasets in which lineages exhibit some morphological similarities but are still
268 geometrically divergent (e.g., Collar et al. 2014, Grossnickle et al. 2020). Further, researchers are most likely to apply convergence measures to lineages with shared ecological or behavioral
270 traits, which often result in morphological similarities, making our simulations more applicable to empirical research. Nonetheless, we also simulated 'unconstrained divergence' for a smaller
272 subset of evolutionary scenarios (see discussion on subset analyses below) in which the 'glider' trait optimum values were not limited to positive values.

274 In total, we generated and analyzed 960 simulated datasets: 540 that simulated trait convergence and 420 that simulated trait divergence.

276

***C*-measures**

278 We applied the *C*-measures (Stayton 2015A) to focal lineages ('gliders') in the simulated datasets. The primary measure, *C*1, is calculated as $1 - (D_{tip}/D_{max})$, with D_{tip} representing the
280 distance between phylogenetic tips of focal taxa and D_{max} representing the maximum distance between any tips or ancestral nodes of those lineages. The *C*1 value is one for complete
282 convergence zero for divergence (i.e., D_{max} is D_{tip}). *C*2 is the difference between D_{max} and D_{tip},

and it captures the absolute magnitude of convergent change. *C*3 and *C*4 are standardized
284 versions of *C*2 that are calculated by dividing *C*2 by the phenotypic change along branches leading to the focal taxa (*C*3) or the total amount of phenotypic change in the entire clade (*C*4).
286 See Stayton (2015A) for full descriptions of *C*1–*C*4. To calculate *C*-measure scores, we used functions in the *R* script from Zelditch et al. (2017), which are computationally faster than the
288 functions in v1.0 through v1.3 of the *convevol R* package (Stayton 2015A, Stayton 2018). *C*1–*C*4 scores were calculated for all simulated datasets, but due to computational limits of analyzing a
290 large number of simulated datasets, we only calculated simulation-based *p*-values for a smaller subset of datasets used for subsequent analyses (see the following subsection).

292 *C*-measures quantify phenotypic convergence between individual phylogenetic tips, not between clades with multiple tips (Stayton 2015A). Thus, we calculated average phenotypes for
294 taxa of focal clades. For example, one glider clade includes six flying squirrel species, so for each of the six simulated traits we calculated mean values for the six species. The averages were
296 then used as the representative flying squirrel lineage. Thus, *C*-measures were measured for five 'glider' lineages, each representing an independent evolution of gliding. The species' traits
298 were not averaged prior to the other types of convergence analyses described below.

300 **Additional measures of convergence**

Subset of simulated datasets. We applied additional measures of convergence (OU model-
302 fitting, θ, and Wheatsheaf index) to a smaller subset of 165 simulated datasets. This subset only includes datasets in which four of six traits were simulated to converge on a specific trait
304 optimum or diverge toward multiple optima, with the remaining two traits evolved by BM (see

Evolutionary simulations subsection for more details). We did not use datasets in which all six
306 traits are convergent because this leads to nearly complete convergence on a trait optimum,
and complete convergence appears to be very rare among empirical analyses (e.g., Grossnickle
308 et al. 2020, Law et al. 2022). Nonetheless, four of six traits being convergent on an optimum
often results in strong convergence (i.e., statistically significant distance-based measures of
310 convergence and strong fits of multiple-peak OU models) among lineages, especially when trait
optima are outliers in morphospace (see Results & Discussion). For convergence simulations,
312 we used the 15 simulated datasets each from the sets of simulations where trait optima were
set at 0, 20, 40, 60, and 80. These represent simulations in which focal lineages evolve toward
314 the ancestral morphology (optimum = 0), evolve just beyond the outer edge of the
morphospace region of BM-evolved lineages (optimum = 20), and evolve far into outlying
316 morphospace (optima = 40, 60, and 80). For selection-based divergence simulations (both
'constrained' and 'unconstrained'), we used the 15 simulated datasets with optima of 40, 60,
318 and 80. (Using trait optima of 0 or 20 could mistakenly simulate convergence toward ancestral
morphologies.) Thus, the subset of datasets includes 75 convergence simulations (15 datasets
320 each for five trait optima) and 45 simulations each for unconstrained and constrained
divergence simulations (15 datasets each for three optima). The following methods were only
322 applied to this subset of datasets.

Evolutionary model-fitting analyses. We performed evolutionary model-fitting analyses
324 using both multivariate models via functions in the *mvMORPH R* package (Clavel et al. 2015)
and univariate models via functions in the *OUwie R* package (Beaulieu et al. 2012). The
326 multivariate *mvMORPH* models offer the benefit of incorporating all data, whereas the

univariate *OUwie* models can be set to allow rate (σ^2) and selection (α) parameters to vary
328 between regimes and therefore can better represent the underlying simulated data, which include both BM-evolved and OU-evolved lineages.

330 We fit four multivariate models to all six simulated traits using functions within the *mvMORPH R* package. The first three models were a single-rate multivariate BM model
332 (mvBM1) that assumes trait variance accumulates stochastically but proportionally to evolutionary time, a single-optimum Ornstein-Uhlenbeck model (mvOU1) that modifies the BM
334 model to constrain each trait to evolve toward a single optimum, and an early burst model (mvEB) that modifies the BM model to decrease evolutionary rate with time. Note that all of
336 these models do not model 'glider' lineages as evolving to a distinct adaptive peak and thus are not expected to reflect convergence of 'gliders' on an adaptive peak (e.g., Fig. 1B). In all model-
338 fitting analyses, the mvEB model collapsed to a mvBM model (i.e., the change-in-rate parameter was zero). Thus, we did not report the mvEB results, and we did not include it when
340 calculating Akaike weights.

We then fit a multivariate OU model with two selective regimes (mvOU2) that allowed
342 'gliders' and 'non-gliders' to exhibit different trait optima (θ). Support for mvOU2 would provide evidence of convergence by indicating that selective forces are driving 'glider' lineages
344 to a shared adaptive peak (Fig. 1B). Note that in the simulations 'non-gliders' evolved via BM, and thus any support for the mvOU2 model is likely to be driven by the 13 'glider' lineages.
346 Relative support for each of the three models was assessed through computation of small-sample corrected Akaike weights (AICcW; Akaike 1974; Hurvich and Tsai 1989). We calculated
348 AICcW for each of the 15 base simulations, and we report the mean values. Prior to fitting the

mvOU2 model, we mapped the gliding regime character state onto the 'glider' branches of the
350 phylogeny (including the subtending branches below 'glider' clades) using the *paintSubTree* function in the *phytools R* package (Revell 2012).

352 In addition to fitting models to multivariate data, we fit models to univariate data (PC1 scores) using functions in the *OUwie R* package (Beaulieu et al. 2012). Although fitting
354 evolutionary models to PC scores may generate biased results (Uyeda et al. 2015, Adams and Collyer 2018), we performed these analyses because they allowed us to fit multiple-regime OU
356 models with varying evolutionary rates (σ) and/or attractions (α) to optima between regimes, which is not a feature of the multivariate *mvMORPH* models. By allowing parameters to vary
358 between regimes, these univariate models are expected to better fit the underlying simulated data compared to the *mvMORPH* models. We fit seven univariate evolutionary models to the
360 subset of simulated datasets, including single-regime models (BM1 and OU1) and a suite of two-regime OU models (i.e., 'OUM' models of Beaulieu et al. 2012). The OU2 model keeps α
362 and σ constant for both 'glider' and 'non-glider' regimes, the OU2A model allows α to vary between regimes, the OU2V model allows σ^2 to vary between regimes, and the OU2VA model
364 allows both σ and α to vary between regimes. See the Supplemental Methods for additional discussion on the univariate model-fitting analyses.

366 *Additional distance-based convergence measures*. We applied two other measures of convergence to the subset of 30 simulated datasets (using all six traits): Wheatsheaf index,
368 which was implemented via the *R* package *windex* (Arbuckle et al. 2014, Arbuckle and Minter 2015), and θ_{real}, which was implemented using the *RRphylo* package (Castiglione et al. 2018,
370 Castiglione et al. 2019). The Wheatsheaf index measures pairwise morphological distances

between putatively convergent taxa, with distances corrected for the degree of phylogenetic relatedness of lineages. These distances are compared to pairwise distances between other lineages in the sample to determine whether putatively convergent lineages are more similar to each other than expected. The θ measurement is the angle between the phenotypic vectors of putatively convergent lineages (note that this θ is different from the θ parameter of OU models), and it is based upon phylogenetic ridge regression. We report the angle obtained by all pairwise comparisons between putatively convergent clades (θ_{real}), standardized by the phylogenetic distance separating them (i.e., expected divergence under a BM model). Significance tests compare standardized θ_{real} values of putatively convergent taxa to values computed for randomly selected tip pairs.

RESULTS & DISCUSSION

C-measure issues

Our analysis of the *C*-measure calculations reveals that the measures do not always perform as intended (Stayton 2015A), especially when putatively convergent lineages are outliers in morphospace. This critical problem can manifest in at least three ways, which we illustrate in Figure 2. First, the more outlying the morphologies are in phylomorphospace (and all else being equal), the greater the *C* scores, indicating stronger convergence. We demonstrate this with a conceptual illustration in Figure 2A (note that D_{tip} remains constant in both scenarios). This does not align with our working definition of convergence and in Stayton (2015A); the distances between ancestral nodes and the distances between descendants are unchanged between the scenarios, and thus we could expect *C* scores to be the same for both scenarios. The pattern of

greater convergence in outliers is also demonstrated by results of applying *C*-measures to evolutionary simulations (Figs. 3 and S2); although this pattern only occurs when some simulated traits are evolved by BM (see the *Relatively greater convergence in morphological outliers* subsection below).

