A cautionary note on quantitative measures of phenotypic convergence

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

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

*Correspondence: dmgrossn@uw.edu

ABSTRACT

1. 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 in phylomorphospace, and Ornstein-Uhlenbeck evolutionary model-fitting analyses to test whether lineages have convergently evolved toward adaptive peaks.
2. We test the performance of *C*-measures and other convergence measures under various evolutionary scenarios. We reveal critical issues with *C*-measures, which we help to address by developing novel convergence measures (*Ct*1–*Ct*4-measures) that measure distances between lineages at specific points in time.

3. The most substantial issue with *C*-measures is that they will often misidentify divergent lineages as convergent; this is most common when focal taxa are morphological outliers. In contrast, our new *Ct*-measures minimize the possibility of 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 all or most of the evolutionary histories of the lineages overlap in time. Beyond *C*-measures, we demonstrate issues with other convergence measures. We find that all distance-based convergence measures are influenced by the position of putatively convergent taxa in morphospace, with morphological outliers often statistically more likely to be categorized as convergent by chance. Further, we demonstrate that multiple-regime Ornstein-Uhlenbeck models often outperform simpler models when fit to divergent lineages, highlighting that model support for multiple-regime models should not always be assumed to reflect convergence among focal lineages.
4. The issues with convergence measures highlighted here are especially relevant because they influence the degree of inferred convergence in many past studies, raising the concern that many lineages have been mistakenly identified as convergent. Our new convergence measures provide researchers with an improved comparative tool for future studies. Nonetheless, we emphasize that all available convergence measures are imperfect, and researchers should recognize the limitations of these methods and use multiple lines of evidence when inferring and measuring convergence.

KEYWORDS: adaptive evolution, convergent evolution, evolutionary models, Ornstein-Uhlenbeck models, phylomorphospace

INTRODUCTION

Phenotypic convergence among distantly related taxa is commonly associated with adaptive 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). 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 tied to similar ecological or functional roles. Thus, quantitative examinations of phenotypic convergence are important; they assist researchers in identifying adaptive morphological changes that are driven by shared selective pressures and/or developmental constraints. Novel methods for identifying and measuring convergence have recently been developed (Mahler et al. 2013, 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 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 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 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. 2021, Rovinsky et al. 2021, Tamagnini et al. 2021, Alfieri et al. 2022, Bennion et al. 2022, Canale et al. 2022, Law 2022).

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 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 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 morphology (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 phenotypic 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 phylomorphospace 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
100 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
102 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 –
104 (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
106 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
108 *C*1 score of 0.7.

One reason for the popularity of *C*-measures is that they can distinguish between
110 convergence and conservatism, which both result in distantly-related phylogenetic tips
with similar phenotypes. The key difference between convergence and conservatism
112 centers on the ancestral morphologies of the lineages. Whereas convergence involves
ancestors that were less morphologically similar to each other than their descendant
114 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
116 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
118 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-
120 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 test for 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), and this attraction and any resulting convergence is often assumed to be due to selective pressures toward adaptive peaks (Fig. 1B). Convergence is identified when the best-supported model indicates that two or more focal lineages have independently evolved toward the same trait optimum. Convergence can be tested by examining model support for multiple-regime OU models in which focal taxa are classified into regimes *a priori* or by examining 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, or with other *R* packages designed to identify peak shifts (e.g., *Bayou*, Uyeda and Harmon 2014; *l1ou*, Khabbazian et al. 2016; *phyloEM*, Bastide et al. 2018). However, an OU process does not require that independent lineages shifting toward a peak come from more dissimilar ancestors and

thus may not meet the criteria of convergence used in this paper, and interpreting
146 convergence from model support for multiple-regime OU models should be made with caution because OU models are prone to overfitting to data (Cooper et al. 2016, Adams
148 and Collyer 2018). Further, when focal taxa exhibit the ancestral morphology, it may be challenging to use OU model-fitting analyses to determine whether the lineages exhibit
150 the morphology due to convergence or conservatism; both scenarios reflect expectations of an OU process (i.e., selection toward a shared morphology).
152 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,
154 therefore providing valuable information about the evolutionary process that may be driving convergence.