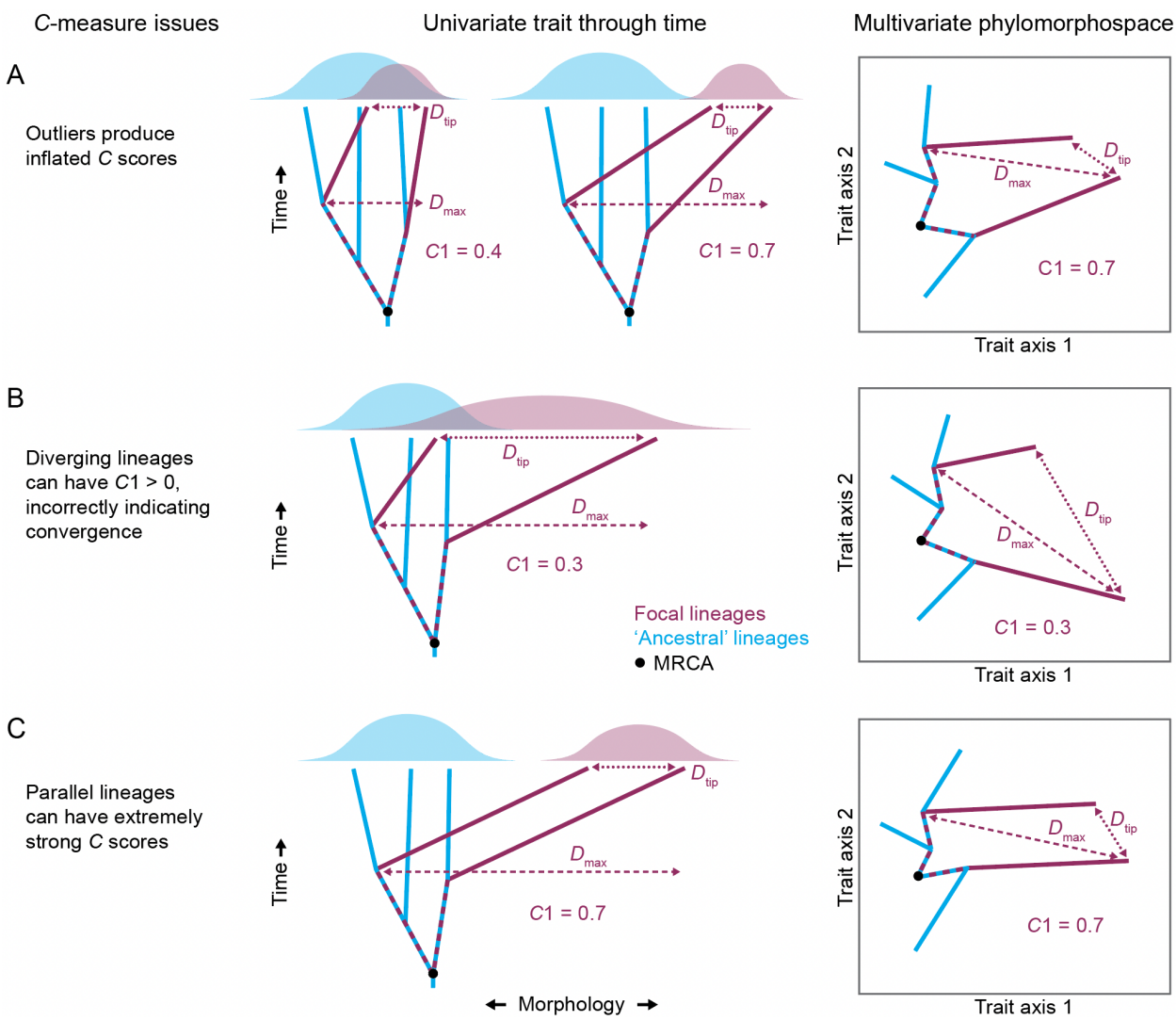

Figure 2. Conceptual illustrations of *C*-measure issues. *C*1 scores are greater than zero for the divergent (*B*) and parallel (*C*) focal lineages ('maroon'), incorrectly indicating that the lineages are convergent. See the main text and Figure 1 for more information on *C*1, D_{max}, and D_{tip}. The *C*-measure issues highlighted here also apply to evolutionary model-fitting analyses. The distribution curves in univariate illustrations could represent adaptive peaks, and OU model-fitting analyses are more likely to identify unique shared adaptive peaks when a peak is farther from the ancestral morphology (Table 1).

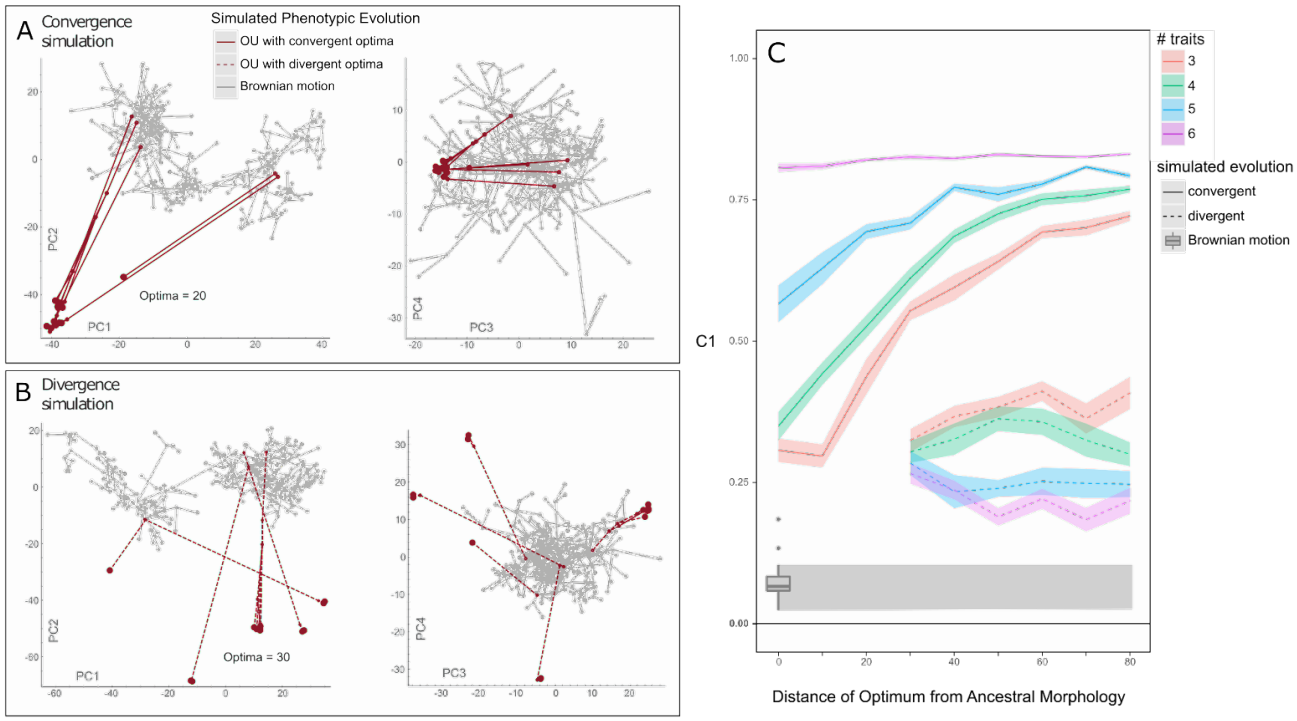

Figure 3. PCA phylomorphospaces for example datasets that simulate convergence (*A*) and divergence (*B*) of all six traits of five focal clades ('gliders'). Traits were selected toward optima via an OU process, with all traits of convergent taxa selected toward a value of 20 and traits of divergent taxa selected toward varying values that result in evolutionary change equal to that of convergent taxa (see Methods). (*C*) *C*1 scores for simulated convergent lineages (solid lines) and divergent lineages (dashed lines), using datasets in which focal taxa have varying numbers of convergent/divergent traits (of six total) and trait optima positions. Any traits and lineages not selected to be convergent/divergent were evolved by Brownian motion (BM). Focal taxa evolved toward trait optima after they originated. Divergent trait optima are randomized, but they are limited to being positive numbers (whereas BM-evolved traits can be positive or negative), resulting in divergent lineages evolving in the same direction along PC1 (e.g., *B*) but otherwise being divergent (unless convergence occurs by chance). *C*1 values above zero indicate convergence, and a value of one would reflect complete convergence (i.e., phenotypically identical tips). We did not simulate divergence to trait optima of 0, 10, and 20 because the simulations might mistakenly generate convergent lineages near the middle of morphospace. As a second means of simulating divergence, we allowed the focal lineages to evolve via BM, and these results are displayed as a box-and-whisker plot in *C*. *C*1 results are means and standard errors of 15 simulated datasets.

The second and third issues with the *C*-measures are more problematic: divergent and parallel lineages can have *C*1 scores that are greater than zero, incorrectly indicating that the lineages are convergent (Fig. 2B, C). In Figure 2B we illustrate lineages that are diverging morphologically (in univariate and multivariate morphospace), but which have a *C*1 score of 0.3, incorrectly suggesting that the lineages have experienced substantial convergence (i.e., closing about 30% of the maximum distance between lineages). To further test this issue, we measured *C*1 in lineages simulated to have divergent traits (Fig. 3B), and *C*1 values are consistently greater than zero (and statistically significant) when applied to the 'constrained divergence' simulations (Fig. 3C; Table 1), incorrectly indicating convergence instead of divergence. The *C*1 scores for divergent focal lineages remain around 0.3 at all morphospace positions (Fig. 3B), although see the Supplemental Results and Figure S4 for discussion on how these results for different morphospace positions may be sensitive to methodological choices.