156 Here, we test the performances of *C*-measures and other convergence measures by applying the methods to simulated data in which a subset of lineages are modeled as
158 truly convergent or truly divergent. Our results highlight a critical concern with the *C*-measures: in some circumstances they may lead to misclassification of divergent
160 lineages as convergent, and this is more common when those lineages are outliers in morphospace. We find that θ (i.e., the angle between phenotypic vectors; Castiglione et
162 al. 2019) is also biased toward misclassifying morphological outliers as convergent, whereas the Wheatsheaf index (Arbuckle et al. 2014, Arbuckle and Minter 2015) is
164 biased but in the opposite direction, indicating greater convergence among lineages that retain their shared ancestral morphology. Further, we find that multiple-regime OU
166 models can outperform simpler models as fits to divergent lineages, highlighting that model support for multiple-regime OU models should not be assumed to always be

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

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 divergence' of 'gliders' via selection (OU-evolved traits, limited to positive values), and 4) 'unconstrained divergence' of focal taxa via selection (OU-evolved traits;

212 no limits on traits). The following text describes these four types of simulations, and Figure S1 provides a graphical summary.

214 *Convergence simulations.* For the convergence simulations, we systematically altered 1) the number of traits (of the six total traits) that were subjected to convergent 216 selection (three 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 218 optimum (target). The convergent traits were subjected to directional OU-selection for their full branch lengths, which allowed most 'glider' tips (besides those with the shortest 220 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 222 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 224 being convergent thus represents very strong convergence, whereas three of six traits being convergent represents weaker convergence. Further, we alter the number of 226 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 228 bone lengths versus long bone widths in Grossnickle et al. 2020).

By changing the values of trait optima ('targets') we systematically controlled the 230 morphological distance of 'gliders' from their ancestral condition, allowing us to test how morphospace position influences the results of applied convergence measures (Fig. 232 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 234 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 unrealistically far from ‘non-gliders’ in morphospace and thus unlikely to reflect empirical scenarios, but we include them here to help infer trends in how morphospace position 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 15 replicates.

Divergence simulations. To simulate divergence via drift (i.e., BM-evolved ‘gliders’), we 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 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 contrast to all other simulations, gliders in these datasets did not possess any OU-evolved traits.

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. 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 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 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 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 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 selected lineages are allowed to fully reach their trait optima.

For our primary selection-based divergence simulations, we limited the OU-evolved ‘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-quadrant) of multidimensional 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 geometrically divergent (e.g., Grossnickle et al. 2020, Collar et al. 2014). Further, researchers are most likely to apply convergence measures to lineages that share some similarities, making these simulations more applicable to empirical research. Nonetheless, we also simulated ‘unconstrained divergence’ for a smaller subset of evolutionary scenarios (see discussion on subset analyses below) in which the ‘glider’ trait optimum values were not limited to positive values. The results of convergence measures applied to these simulations are provided in the Supplemental Results.

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

***C*-measures**

We applied the *C*-measures (Stayton 2015A) to focal lineages ('gliders') in the simulated datasets. The primary measure, *C*1, is the distance between phylogenetic tips of focal taxa (D_{tip}) divided by the maximum distance between any tips or ancestral nodes of those lineages (D_{max}). The resulting proportion is subtracted from one; the *C*1 value is one for complete 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 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). 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 functions in 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 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).

C-measures quantify phenotypic convergence between individual phylogenetic tips, not between clades with multiple tips (Stayton 2015A). Thus, we calculated average phenotypes for 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 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 were not averaged prior to the other types of convergence analyses described below.

Additional measures of convergence

Subset of simulated datasets. We applied additional measures of convergence (OU model-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 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 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 (Grossnickle et al. 2020). Nonetheless, four of six traits being convergent on an optimum often results in strong convergence (i.e., statistically significant distance-based measures of 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, 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 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 morphospace (optima = 40, 60, and 80). For selection-based divergence

simulations (both ‘constrained’ and ‘unconstrained’), we used the 15 simulated datasets
328 with optima of 40, 60, and 80. (Using trait optima of 0 or 20 could mistakenly simulate
convergence toward ancestral morphologies.) Thus, the subset of datasets includes 75
330 convergence simulations (15 datasets each for five trait optima) and 45 simulations
each for unconstrained and constrained divergence simulations (15 datasets each for
332 three optima). The following methods were only applied to this subset of datasets.

Evolutionary model-fitting analyses. We fit four multivariate models to all six
334 simulated traits using functions within the *mvMORPH R* package (Clavel et al. 2015).
The first three models were a single-rate multivariate BM model (mvBM1) that assumes
336 trait variance accumulates stochastically but proportionally to evolutionary time, a
single-optimum Ornstein-Uhlenbeck model (mvOU1) that modifies the BM model to
338 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. Model
340 support for any of these three models would indicate a lack of strong convergence
among the taxa simulated as convergent or divergent, due to the lack of evidence for a
342 distinct adaptive peak associated with ‘gliders.’ In all model-fitting analyses, the mvEB
model collapsed to a mvBM model (i.e., the change-in-rate parameter was zero). Thus,
344 we did not report the mvEB results, and we did not include it when calculating Akaike
weights.