454 **Table 1.** Tests of convergence among focal lineages ('gliders') of the simulated datasets using distance-based measures. Results are means of the 15 simulated datasets for each
456 optimum. For θ results, we report θ_{real} standardized to phylogenetic distance between clades. Note that relatively smaller θ_{real} values (i.e., smaller angles between phenotypic
458 vectors) suggest greater convergence, whereas relatively larger Wheatsheaf index, *C*1, and *Ct*1 values indicate greater convergence. Statistical significance (*, $p \leq 0.05$; **, $p \leq 0.01$;
460 ***, $p \leq 0.001$) for *C*1 and *Ct*1 is based on comparisons to results of 100 simulations via a BM model, and for the Wheatsheaf index it is based on bootstrapping with 1000 replicates.
462 Significance of standardized θ_{real} values is based on bootstrapping with 1000 replicates for each pairwise comparison between the five 'glider' clades (except the monospecific clade).
464 In all cases, the reported significance is based on means of 15 analyses for a given trait optimum. The 'constrained divergence' simulations limited the simulated traits of 'gliders'
466 to be positive values, thus constraining them to one region of morphospace (see Methods).

Simulations	**Convergence measure**	**Trait optimum**				
		0	**20**	**40**	**60**	**80**
Convergence	θ (*RRphylo*)	0.393	0.126*	0.071**	0.044**	0.040**
	Wheatsheaf index	1.976***	1.715***	1.314	0.905	0.626
	C 1	0.350*	0.526***	0.685***	0.751***	0.769***
	Ct 1	0.176*	0.233**	0.367***	0.497***	0.483***
Divergence ('constrained')	θ (*RRphylo*)	–	–	0.220	0.242	0.229
	Wheatsheaf index	–	–	0.826	0.581	0.434
	C 1	–	–	0.363**	0.325**	0.310**
	Ct 1	–	–	0.035	0.011	-0.003

468

470 Similarly, outlying lineages evolving along parallel phylomorphospace trajectories from a similar ancestral condition have extremely strong *C*1 scores (Fig. 2C). This is unexpected
472 because the ancestral nodes of both lineages are the same morphological distance from one another as the distance between tips; this is not convergence according to the definition of

474 convergence adopted here or in Stayton 2015A (see Introduction; Losos 2011, Mahler et al. 2017).

476 The second and third issues with *C*-measures may commonly occur when at least one focal lineage has diverged rapidly from other taxa, which is a scenario that can often generate 478 morphological outliers (Fig. 2). However, these issues are not limited to outliers. For instance, the scenarios for focal ('maroon') lineages in Figures 2B and 2C could occur in any region of 480 morphospace. Thus, these issues have major implications for many empirical studies (see discussion of examples below); divergent or parallel lineages may often be incorrectly 482 interpreted to be convergent, no matter their position in morphospace.

The possibility of diverging and parallel lineages having *C*1 scores that incorrectly 484 indicate convergence (i.e., are greater than zero) stems from the D_{max} measurement (as defined by Stayton 2015A and calculated in versions 1.0 through 1.3 of the *convevol R* package), which 486 can be erroneously inflated, especially when lineages are morphological outliers. D_{max} can be measured between ancestral nodes (e.g., see the illustration in Figure 1A), between tips, or 488 between a node and a tip (which is the case in all examples in Figure 2). For converging lineages, D_{max} is expected to be longer than D_{tip} (Stayton 2015A). For diverging lineages, in 490 contrast, D_{max} is expected to be the morphological distance between the tips, meaning that D_{max} equals D_{tip} (and $C1 = 0$). However, this is not always the case; divergent lineages can have a 492 D_{max} length that is not between tips, as illustrated in Figure 2B. Thus, D_{max} can be greater than D_{tip} (indicating convergence) even when lineages are divergent. Although we illustrate this issue 494 using diverging phylogenetic tips (Fig. 2B), the problem could also arise if there are internal nodes that are similarly divergent and outlying in morphospace (and branching lineages from

496 those nodes do not converge on other focal lineages); thus, this issue is not solely due to allowing D_{max} to be measured to tips.

498

Other measures of convergence

500 To test whether other convergence measures also experience similar issues as those of the *C*-measures, we applied two other 'distance-based' metrics (Wheatsheaf index [Arbuckle et al. 2014, Arbuckle and Minter 2015] and θ [Castiglione et al. 2019]) and OU model-fitting analyses to a subset of simulated datasets (Table 1).

504 *Distance-based convergence measures*. Consistent with the *C*-measures, the θ_{real} results (standardized to phylogenetic distance between clades) indicate greater convergence in 506 morphological outliers (Table 1). That is, the angle between phenotypic vectors, θ_{real}, decreases when lineages evolve toward optima that are farther from the ancestral morphology. This is 508 unsurprising because a relatively farther trait optimum results in greater trait values in the lineages simulated to be convergent, and, all other variables being equal, greater trait values 510 should result in smaller angles between phenotypic vectors. However, unlike the *C*-measures, θ_{real} does not consistently identify simulated divergent lineages as convergent (Tables 1 and S), 512 although many 'constrained divergence' simulations (~40%) still returned significant results. In contrast to the *C*-measures and standardized θ_{real}, the Wheatsheaf index measures less 514 convergence in outliers relative to non-outliers; values decrease when convergent taxa are farther from the ancestral morphology in morphospace.

516 *Evolutionary model-fitting analyses*. When 'gliders' are simulated to be morphologically convergent, multiple-regime OU models (with 'gliders' assigned to one regime) are the best-

fitting models (Table 2), correctly suggesting that the ‘glider’ lineages are convergently evolving toward a shared adaptive peak (Fig. 1B). However, we obtain similar results when ‘gliders’ are simulated to be divergent – multiple-regime OU models outperform simpler models even though the ‘gliders’ are not evolving toward a shared adaptive peak (Table 2). The following text explains these results in greater detail.

544 **Table 2.** Evolutionary model-fitting analyses in which multivariate models (via functions in the *mvMORPH R* package) were fitted to all six simulated traits and univariate models (via
546 functions in the *OUwie R* package) were fitted to PC1 scores. Model-fitting results for each trait optimum are the mean AICcWs of 15 simulated datasets. Model support for the two-
548 regime models (mvOU2 and all OU2 models) could be interpreted as evidence for convergence because these are modeling evolution of focal lineages toward a shared
550 adaptive peak (although see the Results & Discussion for cautionary notes on this assumption). The 'constrained divergence' simulations limited the simulated traits of focal
552 taxa ('gliders') to be positive values, thus constraining them to one region of morphospace (see Methods). See the Supplemental Results for 'unconstrained divergence' results, and
554 see the Methods for descriptions of the models. Abbreviations: AICcW, small-sample corrected Akaike weights; mvBM, multivariate Brownian motion model; mvOU, multivariate
556 Ornstein-Uhlenbeck model.

	Simulations	**Model**	**Trait optimum of focal taxa**				
			0	**20**	**40**	**60**	**80**
Multivariate (via *mvMORPH*)	**Convergence**	mvBM1	1.000	0.000	0.000	0.000	0.000
		mvOU1	0.000	0.000	0.000	0.000	0.000
		mvOU2	0.000	1.000	1.000	1.000	1.000
	Divergence ('constrained')	mvBM1	–	–	0.000	0.000	0.000
		mvOU1	–	–	0.000	0.000	0.000
		mvOU2	–	–	1.000	1.000	1.000
Univariate (via *OUwie*)	**Convergence**	BM1	0.087	0.051	0.000	0.000	0.000
		OU1	0.032	0.022	0.000	0.000	0.000
		OU2	0.045	0.023	0.000	0.000	0.000
		OU2A	0.024	0.010	0.001	0.000	0.000
		OU2V	0.038	0.040	0.000	0.000	0.000
		OU2VA	0.773	0.854	0.999	1.000	1.000
	Divergence ('constrained')	BM1	–	–	0.000	0.000	0.000
		OU1	–	–	0.000	0.000	0.000
		OU2	–	–	0.000	0.000	0.000
		OU2A	–	–	0.105	0.079	0.214
		OU2V	–	–	0.046	0.000	0.000
		OU2VA	–	–	0.849	0.921	0.786

558

Our simulations and analyses mimic an empirical scenario in which researchers test an *a*
560 *priori* hypothesis that specific focal taxa (in this case, mammalian 'glider' species; *sensu* Grossnickle et al. 2020) are evolving toward, and possibly converging on, an adaptive peak. We
562 have assigned the 'gliders' to one selective regime and the 'non-gliders' to a second regime, and then use model support for multiple-regime OU models as evidence for convergence on
564 adaptive peaks (Fig. 1B). In this type of analysis, support for multiple-regime OU models could suggest that the focal lineages (assigned to a single regime) have convergently shifted toward a
566 shared adaptive peak for that regime. This procedure is prevalent in the literature (e.g., Collar et al. 2014, Friedman et al. 2016, Baliga and Mehta 2018, Grossnickle et al. 2020, Rincon-
568 Sandoval et al. 2020, Law 2022), although other studies assess convergence without *a priori* hypotheses for lineages using alternative model-fitting procedures (e.g., using functions in the
570 *SURFACE R* package; Ingram and Mahler 2013), which is not examined here (see Adams and Collyer 2018).

572 The simulated datasets used in model-fitting analyses include lineages that were evolved by both BM ('non-gliders') and by a mix of OU and BM processes ('gliders'; two traits
574 BM-evolved and four traits OU-evolved), meaning that we do not expect any of the tested multivariate *mvMORPH* models (mvBM1, mvOU1, mvEB, mvOU2) to be good fits to the data
576 because they do not include a mix of BM and OU processes. Nonetheless, we fit these models because the multivariate models enable us to incorporate all data into the analyses. Conversely,
578 the univariate *OUwie* models are expected to be relatively better fits to the data because they

allow parameters to vary between regimes, but they are only fit to PC1 scores and thus do not model all data.