346 We then fit a multivariate OU model with two selective regimes (mvOU2) that
allowed ‘gliders’ and ‘non-gliders’ to exhibit different trait optima (θ). Support for mvOU2
348 would provide evidence of convergence by indicating that selective forces are driving
‘glider’ lineages to a shared adaptive peak (Fig 1B). Note that in the simulations 'non-

350 gliders' evolved via BM, and thus any support for the mvOU2 model is likely to be driven by the 13 ‘glider’ lineages. Relative support for each of the three models was assessed
352 through computation of small-sample corrected Akaike weights (AICcW; Akaike 1974; Hurvich and Tsai 1989). We calculated AICcW for each of the 15 trees, and we report
354 the mean values for these trees. Prior to fitting the mvOU2 model, we mapped the gliding regime character state onto the ‘glider’ branches of the phylogeny (including the
356 subtending branches below ‘glider’ clades) using the *paintSubTree* function in the *phytools R* package (Revell 2012).

358 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
360 fitting 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
362 to fit multiple-regime OU models with varying evolutionary rates (σ) and/or attractions to optima (α) between regimes, which is not a feature of the multivariate *mvMORPH*
364 models. See the Supplemental Methods for additional information.

Additional distance-based convergence measures. We applied two other
366 measures of convergence to the subset of 30 simulated datasets (using all six traits): Wheatsheaf index, which was implemented via the *R* package *windex* (Arbuckle et al.
368 2014, Arbuckle and Minter 2015), and θ_{real}, which was implemented using the *RRphylo* package (Castiglione et al. 2018, Castiglione et al. 2019). The Wheatsheaf index
370 measures pairwise morphological distances between putatively convergent taxa, with distances corrected for the degree of phylogenetic relatedness of lineages. These
372 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); taxa evolving to trait optima farther from

396 the ancestral morphology have greater *C*1–*C*4 scores. The only exception is the *C*1 set of results when all six simulated traits are convergent. In this case, *C*1 scores remain
398 consistently around 0.8 regardless of the position of trait optima (Fig. 3C).

The second and third issues with the *C*-measures are more problematic:
400 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
402 illustrate lineages that are diverging morphologically (in univariate and multivariate morphospace), but they have a *C*1 score of 0.3, incorrectly suggesting that the lineages
404 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
406 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'
408 simulations (Fig. 3C; Table 1), incorrectly indicating strong convergence instead of divergence.

410 Similarly, outlying lineages evolving along parallel phylomorphospace trajectories from a similar ancestral condition have extremely strong *C*1 scores (Fig. 2C). This is
412 unexpected because the ancestral nodes of both lineages are the same morphological distance from one another as the distance between tips; this is not convergence
414 according to the definition of convergence adopted here or in Stayton 2015A (see Introduction; Losos 2011, Mahler et al. 2017).

416 Although the second and third issues with *C*-measures may be more common when at least one focal lineage has diverged rapidly from other focal taxa and is likely to
418 be a morphological outlier (e.g., see examples in Figures 2B and 2C), 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 morphospace. Thus, these issues have major implications for many empirical studies (see discussion of examples below); divergent or parallel lineages may often be incorrectly interpreted to be convergent, no matter their position in morphospace.

The possibility of diverging and parallel lineages having *C*1 scores that incorrectly 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 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 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 contrast, D_{max} is expected to be the morphological distance between the tips, meaning that D_{max} equals D_{tip} (and *C*1 = 0). However, this is not always the case; divergent lineages can have a 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 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 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.

Other measures of convergence show biased results

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).

Distance-based convergence measures. Consistent with the *C*-measures, the θ_{real} results (standardized to phylogenetic distance between clades) indicate greater convergence in 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 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 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 S3), 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 convergence in outliers relative to non-outliers; values decrease when convergent taxa are farther from the ancestral morphology in morphospace. The Wheatsheaf index compares the distances between putatively convergent taxa to distances between other tips. Our simulations did not allow all convergent lineages to completely reach trait optima (Fig. 3A), and for the subset of datasets used for Wheatsheaf index analyses, two of the six simulated traits evolved via BM. Together, these two factors mean that simulated convergent lineages did not completely converge on a morphology, and the pairwise distances between many tips of simulated

convergent lineages are farther apart from each other than are the pairwise distances between other, BM-evolved lineages (see the phylomorphospace in Figure 3A, but note that the plot is for data in which all six traits were convergent). If we allowed simulated lineages to completely reach trait optima, then this trend of less convergence in outliers (as measured by the Wheatsheaf index) might disappear. However, complete convergence on morphologies seems especially rare in empirical datasets (Grossnickle et al. 2020); thus, we believe that the Wheatsheaf index is likely to show reduced measures of convergence in morphological outliers of most empirical samples, in line with our simulation results.