For the convergence simulations, the univariate OU model-fitting analyses, which include models with varying rate and attraction parameters between regimes, correctly identify convergence of focal lineages at any optima values (i.e., at any positions in morphospace): the OU2VA model is the best-fitting model to PC1 scores and it allows the selection and rate parameters to vary between regimes (Table 2). The results correctly identify different underlying evolutionary processes for the two regimes, with 'gliders' showing stronger evidence of convergence via an OU process (i.e., greater selection and rate parameters compared to those of the non-glider regime). However, for the multivariate model-fitting analyses in which rate and attraction parameters are kept constant between regimes, the analyses struggle to identify convergence when focal lineages converge on a morphology that is similar to the ancestral morphology (i.e., in the middle of morphospace), with mvBM1 out-performing mvOU2 (Table 2). The results correctly identify different underlying evolutionary processes for the two regimes, with 'gliders' showing stronger evidence of convergence via an OU process (i.e., greater selection and rate parameters compared to those of the non-glider regime). Thus, for robust analyses of convergence via model-fitting analyses, we recommend that researchers include models that allow all model parameters to vary among regimes.

Our results for 'constrained divergence' simulations highlight a potential pitfall of the assumption that model support for multiple-regime OU models is indicative of convergence (Fig. 1B). Contrary to this assumption, the two-regime OU models (both multivariate and univariate) outperformed simpler, single-regime models when fit to data in which lineages of

both regimes (‘gliders’ and ‘non-gliders’) are divergent (via the ‘constrained divergence’ simulations; Table 2). Although the divergent focal lineages (‘gliders’) are partially constrained to one region of morphospace in the simulations (see Methods and further discussion below), they still show considerable divergence in phylomorphospace (Fig. 3C) rather than convergent or parallel attraction toward a shared optimum (a presumptive adaptive peak), which is an assumption of the OU process (Fig. 1B). Further, taxa representing the second regime (‘non-gliders’) were evolved by BM, not an OU process. Thus, taxa of neither selective regime are expected to be well-fit by by an OU model, and yet the two-regime OU model is a substantially better fit to the data than the null, BM1 or mvBM1 model in both the univariate and multivariate analyses (Table 2). For the univariate analyses, the best-fitting model is OU2VA, which allows parameters to vary between ‘glider’ and ‘non-glider’ regimes – this is the model that is most appropriate considering the generating processes of the simulations (i.e., divergent BM and OU processes). However, OU2VA is still a multiple-regime OU model, and its performance could be interpreted as evidence for the presence of two adaptive peaks, one for ‘gliders’ and one for ‘non-gliders,’ even though lineages of each regime were not simulated as evolving toward single, regime-specific adaptive peaks. This result is consistent with the model-fitting results of Collar et al. (2014) for eel morphology – durophagous eels appear to be divergent in phylomorphospace, but two-regime OU models (with durophagous taxa in one regime and non-durophagous taxa in a second regime) were the best-fitting models.

That the two-regime OU model is the best fit to our divergence data, despite the expectation that none of our fitted models fit the simulated data, simply reflects that two-regime OU models are the best-fitting of bad-fitting models. Knowing the underlying processes

that generated the simulated data, we recognize that the multiple-regime OU models are not a good representation of the underlying data. However, we reiterate that we are mimicking specific empirical scenarios in which researchers use *a priori* regime assignments (e.g., based on gliding behavior) and do not know the evolutionary processes that generated their trait data. Our results here highlight how researchers using OU model-fitting analyses in the same manner should be cautious when interpreting results of model-fitting analyses – if multiple-regime OU models are best-fitting to data, it does not necessarily indicate that focal lineages experienced convergence.

The use of OU model-fitting analyses to test for convergence can offer a valuable complement to distance-based metrics, but researchers should be cautious both when choosing models to fit to data and when interpreting results, especially considering that empirical datasets often include complex evolutionary patterns and small sample sizes for some regimes (Ho and Ané 2014, Cooper et al. 2016, Grabowski et al. 2023). Although not explored in this study, multiple-regime BM (BMM) models may offer alternative options that complement multiple-regime OU models (e.g., Law et al. 2018, Grossnickle et al. 2020, Martinez et al. 2020, Rincon-Sandoval et al. 2020). BMM models allow varying phylogenetic means among regimes and can be fit using functions within some *R* packages, including *mvMORPH* (Clavel et al. 2015). Because BMM models do not model selection toward an optima, support for BMM models over OUM models may suggest that there is limited or no convergence among lineages of interest (e.g., Grossnickle et al. 2020), and in some cases a BMM model might serve as a more appropriate null model than BM1.

644 An additional OU model-fitting approach involves testing for convergent shifts toward shared adaptive peaks without *a priori* regime classifications for focal lineages (e.g., using
646 functions in the *SURFACE R* package; Ingram and Mahler 2013). We did not apply this method in part because results are challenging to interpret across many simulated datasets, especially
648 when peak shifts are likely to be identified in taxa that evolved via BM (Adams and Collyer 2018), and because the method may be less appropriate than our method (with *a priori* regime
650 classifications for 'gliders' and 'non-gliders') for testing the specific hypothesis of this study. Nonetheless, this method provides a valuable research tool for examining convergence,
652 especially for analyses that are exploratory (i.e., without hypotheses for specific lineages), and we recommend applying it to empirical datasets as a supplement to other convergence
654 measures (e.g., Grossnickle et al. 2020, Rincon-Sandoval et al. 2020).

Our goal with divergence simulations was to mimic empirical samples with lineages that
656 might share some similarities and thus be tested for convergence. Therefore, in our primary divergence simulations (i.e., 'constrained divergence'), the divergent focal lineages ('gliders')
658 were partially constrained to one region of morphospace by limiting the randomized optima to be positive values for the four OU-evolved divergent traits (the other two traits were evolved
660 by BM and could be negative; see Methods). This methodological decision helps to explain the relatively strong fits of the two-regime OU models to the simulated divergence datasets.
662 Because 'glider' lineages are shifting in similar directions (e.g., all divergent lineages are in negative PC2 space in Figure 3B), the 'gliders' may be modeled as evolving toward a broad
664 adaptive peak that occupies a large region of one side of morphospace, especially because an OU process does not require that independent lineages shifting toward a peak come from more

666 dissimilar ancestors. Figure 2B provides a conceptual illustration of this scenario; the focal lineages ('maroon') are diverging but still evolved toward a broad adaptive peak, which is
668 broader than the adaptive peak of the ancestral ('blue') lineages. This could be a similar scenario to the divergent outlier lineages in our simulated dataset (Fig. 3C). The durophagous
670 and non-durophagous eels examined by Collar et al. (2014) are best fit by a two-peak model, and they may provide a possible empirical example of this scenario; durophagous eels are
672 morphologically divergent but appear to be 'constrained' to one region of morphospace, a scenario similar to our simulations.

674 For 'unconstrained divergence' simulations, we allowed the 'gliders' to diverge in all directions in morphospace, and all convergence measures, including OU model-fitting analyses,
676 correctly identify the taxa as being divergent (Supplemental Results; Table S2). Thus, the central issue highlighted here for OU model-fitting analyses (i.e., that multiple-regime OU models are
678 best-fitting to divergent lineages) is likely limited to scenarios in which divergent focal taxa share some morphological similarities.

680 *Summary of issues with convergence measures.* We found that all convergence measures, including both distance-based and model-fitting approaches, correctly identified
682 convergence among lineages simulated to be convergent (Tables 1 and 2, Figs. 3C and 5A). However, we uncovered a critical issue with *C*-measures: they often misidentify divergent
684 lineages as being convergent (Fig. 2B, Fig. 3C, Table 1). In addition, all convergence measures appear to be influenced by the positions of focal taxa in morphospace. *C*-measures, θ_{real}, and
686 some OU model-fitting analyses (those in which rate and attraction parameters are not allowed to vary between regimes) all result in stronger measures of convergence when simulated

688 convergent taxa evolve toward relatively farther trait optima. In contrast, the Wheatsheaf index shows weaker convergence in outliers (Table 1). Further, model support for multiple-regime OU
690 models, which is often interpreted as support for convergent shifts toward shared adaptive peaks, should be considered with caution because in some scenarios these models may be the
692 best fits to divergent lineages (Table 2). To help address this issue, we recommend that researchers fit a broad range of models that allow parameters to vary among regimes and
694 perform model-fitting analyses in conjunction with distance-based convergence measures.

696 **Measuring convergence through time via *Ct*-measures**

Despite any shortcomings, *C*-measures have benefits over other convergence measures,
698 including the ability to distinguish between convergence and conservatism (Stayton 2015A). Thus, our objective is not to discourage the use of distance-based metrics like *C*-measures but
700 rather to identify issues and encourage the development of improved measures.

We help to address the *C*-measure issues by presenting novel distance-based
702 convergence measures that are derived from the *C*-measures. The new measures are calculated using the same equations as those for *C*1–*C*4 (except with a change to *C*4; see below and
704 Supplemental Methods), but we limit the candidate D_{max} measurements to distances between lineages at synchronous 'time slices' coinciding with internal phylogenetic nodes (Fig. 4). We
706 refer to the new measures as *Ct*-measures (or *Ct*1–*Ct*4) and D_{max} as $D_{max.t}$ because time (*t*) is incorporated when measuring morphological distances between lineages, unlike the *C*-
708 measures. *Ct*1 scores can be interpreted in the same way as *C*1 scores were intended to be interpreted (Stayton 2015A): positive *Ct*1 scores represent a proportion of the maximum

710 morphological distance between lineages that has been closed by convergent evolution, with a value of one representing complete convergence. Like *C*-measures, statistical significance for

712 *Ct*-measures is based on comparison with expectations for evolution proceeding entirely on a BM model, with simulations used to generate the expectations.