Evolutionary model-fitting analyses. 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 BM-evolved and four traits OU-evolved), meaning that we do not expect any of the four tested models (mvBM1, mvOU1, mvEB, mvOU2) to be good fits to the data because they do not include a mix of BM and OU processes. Nonetheless, we fit these models because our goal is to mimic how researchers are likely to test for convergence if given the simulated morphological data without prior knowledge of the underlying evolutionary processes (e.g., Friedman et al. 2016, Baliga and Mehta 2018, Grossnickle et al. 2020, Martinez et al. 2020, Rincon-Sandoval et al. 2020). We have assigned the 'gliders' to one regime and the 'non-gliders' to a second regime, and then use model support for multiple-regime OU models (e.g., mvOU2 in this study) as evidence for convergence (Fig. 1B). Support for multiple-regime OU models could suggest that lineages of one or both regimes have convergently shifted toward a shared adaptive peak.

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 outperformed simpler 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 model (AICcW for mvOU2 is 1.0 when ‘glider’ trait optima are 40, 60, and 80; Table 2). These results are also supported by the univariate model-fitting results, with the two-regime OU2VA model being the best-fitting model (Table S1). This indicates evidence of two adaptive peaks, one for ‘gliders’ and one for ‘non-gliders,’ even though neither of those groups were simulated as evolving toward an adaptive peak.

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, may simply reflect that two-regime OU models are the best-fitting of bad-fitting models. Further, multiple-regime OU models are often incorrectly favored over simpler models (Ho and Ané 2014, Cooper et al. 2016), especially when sample sizes are small, and this may be the case

with our divergence datasets. The strength of an OU process can be assessed by
512 examining the phylogenetic half-lives (ln(2)/α) of traits (Ho and Ané 2014, Cooper et al.
2016). To examine half-lives for our datasets, we used the univariate, two-regime OU
514 models (via the *OUwie R* package) because, unlike the multivariate models (via the
mvMORPH R package), they permit the α value parameter to vary between regimes
516 (see Supplemental Methods), which then allows us to calculate the phylogenetic half-life
specifically for the 'glider' regime. For simulations with divergent trait optima of 40, the
518 fitted α value (mean of 15 trees) of the best-fitting univariate model (OU2VA; Table S1)
to PC1 scores indicates a phylogenetic half-life for the simulated 'gliders' of 46 million
520 years. Four of the five glider clades originated less than 46 million years ago. Thus, the
relatively long half-life suggests an especially poor fit of the OU2VA model to the data,
522 despite this model being a better fit than the BM1 and OU1 models according to the
AICcW comparisons.

524 Considering that empirical datasets often include complex evolutionary patterns
and small sample sizes for some regimes, researchers should be cautious both when
526 choosing models to fit to data and when interpreting results (Cooper et al. 2016).
Although not explored in this study, multiple-regime BM (BMM) models may offer
528 alternative options that complement multiple-regime OU models (e.g., Grossnickle et al.
2020, Martinez et al. 2020, Rincon-Sandoval et al. 2020). BMM models allow varying
530 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
532 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.,

534 Grossnickle et al. 2020), and in some cases a BMM model might serve as a more appropriate null model than BM1.

536 An additional OU model-fitting approach involves testing for convergent shifts toward shared adaptive peaks without *a priori* regime classifications for focal lineages 538 (e.g., using 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 540 datasets, especially 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 542 our method (with *a priori* regime classifications for 'gliders' and 'non-gliders') for testing the specific hypothesis of this study. Nonetheless, this method provides a valuable 544 research tool for examining convergence, especially for analyses that are exploratory (i.e., without hypotheses for specific lineages), and we recommend applying it to 546 empirical datasets as a supplement to other convergence measures (e.g., Grossnickle et al. 2020, Rincon-Sandoval et al. 2020).