714

716

718

720

722

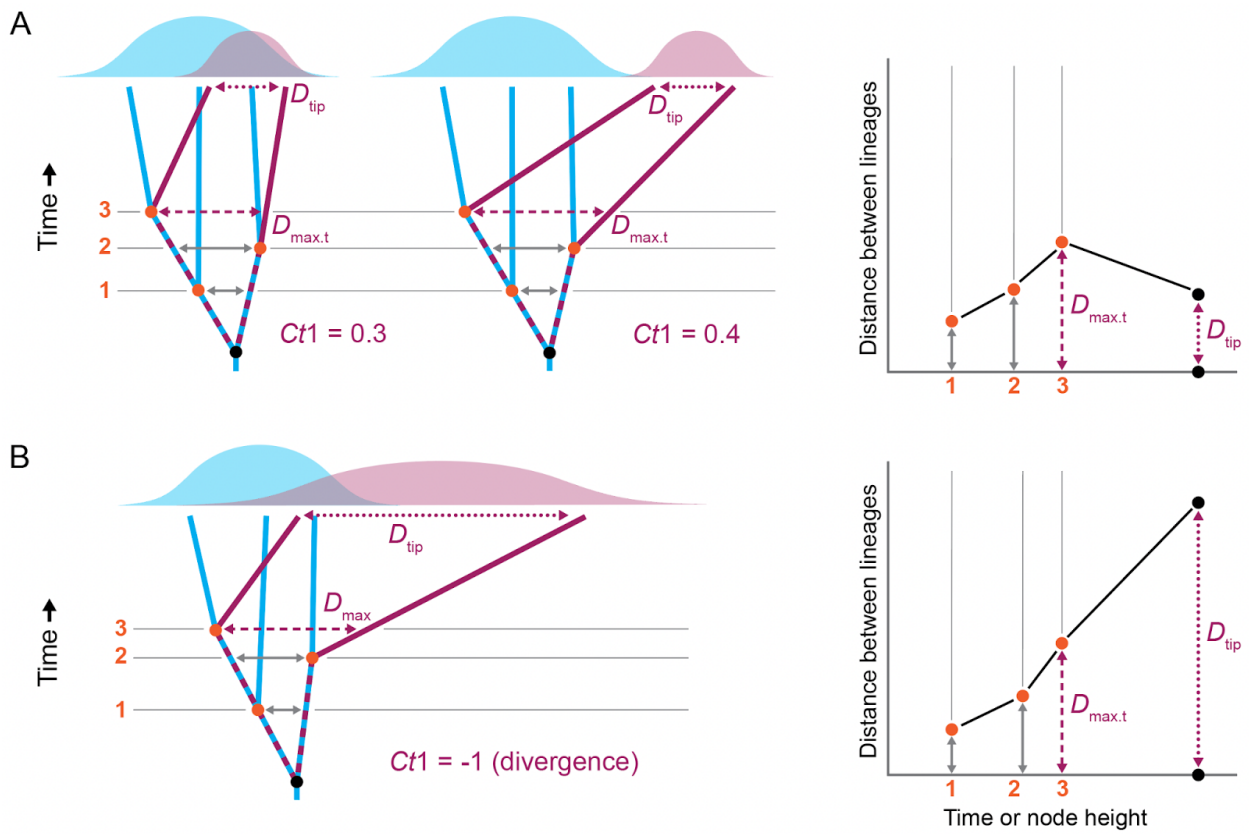

Figure 4. Conceptual illustration of our new *Ct*1 convergence measure, which is calculated like *C*1 of Stayton (2015A) but candidate $D_{max.t}$ measurements are limited to 'time slices' at internal phylogenetic nodes. The plots on the right show the three candidate $D_{max.t}$ measurements and the distance between lineages at the tips (D_{tip}). The scenarios in *A* and *B* are the same as those in Figures 2A and 2B, respectively. In contrast to *C*1, *Ct*1 correctly identifies divergence (negative score) in the scenario in *B*. Although the *Ct*1 score is greater when lineages are outliers (*A*), note that the *Ct*1 scores (0.3 for non-outliers and 0.4 for outliers) are more similar to each than are the *C*1 scores in the same scenarios (0.4 and 0.7; Fig. 2A), indicating that *Ct*-measures are less influenced by positions of taxa in morphospace compared to *C*-measures.

By limiting the candidate $D_{max.t}$ measurements to time slices, the *Ct*-measures minimize
740 the possibility of $D_{max.t}$ being erroneously inflated by divergent tips. This is conceptually illustrated in Figures 4A and 4B, which show the same scenarios as in Figure 2A and 2B.
742 Whereas the *C*1 score in Figure 2B incorrectly indicates convergence (i.e., *C*1 is greater than zero), the *Ct*1 score in Figure 4B correctly indicates divergence (i.e., the value is negative; unlike
744 the *C*-measures, the *Ct*-measures allow divergence results to be negative).

Unlike the D_{max} measurement, the D_{tip} measurement has not been altered from its
746 original implementation in *C*-measures (Stayton 2015A) and is not limited to a synchronous time slice, thus allowing for distances between tips to be compared even if the tips vary in
748 geologic age (e.g., comparison of an extinct taxon and an extant taxon). However, unlike the *C*-measures, the *Ct*-measures do not allow $D_{max.t}$ to be measured between tips (i.e., $D_{max.t}$ cannot
750 equal D_{tip}). This means that divergent taxa will have negative *Ct* scores, whereas *C*-measures (as they were initially intended) will measure divergent taxa as having scores of zero (i.e., D_{max}
752 equals D_{tip}). See the Supplemental Methods for more information on the *Ct*-measures.

In addition to developing the *Ct*-measures, we added several new features to the
754 *convevol R* package (Stayton 2018). This includes allowing *C* and *Ct*-measures to compare clades that contain multiple lineages, whereas the original routines were limited to comparisons of
756 individual lineages (see Methods; Stayton 2015A). Clade comparisons are enabled by 1) excluding pairwise comparisons between within-clade lineages (e.g., two flying squirrel species)
758 and 2) weighting of *C* and *Ct* scores and *p*-values based on the number of pairwise comparisons between focal clades (see Supplemental Methods). Further, *C* and *Ct*-measures can be
760 measured using single traits (the original routines only permitted measures of multivariate

distances, although they were adapted for univariate analyses in some studies; Spear and Williams 2020, Law 2022), and we updated the *C*4 (and *Ct*4) calculation to better match the original description of that measure. See the Supplemental Methods for additional information on these updates. We used an *R* script written by Jonathan S. Mitchell and published in Zelditch et al. (2017) as a foundation for the updated functions. The run times for the revised *R* functions (*convrat.t* and *convratsig.t*) are approximately ten times faster than the original functions of Stayton (2015A) when using our simulated datasets. We did not revise *C*5, which is a frequency-based convergence measure that tallies the number of times lineages enter a region of morphospace (Stayton 2015A), because it is not influenced by the issues highlighted here.

We have also developed a new *R* function, *plot.C*, that produces a plot of the distances between lineages through time. This type of plot is conceptually illustrated in Figure 4, and Figure 5B includes a phylogeny and plot produced by the *plot.C* function for an example dataset from our convergence simulations, showing pairwise distances between three 'glider' lineages. An additional example output of *plot.C* is provided in Figure S7 for the 'twig' ecomorphotype lineages of anoles, although we separated convergent and non-convergent pairwise comparisons for ease of interpretation. These plots allow researchers to visualize when the measured $D_{\mathrm{max.t}}$ occurred during the evolutionary history of the lineages, and they may be useful for applications beyond studies of convergence. The candidate $D_{\mathrm{max.t}}$ measurements at time slices are provided as an output of the *calcConvCt* function.