548 Our goal with divergence simulations was to mimic empirical samples with lineages that might share some similarities and thus be tested for convergence. Thus, in 550 our primary divergence simulations (i.e., 'constrained divergence'), the divergent focal lineages ('gliders') were partially constrained to one region of morphospace (e.g., all 552 divergent lineages are in negative PC2 space in Figure 3B) because we limited the randomized optima to be positive values for the four OU-evolved divergent traits (the 554 other two traits were evolved by BM and could be negative; see Methods). This methodological decision helps to explain the relatively strong fits of the two-regime OU 556 models to the simulated divergence datasets. Because 'glider' lineages are shifting in

similar directions (e.g., all divergent lineages are in negative PC2 space in Figure 3B),
558 the ‘gliders’ may be modeled as evolving toward an especially broad adaptive peak that occupies a large region of one side of morphospace, especially because an OU process
560 does not require that independent lineages shifting toward a peak come from more dissimilar ancestors. Figure 2B provides a conceptual illustration of this scenario; the
562 focal lineages (‘maroon’) are diverging but still evolved toward a broad adaptive peak, which is broader than the adaptive peak of the ancestral (‘blue’) lineages. This could be
564 a similar scenario to the divergent outlier lineages in our simulated dataset (Fig. 3C).

For ‘unconstrained divergence’ simulations, we allowed the ‘gliders’ to diverge in
566 all directions in morphospace, and all convergence measures, including OU model-fitting analyses, correctly identify the taxa as being divergent (Supplemental Results;
568 Table S3). Thus, the issues highlighted here for OU model-fitting analyses are likely limited to scenarios in which focal taxa share some morphological similarities.

570 Our results highlight a second potential issue with using the fits of multiple-regime OU models to measure convergence: OU model-fitting analyses may struggle to
572 identify convergence when lineages converge on a morphology that is similar to the ancestral morphology (Table 2). This is reflected by the strong performance of mvBM1
574 (a uniform BM model) when optima are zero (AICcW = 0.991). Similarly, model-fitting analyses may be biased toward measuring stronger convergence when lineages evolve
576 farther from the ancestral morphology, consistent with *C*-measures and standardized θ_{real}. This is demonstrated by model support for mvOU2, which strengthens when
578 lineages evolve farther from the ancestral morphology (Table 2).

However, the aforementioned issue can be addressed in part by fitting more appropriate multiple-regime OU models to the data, such as models that allow for parameters to vary between regimes. This is demonstrated by our univariate results; the OU2VA allows the selection and rate parameters to vary between regimes, and it is the best-fitting model to PC1 scores at all trait optima values (Table S1). 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 parameters to vary across regimes.

Although results of both *C*-measures and OU model-fitting analyses can incorrectly suggest that divergent lineages are convergent, the reasons for this issue are different for the two methods; they are fundamentally different in how they test for convergence. Stayton (2015A, 2015B) highlighted that distance-based measures rely on a pattern-based definition of convergence that does not assume a specific mechanism is driving convergence (although see Mahler et al. 2017 for an opposing view), whereas OU model-fitting analyses assume that a specific mechanism, selective pressure (modeled as the α parameter), is driving convergence, and thus rely on a process-based definition of convergence. A further distinction between these types of convergence measures is that distance-based measures assess *morphological convergence of lineages* (i.e., the focus is on whether lineages are evolving toward each other), whereas OU model-fitting analyses test for *convergence on a morphology* (i.e., the focus is on whether lineages are evolving toward a trait optimum or adaptive

peak, not toward each other). This distinction between these two types of convergence measures is important: OU model-fitting analyses are not testing for similarities of lineages but rather similarities of lineages to a morphology; thus, they are less directly testing for convergence compared to distance-based measures, and instead are testing for occupation of similar adaptive zones, which may permit greater morphological variation than strict convergence towards a shared morphology.

Summary of issues with convergence measures. The most critical issue uncovered by our results is that *C*-measures often misidentify divergent lineages as being convergent (Fig. 2B, Fig. 3C, Table 1). However, all convergence measures show some bias, especially when examining morphological outliers, albeit for different underlying reasons for each method. *C*-measures, θ_{real}, and some OU model-fitting analyses (those in which parameters are not allowed to vary between regimes) all result in stronger measures of convergence when simulated convergent taxa evolve toward relatively farther trait optima. In contrast, the Wheatsheaf index shows weaker convergence in outliers, although the magnitude of this bias may be influenced by our simulation methods (Table 1). Further, model support for multiple-regime OU 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 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, examine the phylogenetic half-lives measured by OU models to evaluate model fit, and perform model-fitting analyses in conjunction with distance-based convergence measures.

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

Despite any shortcomings, *C*-measures have benefits over other convergence

628 measures, including the ability to distinguish between convergence and conservatism (Stayton 2015A). Thus, our objective is not to discourage the use of distance-based

630 metrics like *C*-measures but rather to identify issues and encourage the development of improved measures.

632 We help to address the *C*-measure issues by presenting novel distance-based convergence measures that are derived from the *C*-measures. The new measures are

634 calculated using the same equations as those for *C*1–*C*4 (except with a change to *C*4; see below and Supplemental Methods), but we limit the candidate D_{max} measurements

636 to distances between lineages at synchronous 'time slices' coinciding with internal phylogenetic nodes. We refer to the new measures as *Ct*-measures (or *Ct*1–*Ct*4) and

638 D_{max} as $D_{max.t}$ because time (*t*) is incorporated when measuring morphological distances between lineages, unlike the *C*-measures. *Ct*1 scores can be interpreted in the same

640 way as *C*1 scores were intended to be interpreted (Stayton 2015A): positive *Ct*1 scores represent a proportion of the maximum morphological distance between lineages that

642 has been covered by convergent evolution, with a result of one representing complete convergence. Like *C*-measures, statistical significance for *Ct*-measures is based on

644 comparison with expectations for evolution proceeding entirely on a BM model, with simulations used to generate the expectations.

646 By limiting the candidate $D_{max.t}$ measurements to time slices, the *Ct*-measures minimize the possibility of $D_{max.t}$ being erroneously inflated by divergent tips. This is

conceptually illustrated in Figures 4A and 4B, which are the same scenarios as in Figure 2A and 2B. 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 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 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 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 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} 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 *convevol R* package (Stayton 2018). This includes allowing *Ct*-measures to compare clades that contain multiple lineages, whereas the *C*-measures are limited to comparisons of 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) and 2) weighting of *Ct* scores and *p*-values based on the number of pairwise comparisons between focal clades (see Supplemental Methods). Further, *Ct*-measures can be measured using single traits (*C*-measures 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 (now
672 *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
674 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
676 *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
678 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
680 highlighted here.

We have also developed a new *R* function, *plot.C*, that produces a plot of the
682 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
684 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
686 Figure S6 for the 'twig' ecomorphotype lineages of anoles, although we separated
convergent and non-convergent pairwise comparisons for ease of interpretation. These
688 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
690 studies of convergence. The candidate $D_{\mathrm{max.t}}$ measurements at time slices are provided
as an output of the *convrat.t* function.

692 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

694 (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
696 drift/BM and selection/OU) exhibit *Ct*1 scores that are negative, correctly indicating divergence. Like *C*-measures, the *Ct*-measures do measure more convergence in
698 morphological outliers (Figs. 4A and 5A), but this pattern appears to be less pronounced than with *C*-measures (Figs. 2A and 3C). Although in some cases the *Ct*1 score is
700 greater than zero (indicating convergence; Fig. 5A), the *Ct*1 scores are not statistically significant when applied to divergence datasets (Table 1), which is in contrast to the
702 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
704 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
706 associated with *Ct*-measures when evaluating convergence in their samples.

Different origination ages of convergent clades might also inflate *Ct* scores in
708 morphological outliers, especially if the oldest lineage evolves rapidly into outlying morphospace and away from other putatively convergent lineages. This is illustrated in
710 Figure S4 and discussed in the Supplemental Results. To help address this issue, we added an optional feature to the *convrat.t* function that limits candidate $D_{\mathrm{max.t}}$
712 measurements to the time prior to the evolution of the focal lineages (e.g., prior to the evolution of the earliest glider clade). We recommend that researchers use this option
714 as a supplement to regular *Ct*-measures when their clades of interest have very different origination ages (see Supplemental Results).

716

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 they observed conflicting results. Statistically significant *C*-measure scores indicated strong convergence, but other analyses (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), 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 convergence. In some instances, the *Ct*1 scores are only slightly negative and have significant *p*-values (e.g., $Ct1 = -0.01$ and $p < 0.01$ for the comparison of scaly-tailed squirrels and flying 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 groups.

Huie et al. (2021) and Stayton (2015A; using data from Mahler et al. 2013) 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 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, 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 *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 ('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 significant, it was near zero (Table S2), 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 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 S6 and Table S2 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. S4). 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 *convrat.t* 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 evolutionary histories of ichthyosaurs and dolphins overlap from their most recent common ancestor (MRCA; early amniotes) to the ichthyosaur tips, providing considerable time (more than 200 million years) for candidate $D_{max.t}$ measurements. However, there are no candidate $D_{max.t}$ measurements

784 between the extinction of ichthyosaurs and extant dolphins, and if this is the period in which the ancestral dolphin lineages shows the greatest divergence from ichthyosaurs,
786 then $D_{max.t}$ will not capture the maximum divergence of the lineages. That is, much of the evolutionary history of dolphins (and placental mammals more broadly) would be
788 excluded by *Ct*-measures, likely resulting in smaller-than-expected *Ct* scores. (Note, however, that D_{tip} ignores time and would measure the morphological distance between
790 ichthyosaur and dolphin tips.) In these types of cases, alternative methods for measuring convergence (e.g., OU model-fitting analyses) may be more appropriate.