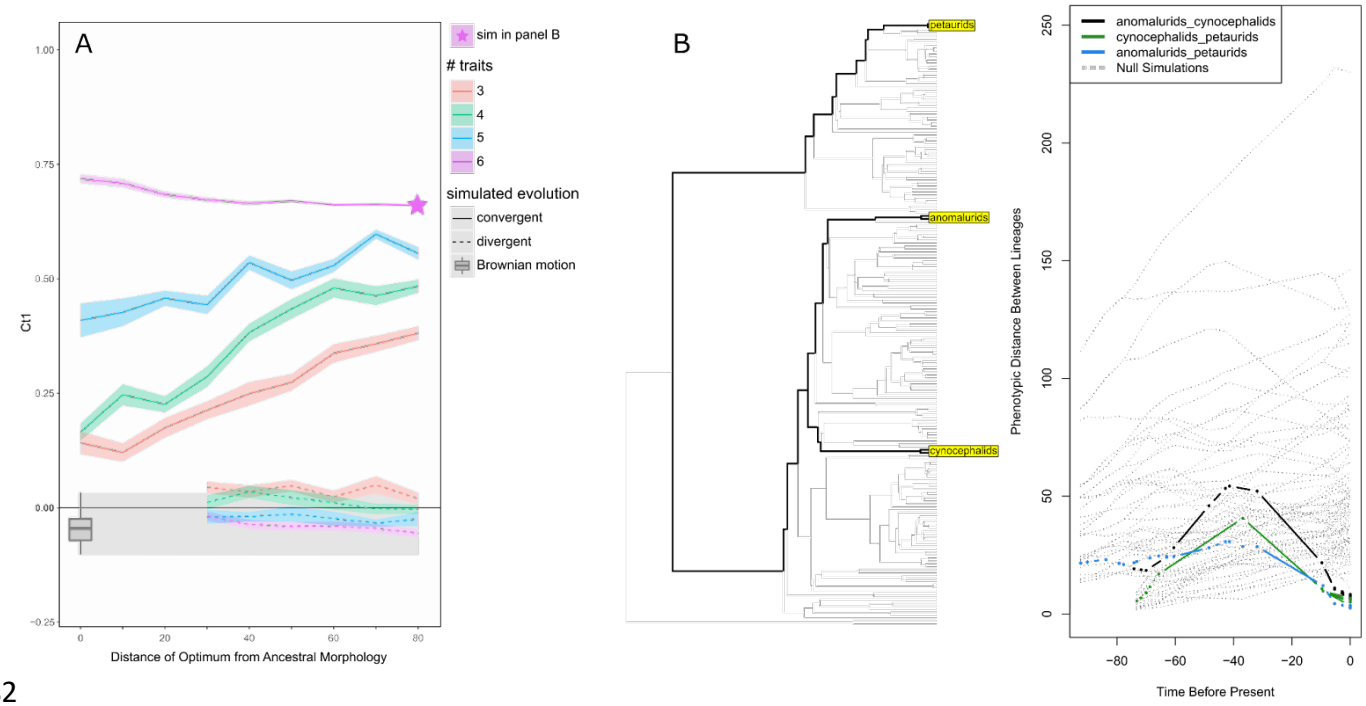

Figure 5. (*A*) *Ct*1 scores for simulated convergent lineages (top results in plot) and divergent lineages (bottom) under varying evolutionary scenarios. See the Methods and Figure 3 caption for more information. Although in some cases the divergence *Ct*1 results are greater than zero (indicating convergence), these results were not statistically significant when we calculated simulation-based *p*-values for a subset of datasets (Table 1). *C*1 results are means and standard errors of 15 simulated datasets. (*B*) An example output from the *plot.C R* function that shows the pairwise distances between lineages with time. Note that although only three 'glider' lineages are highlighted in the plot, five lineages were used for *Ct1* measurements.

We tested the performance of *Ct*-measures by applying them to the simulated datasets using the same methodology as that for *C*-measures. Unlike the *C*-measures (Fig. 3C), the *Ct*-measures do not consistently misidentify divergent lineages as being convergent (Figs. 5A and S3); most of the simulated divergence datasets (via both drift/BM and selection/OU) exhibit *Ct*1

scores that are negative, correctly indicating divergence. Although in some cases the *Ct*1 score is greater than zero when applied to divergence datasets (incorrectly indicating convergence; Fig. 5A), these *Ct*1 scores are not statistically significant (Table 1), which contrasts with the strongly significant *C*1 scores for divergence simulations. Further, the greater-than-zero *Ct*1 results could be due in part to convergence occasionally occurring by chance in our simulated-divergence datasets (e.g., BM-evolved 'glider' lineages evolving toward each other by chance). This indicates the importance of researchers considering the *p*-values associated with *Ct*-measures when evaluating convergence in their samples.

Although both *C*1 and *Ct*1 have a maximum value of one (complete convergence), the two measures scale differently and thus their values should not be directly compared. For instance, *C*1 values for 'constrained' divergence simulations are approximately 3.0 (Fig. 3C), whereas *Ct*1 values for the same divergence simulations are approximately 0.0 (Fig. 5A). Thus, positive *C*1 values reflect both convergence and some degree of divergence (due to the D_{max} issue highlighted in this study), and their values are expected to be greater than those of *Ct*1 for the same scenario. Further, limiting $D_{max.t}$ to specific time slices may dampen the convergence signal (see discussion on this topic below).

Relatively greater convergence in morphological outliers

Like *C*-measures and θ_{real}, the *Ct*-measures often measure more convergence when focal lineages are morphological outliers (Figs. 4A, 5A, and S8). This pattern appears to be more pronounced in *C*-measures compared to *Ct*-measures (Figs. 2A, 3C, and S8), likely due to the *C*-measure issues highlighted in this study (Fig. 2). Further, this trend may occur in *Ct*-measures

(despite the *C*-measure-related issues being eliminated) in part because values can be inflated
822 when focal taxa have different origination ages, and this issue is expected to be more
pronounced in morphological outliers than non-outliers (see Supplemental Results and Figure
824 S5). To help address this issue for *Ct*-measures, we added an optional feature to the *calcConvCt*
function that limits candidate $D_{\mathrm{max.t}}$ measurements to the time prior to the evolution of the
826 focal lineages (e.g., prior to the evolution of the earliest glider clade). We recommend that
researchers use this option as a supplement to regular *Ct*-measures when their clades of
828 interest have very different origination ages (see Supplemental Results).

The trend of greater convergence in outliers largely disappears for *C*- and *Ct*-measures
830 when all six simulated traits are OU-evolved to trait optima (i.e., all six traits are convergent) –
the *C*1 and *Ct*1 values remain relatively constant at any optimum values (Figs. 3C and 5A).
832 Conversely, the trend is more distinct when a greater number of traits are BM-evolved. This
indicates that the pattern of greater convergence in outliers is influenced by the types of
834 simulated traits (BM-evolved or OU-evolved to optima) and not just the aforementioned issues
with *C*- and *Ct*-measures. We believe that trait type influences this pattern because the relative
836 amount of morphological evolution in OU-evolved traits (vs BM-evolved traits) varies
depending on the value of trait optima. For instance, when OU-evolved traits are evolved
838 toward optima near the ancestral morphology, they may undergo less morphological change
relative to the BM-evolved traits. Thus, there is a relatively greater influence of BM-evolved
840 traits (i.e., divergent traits) on *C*- and *Ct*-measures at smaller optima values, resulting in weaker
convergence signals via *C*- and *Ct*-measures. Conversely, OU-evolved traits (i.e., convergent
842 traits) of outliers have undergone relatively greater morphological change, evolving farther

from the ancestral morphology, and therefore the OU-evolved traits have a relatively larger influence on *C*- and *Ct*-measures. This pattern is likely to reflect empirical scenarios because the strength of natural selection is expected to vary among traits (e.g., Grossnickle et al. 2020). Thus, when different traits of taxa have undergone varying amounts of adaptive morphological change, researchers should expect to observe relatively stronger evidence of convergence in outliers when using most convergence measures.

Empirical examples – *C*1 vs *Ct*1

The *C*-measure issues highlighted here are relevant to the many studies that have employed (or will employ) the *C*-measures. In many cases, erroneous *C*-measure results may have led researchers to either infer convergence in lineages that are divergent or infer inflated degrees of convergence. For instance, Grossnickle et al. (2020) tested for convergence among gliding mammal lineages using limb measurements and observed conflicting results. Statistically significant *C*-measure scores indicated strong convergence, but other analyses (i.e., evolutionary model-fitting, morphological disparity, phylomorphospace trajectories) suggested parallel evolutionary patterns. The authors concluded that the conflicting lines of evidence indicated weak, incomplete convergence. But considering the issues highlighted here, the strong *C*-measure results in Grossnickle et al. (2020) are probably misleading. For instance, the *C*-measure scores were likely inflated due to the outlying morphologies of some gliders (e.g., dermopterans) because outliers produce greater *C* scores (Figs. 2A and 3C), meaning that the gliders are probably less convergent than the authors concluded. We re-analyzed the data from Grossnickle et al. (2020) using the *Ct*-measures, and in contrast to strong *C*-measure scores, we

found that all glider comparisons have negative *Ct*1 scores, indicating divergence instead of
866 convergence. In some instances, the *Ct*1 scores are only slightly negative and have significant *p*-
values (e.g., *Ct*1 = -0.01 and $p < 0.01$ for the comparison of scaly-tailed squirrels and flying
868 squirrels), which is congruent with the other lines of evidence examined in Grossnickle et al.
(2020) that suggested parallel evolutionary changes rather than convergence for most glider
870 groups.

Huie et al. (2021) and Stayton (2015A; using data from Mahler et al. 2013)
872 independently analyzed *Anolis* lizard morphologies using distinct datasets, and both found that
the ecomorphotypes with the greatest *C*1 scores are those in the outermost regions of
874 morphospace (‘crown-giant,’ ‘grass-bush,’ and ‘twig’; see Figure 3 of Huie et al. 2021). The *C*1
values for these ecomorphotypes ranged from 0.31 to 0.43 in these studies, whereas other,
876 non-outlying ecomorphotypes had *C*1 values ranging from 0.09 to 0.25 (Stayton 2015A, Huie et
al. 2021). The relatively large *C*1 scores of outlying ecomorphotypes, in addition to the positive
878 *C*1 scores for all pairwise comparisons, may be due in part to the biases in the *C*-measures. We
evaluated this possibility by applying *Ct*-measures to one of the outlying ecomorphotypes
880 ('twig') from the anole dataset of Mahler et al. (2013; ten standardized measurements of body
dimensions, e.g., tail length). We found that, although the overall *Ct*1 score was statistically
882 significant, it was near zero (Table S1), in contrast to the *C*1 score being 0.36 (Stayton 2015A).
Interestingly, there was considerable disparity in the pairwise *Ct* results for the five twig
884 lineages, with *Ct*1 scores ranging from 0.346 (*A. paternus* vs. *A. valencienni*) to -0.763 (*A.
occultus* vs. *A. paternus*) and six of ten pairwise comparisons not significant. (See Figure S7 and
886 Table S1 for full results and plotted pairwise distances through time.) Thus, these results

highlight not only the issues with the *C*-measures, namely the inflation of *C* scores among outliers, but also the importance of considering pairwise comparisons when evaluating convergence among multiple focal lineages.