792 The restriction of candidate $D_{max.t}$ measurements to coincide with internal nodes exacerbates an issue inherent to many phylogenetic comparative methods: the reliance
794 on inferred ancestral states. $D_{max.t}$ is the critical value that enables the *Ct*-measures to diagnose convergence, and it is drawn entirely from ancestral state data, which are
796 estimated from tip values assuming a BM model of evolution. The consequence is that ancestral reconstructions are likely to reflect average morphologies of the sampled taxa,
798 decreasing the chance of measuring convergence via the *Ct*-measures because $D_{max.t}$ estimates may be artificially shorter than the 'real' $D_{max.t}$ values. For instance, despite
800 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% of $D_{max.t}$ has been closed by
802 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 measures to simulations in which we
804 used the generated ('true') ancestral traits. For 'glider' trait optima of 50, the *Ct*1 score is 0.67 when using estimated node reconstructions and it increases to 0.78 when using
806 the generated ancestral values (Table S4). This helps to confirm that the convergence

signal of *Ct*-measures might often be diluted due to the assumption of a BM model of evolution.

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 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 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 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 to improve ancestral reconstructions at internal nodes.

The number of phenotypic traits used to assess convergence is likely of increased 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 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 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, 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 S5, highlighting that increasing the number of BM-evolved traits (which are expected to be mostly non-convergent) in

simulations results in relatively greater increases of D_{tip} scores compared to $D_{max.t}$ scores. Therefore, an increased number of 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 recommend that researchers carefully choose traits (or landmarks if using geometric morphometrics) based on the specific hypothesis being tested, and analyze individual traits or subsets of traits whenever feasible to tease apart unique patterns among traits.

Many of the aforementioned factors that influence *Ct*-measures, especially the assumption of a BM mode of evolution in ancestral lineages, could contribute to the *Ct*-measures being conservative in their measures of convergence. This should be considered by researchers who use the *Ct*-measures.

As highlighted throughout this study, convergence measures can be biased based on the location of taxa in morphospace, with outliers tending to show greater convergence when using the *C*-measures, *Ct*-measures (although to a lesser degree than *C*-measures), θ, and some OU-model-fitting analyses (when model parameters are limited to being constant across regimes), and less convergence when using the Wheatsheaf index (Figs. 2A and 4A, Tables 1 and 2). We consider the greater observed convergence in morphological outliers via most methods to be an issue (Fig. 2A) because it is inconsistent with our working definition of convergence, which has the precision that allows for quantitative comparisons. However, under looser definitions of convergence this pattern could be interpreted as a reflection of the amount of evolutionary change of the convergent lineages. Outliers have undergone greater morphological change, evolving farther from the ancestral morphology, and thus their

tendency to appear ‘convergent’ could be an emergent property of the evolution of outlying morphologies (e.g., see Collar et al. [2014] for a discussion of ‘imperfect convergence’ in divergent, outlying lineages). This, however, is not what the measures of convergence have been defined to test, and we emphasize that researchers should ensure that their chosen convergence metrics and interpretations of results align with their *a priori* definition of convergence. For instance, researchers should think critically about whether they are trying to measure convergence toward a shared morphology or a shared trajectory toward a broad adaptive zone. In any case, researchers should expect to observe relatively stronger evidence of convergence in outliers when using most convergence measures.

Summary

The *C*-measures are a popular means of identifying and quantifying phenotypic convergence, in 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 (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 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 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, 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 convergence methods. Researchers should also recognize the benefits and drawbacks of their chosen methods when interpreting results, and they should put greater focus on methods that are most appropriate for the data being analyzed 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.

LITERATURE CITED

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

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

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. Evol. 29:283–312.

Arbour, V. M., and L. E. Zanno. 2020. Tail Weaponry in Ankylosaurs and Glyptodonts: An 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 convergent evolution. Methods Ecol. Evol. 5:685–693.

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

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

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

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.

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).

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

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

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.

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.

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.

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.

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

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

Ho, L. S. T., and Ané, C. 2014. Intrinsic inference difficulties for trait evolution with Ornstein-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 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 samples. Biometrika 76:297–307.