_Ct_-measures – recommendations and limitations

In contrast to *C*-measures, the *Ct*-measures are influenced by the timing of evolutionary change because they limit candidate $D_{max.t}$ measurements to specific time slices. This feature should be considered by researchers who apply the *Ct*-measures because it may alter expectations about the degree of measured convergence. For instance, if different focal lineages evolve toward a specific morphology (or adaptive peak) at different points in time, then the $D_{max.t}$ measurement may not measure the morphologically farthest distances between the lineages, possibly resulting in lower-than-expected *Ct* scores. Conversely, and as noted above, if the putatively convergent taxa evolve toward outlying regions of morphospace, then the asynchronous origins of the clades could inflate the *Ct*-measures (Supplemental Results; Fig. S5). To help mitigate this issue, we recommend that researchers generate and assess phylomorphospace and distances-between-lineages-through-time plots, and compare default *Ct* results to those generated when using the alternative option of the *calcConvCt* function that limits candidate $D_{max.t}$ measurements to the period in which lineages of interest overlap in time (see Supplemental Methods).

The *Ct*-measures can be applied to taxa of varying geologic ages (assuming there is an internal node for a $D_{max.t}$ measurement), but they may perform poorly when the tips of focal taxa are very different in geologic age because candidate $D_{max.t}$ measurements are restricted to

the period in which the evolutionary histories of the lineages overlap in time. For instance, the
910 evolutionary histories of ichthyosaurs and dolphins overlap from their most recent common ancestor (MRCA; early amniotes) to the ichthyosaur tips, providing considerable time (more
912 than 200 million years) for candidate $D_{max.t}$ measurements. However, there are no candidate $D_{max.t}$ measurements between the extinction of ichthyosaurs and extant dolphins, and if this is
914 the period in which the ancestral dolphin lineages shows the greatest divergence from ichthyosaurs, then $D_{max.t}$ will not capture the maximum divergence of the lineages. That is,
916 much of the evolutionary history of dolphins (and placental mammals more broadly) would be excluded by *Ct*-measures, likely resulting in smaller-than-expected *Ct* scores. (Note, however,
918 that D_{tip} ignores time and would measure the morphological distance between ichthyosaur and dolphin tips.) In these types of cases, alternative methods for measuring convergence (e.g., OU
920 model-fitting analyses) may be more appropriate.

The restriction of candidate $D_{max.t}$ measurements to coincide with internal nodes
922 exacerbates an issue inherent to many phylogenetic comparative methods: the reliance on inferred ancestral states. $D_{max.t}$ is the critical value that enables the *Ct*-measures to diagnose
924 convergence, and it is drawn entirely from ancestral state data, which are estimated from tip values assuming a BM model of evolution. The consequence is that ancestral reconstructions
926 are likely to reflect average morphologies of the sampled taxa, decreasing the chance of measuring convergence via the *Ct*-measures because $D_{max.t}$ estimates may be artificially shorter
928 than the 'real' $D_{max.t}$ values. For instance, despite simulating extremely strong convergence on a trait optimum for all six traits, the greatest *Ct*1 scores are around 0.7, indicating that about 70%
930 of $D_{max.t}$ has been closed by convergent evolution. Based on the simulation methods, we

expected *Ct*1 values to be closer to 1.0. We tested this issue by applying *Ct*1 (and *C*1) measures
932 to simulations in which we used the generated ('true') ancestral traits. *Ct*1 scores are consistently greater when 'true' ancestral states are used in calculations (Fig. S8), helping to
934 confirm that the convergence signal of *Ct*-measures might often be diluted due to the assumption of a BM model of evolution. Note that because *C*-measures rely on the same
936 ancestral state reconstructions to infer morphology at internal nodes, this issue is not unique to the *Ct*-measures (Fig. S8).

938 The aforementioned issue is likely to be exacerbated under conditions where there are relatively few intervening nodes between putatively convergent lineages (i.e., there is a small
940 sample of candidate $D_{max.t}$ measurements), when those putatively convergent lineages are subtended by long branches (i.e., distances from which to draw $D_{max.t}$ are biased toward deeper
942 nodes), and when only contemporary tips are included (i.e., there is a lack of fossil data informing reconstructions at internal nodes). Therefore, the *Ct*-measures may be most
944 appropriate for well-sampled study systems that include a substantial number of internal nodes and relatively few long branches, and researchers should include fossil taxa whenever possible
946 to improve ancestral reconstructions at internal nodes.

The number of phenotypic traits used to assess convergence is likely of increased
948 importance when using *Ct*-measures. In multivariate datasets, some traits may be convergent and others non-convergent (i.e., divergent, parallel, or conservative). While including a greater
950 number of non-convergent traits in analyses is expected to decrease the overall convergence signal of any convergence measure, it may also exacerbate the *Ct*-related issues raised in this
952 section. In general, adding traits increases the measured distances between tips and internal

nodes. However, ancestral inference via BM tends to average variation at internal nodes; thus,
954 D_{tip} typically increases at a higher rate than $D_{max.t}$ for each non-convergent trait that is added to
a dataset. This pattern is illustrated in Figure S6, highlighting that increasing the number of BM-
956 evolved traits (which are expected to be mostly divergent) in simulations results in relatively
greater increases of D_{tip} scores compared to $D_{max.t}$ scores. Therefore, an increased number of
958 traits in analyses (with all else equal) could result in a relative decrease in *Ct* scores compared
to datasets with fewer traits, unless the additional traits are strongly convergent. We
960 recommend that researchers carefully choose traits based on the specific hypothesis being
tested, and analyze individual traits or subsets of traits whenever feasible to tease apart unique
962 patterns among traits.

964 **Summary**

The *C*-measures are a popular means of identifying and quantifying phenotypic convergence, in
966 part because they can differentiate between convergence and conservatism. However, we
highlight a critical issue with *C*-measures: they can misidentify divergent lineages as convergent
968 (Figs. 2 and 3, Table 1). To help address this issue, we developed improved convergence
measures (*Ct*-measures) that quantify distances between lineages at time slices at internal
970 phylogenetic nodes, minimizing the possibility of divergent taxa being mistakenly measured as
convergent. We have also developed new features (available in the *convevol R* package), such
972 as a function to produce distances-between-lineages-through-time plots and the ability to
compare clades that include multiple taxa. Although *Ct*-measures improve on *C*-measures,
974 researchers should recognize their limitations. For instance, *Ct*-measures may be unreliable if

convergent evolutionary change is asynchronous between lineages of interest (e.g., fossils of very different geologic ages), especially when lineages are morphological outliers. More broadly, we find that multiple methods (including *Ct*-measures) are biased by the location of taxa in morphospace, with most methods measuring greater convergence in morphological outliers. OU model-fitting analyses are a valuable tool for investigating convergence, but we urge caution when applying and interpreting the results of the analyses. Model support for multiple-regime OU models is not always indicative of convergent shifts toward shared adaptive peaks, especially when all of the fitted models are poor fits to the data (Table 2). Because all available methods for identifying and measuring phenotypic convergence are imperfect, we recommend that researchers use multiple methods. Researchers should also recognize the benefits and drawbacks of their chosen methods when interpreting results, and they should rely on methods that are most appropriate for their dataset and the hypotheses being tested.

ACKNOWLEDGEMENTS

For helpful discussion and feedback, we thank two anonymous reviewers, Miriam Zelditch, Gregory Wilson Mantilla, Eli Amson, Kenneth-Dieter Benton, Allison Bormet, Richard Meyn, Sharlene Santana, and members of the 'Tempo & Mode in Isolation' discussion group. Jonathan S. Mitchell wrote the *R* script (published in Zelditch et al. 2017) that we adapted for the updated *convevol* functions, and we greatly appreciate his permission to use the code. Funding for D. M. G. was provided by National Science Foundation (NSF) grants DBI-1812126 and IOS-2017738. C. J. L. was funded by an Early Career Provost Fellowship at the University of Texas at

Austin, the NSF (DBI−1906248 and DBI–2128146), the Gerstner Family Foundation, the Richard Gilder Graduate School, and the Department of Mammalogy at the American Museum of Natural History. L. N. W. was funded by the NSF Graduate Research Fellowship and NSF grant EAR-PF 2052992.

990

LITERATURE CITED

992 Adams, D. C., and M. L. Collyer. 2018. Multivariate phylogenetic comparative methods: Evaluations, comparisons, and recommendations. Syst. Biol. 67:14-31.

994 Akaike, H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19:716–723.

996 Alfieri, F., L. Botton-Divet, J. A. Nyakatura, and E. Amson. 2022. Integrative approach uncovers new patterns of ecomorphological convergence in slow arboreal xenarthrans. J. Mamm. 998 Evol. 29:283–312.

Arbour, V. M., and L. E. Zanno. 2020. Tail Weaponry in Ankylosaurs and Glyptodonts: An 1000 Example of a Rare but Strongly Convergent Phenotype. Anat. Rec. 303:988–998.

Arbuckle, K., C. M. Bennett, and M. P. Speed. 2014. A simple measure of the strength of 1002 convergent evolution. Methods Ecol. Evol. 5:685–693.

Arbuckle, K., and Minter, A. 2015. Windex: Analyzing convergent evolution using the 1004 Wheatsheaf index in R. Evol. Bioinform. 11:EBO-S20968.

Baliga, V. B., and R. S. Mehta. 2018. Phylo-allometric analyses showcase the interplay between 1006 life-history patterns and phenotypic convergence in cleaner wrasses. Am. Nat. 191:E129-E143.

1008 Bastide, P., C. Ané, S. Robin, and M. Mariadassou. 2018. Inference of Adaptive Shifts for Multivariate Correlated Traits. Syst. Biol. 67:662–680.

1010 Beaulieu, J. M., D.-C. Jhwueng, C. Boettiger, and B. C. O'Meara. 2012. Modeling stabilizing selection: Expanding the Ornstein-Uhlenbeck model of adaptive evolution. Evolution 66:2369–2383.

1012

Bennion, R. F., J. A. MacLaren, J. A., E. J. Coombs, F. G. Marx, O. Lambert, and V. Fischer. 2022. Convergence and constraint in the cranial evolution of mosasaurid reptiles and early cetaceans. Paleobiology (First View, https://doi.org/10.1017/pab.2022.27).

1014

1016 Butler, M. A., and A. A. King. 2004. Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. Am. Nat. 164:683–695.

1018 Button, D. J., and L. E. Zanno. 2020. Repeated evolution of divergent modes of herbivory in non-avian dinosaurs. Curr. Biol. 30:158–168.

1020 Baumgart, S. L., P. C. Sereno, and M. W. Westneat. 2021. Wing shape in waterbirds: morphometric patterns associated with behavior, habitat, migration, and phylogenetic convergence. Integr. Org. Biol. 3:obab011.

1022

Canale, J. I., Apesteguía, S., Gallina, P.A., Mitchell, J., Smith, N.D., Cullen, T.M., Shinya, A., Haluza, A., Gianechini, F.A., Makovicky, P.J. 2022. New giant carnivorous dinosaur reveals convergent evolutionary trends in theropod arm reduction. Curr. Biol. 32:3195–3202.

1024

1026 Castiglione, S., G. Tesone, M. Piccolo, M. Melchionna, A. Mondanaro, C. Serio, M. De Febbraro, and P. Raia. 2018. A new method for testing evolutionary rate variation and shifts in phenotypic evolution. Methods. Ecol. Evol. 9:974–983.

1028

Castiglione, S., C. Serio, D. Tamagnini, M. Melchionna, A. Mondanaro, M. De Febbraro, A. Profico, P. Piras, F. Barattolo, P. Raia. 2019. A new, fast method to search for morphological convergence with shape data. PLOS One 16:e0252264.

Chen, M., and G. P. Wilson. 2015. A multivariate approach to infer locomotor modes in Mesozoic mammals. Paleobiology 41:280–312.

Clavel, J., G. Escarguel, and G. Merceron. 2015. mvMORPH: an R package for fitting multivariate evolutionary models to morphometric data. Methods Ecol. Evol. 6:1311–1319.

Collar, D. C., J. S. Reece, M. E. Alfaro, P. C. Wainwright, and R. S. Mehta. 2014. Imperfect morphological convergence: variable changes in cranial structures underlie transitions to durophagy in moray eels. Am. Nat. 183:E168–E184.

Cooper, N., Thomas, G. H., Venditti, C., Meade, A., & Freckleton, R. P. (2016). A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies. Biol. J. Linn. Soc. Lond. 118:64–77.

Da Silva, F. O., A. C. Fabre, Y. Savriama, J. Ollonen, K. Mahlow, A. Herrel, J. Müller, and N. Di-Poï, N. 2018. The ecological origins of snakes as revealed by skull evolution. Nat. Commun. 9:1–11.

Darwin, C. 1859. *On the Origin of Species by Means of Natural Selection*. John Murray, London.

Drummond, A. J., M. A. Suchard, D. Xie, and A. Rambaut, 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:1969–1973.

Friedman, S. T., S. A. Price, A. S. Hoey, and P. C. Wainwright. 2016. Ecomorphological convergence in planktivorous surgeonfishes. J. Evol. Biol. 29:965–978.

1050 Grabowksi, et al. 2023. A cautionary note on "A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies". Syst. Biol. syad012, 1052 https://doi.org/10.1093/sysbio/syad012.

Grossnickle, D. M. 2020. Feeding ecology has a stronger evolutionary influence on functional 1054 morphology than on body mass in mammals. Evolution 74:610–628.

Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 1056 51:1341–1351.

Ho, L. S. T., and Ané, C. 2014. Intrinsic inference difficulties for trait evolution with Ornstein-1058 Uhlenbeck models. Methods Ecol. Evol. 5:1133-1146.

Huie, J. M., I. Prates, R. C. Bell, and K. de Queiroz. 2021. Convergent patterns of adaptive 1060 radiation between island and mainland *Anolis* lizards. Biol. J. Linn. Soc. Lond. 134:85–110.

Hurvich, C. M., and C. L. Tsai. 1989. Regression and time series model selection in small 1062 samples. Biometrika 76:297–307.

Ingram, T., and D. L. Mahler. 2013. SURFACE: detecting convergent evolution from comparative 1064 data by fitting Ornstein-Uhlenbeck models with stepwise Akaike Information Criterion. Methods Ecol. Evol. 4:416–425.

1066 Khabbazian, M., Kriebel, R., Rohe, K. and Ané, C. 2016. Fast and accurate detection of evolutionary shifts in Ornstein-Uhlenbeck models. Methods Ecol. Evol. 7:811–824.

1068 Law, C. J., Slater G. J., and Mehta, R. S. 2018. Shared extremes by ectotherms and endotherms: Body elongation in mustelids is associated with small size and reduced limbs. Evolution. 1070 73:735–749.

Law, C. J. 2022. Different evolutionary pathways lead to incomplete convergence of elongate
1072 body shapes in carnivoran mammals. Syst. Biol. 71:788-796.

Losos, J. B. 2011. Convergence, adaptation, and constraint. Evolution 65:1827–1840.

1074 Mahler, D. L., T. Ingram, L. J. Revell, and J. B. Losos. 2013. Exceptional convergence on the macroevolutionary landscape in island lizard radiations. Science 341:292–295.

1076 Mahler, D. L., M. G. Weber, C. E. Wagner, and T. Ingram. 2017. Pattern and process in the comparative study of convergent evolution. Am. Nat. 190:S13–S28.

1078 Martinez, Q., J. Clavel, J. A. Esselstyn, A. S. Achmadi, C. Grohé, N. Pirot, and P. H. Fabre. 2020. Convergent evolution of olfactory and thermoregulatory capacities in small amphibious
1080 mammals. Proc. Natl. Acad. Sci. 117:8958–8965.

McLean, B. S., K. M. Helgen, H. T. Goodwin, and J. A. Cook. 2018. Trait-specific processes of
1082 convergence and conservatism shape ecomorphological evolution in ground-dwelling squirrels. Evolution 72:473–489.

1084 Pevsner, S. K., D. M. Grossnickle, and Z. X. Luo. 2022. The functional diversity of marsupial limbs is influenced by both ecology and developmental constraint. Biol. J. Linn. Soc. Lond.
1086 135:569–585.

Polly, P. D. 2019. Phylogenetics for Mathematica. Version 6.5. Department of Earth and
1088 Atmospheric Sciences, Indiana University: Bloomington, Indiana. https://pollylab.indiana.edu/software.html.

1090 R Core Team 2020. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

1092 Revell, L. J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. 2:217–223.

1094 Rincon-Sandoval, et al. (2020). Evolutionary determinism and convergence associated with water-column transitions in marine fishes. Proc. Natl. Acad. Sci. 117:33396–33403.

1096 Rovinsky, D. S., A. R. Evans, and J. W. Adams. 2021. Functional ecological convergence between the thylacine and small prey-focused canids. BMC Ecol. Evol. 21:1–17.

1098 Serio, C., P. Raia, and C. Meloro. 2020. Locomotory adaptations in 3D humerus geometry of Xenarthra: testing for convergence. Front. Ecol. Evol. 8:139.

1100 Spear, J. K., and S. A. Williams. 2020. Mosaic patterns of homoplasy accompany the parallel evolution of suspensory adaptations in the forelimb of tree sloths (Folivora: Xenarthra).

1102 Zool. J. Linn. Soc. zlaa154.

Speed, M. P., and K. Arbuckle. 2017. Quantification provides a conceptual basis for convergent

1104 evolution. Biol. Rev. Camb. Philos. Soc. 92:815–829.

Stayton, C. T. 2008. Is convergence surprising? An examination of the frequency of convergence

1106 in simulated datasets. J. Theor. Biol. 252:1–14.

Stayton, C. T. 2015A. The definition, recognition, and interpretation of convergent evolution,

1108 and two new measures for quantifying and assessing the significance of convergence. Evolution 69:2140–2153.

1110 Stayton, C. T. 2015B. What does convergent evolution mean? The interpretation of convergence and its implications in the search for limits to evolution. Interface Focus

1112 5:20150039.

Stayton C. T. 2018. *Convevol: quantifies and assesses the significance of convergent evolution.* R package version 1.3. https:cran.r-project.org/package=convevol.

Tamagnini, D., C. Meloro, P. Raia, and L. Maiorano. 2021. Testing the occurrence of convergence in the craniomandibular shape evolution of living carnivorans. Evolution 75:1738–1752.

Upham, N., J. A. Esselstyn, and W. Jetz. 2019. Inferring the mammal tree: species-level sets of phylogenies for questions in ecology, evolution, and conservation. PLOS Biol. 17:e3000494.

Uyeda, J. C. and L. J. Harmon. 2014. A novel Bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data. Syst. Biol. 63:902–918.

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

Weaver, L. N., and D. M. Grossnickle. 2020. Functional diversity of small-mammal postcrania is linked to both substrate preference and body size. Curr. Zool. 66:539–553.

Zelditch, M. L., J. Ye, J. S. Mitchell, and D. L. Swiderski. 2017. Rare ecomorphological convergence on a complex adaptive landscape: Body size and diet mediate evolution of jaw shape in squirrels (Sciuridae). Evolution 71:633–649.