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

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.

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

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

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.

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.

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 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 convergence and conservatism shape ecomorphological evolution in ground-dwelling squirrels. Evolution 72:473–489.

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. 135:569–585.

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

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

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

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

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.

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

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). Zool. J. Linn. Soc. zlaa154.

Speed, M. P., and K. Arbuckle. 2017. Quantification provides a conceptual basis for convergent evolution. Biol. Rev. Camb. Philos. Soc. 92:815–829.

Stayton, C. T. 2008. Is convergence surprising? An examination of the frequency of convergence in simulated datasets. J. Theor. Biol. 252:1–14.

Stayton, C. T. 2015A. The definition, recognition, and interpretation of convergent evolution, and two new measures for quantifying and assessing the significance of convergence. Evolution 69:2140–2153.

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 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.

FIGURES

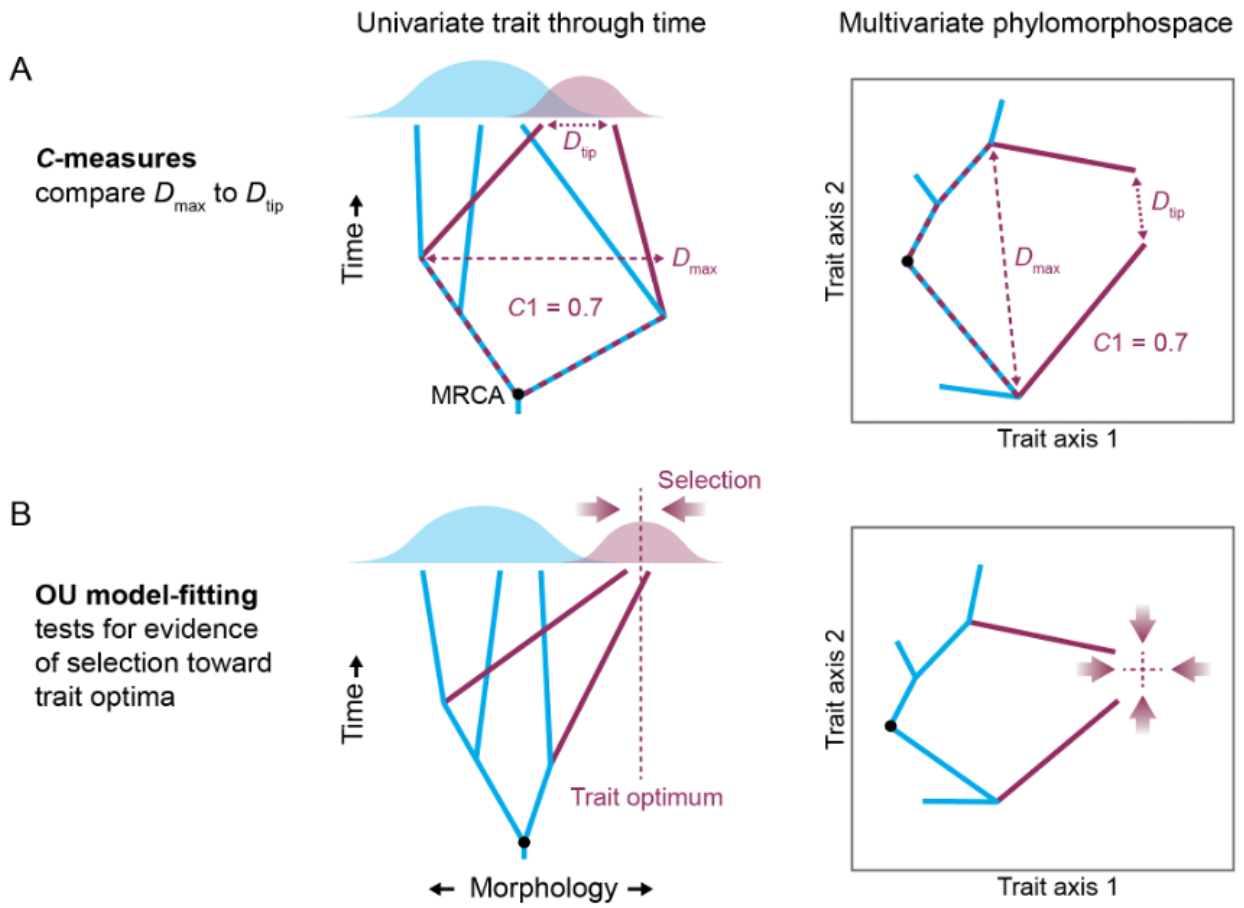
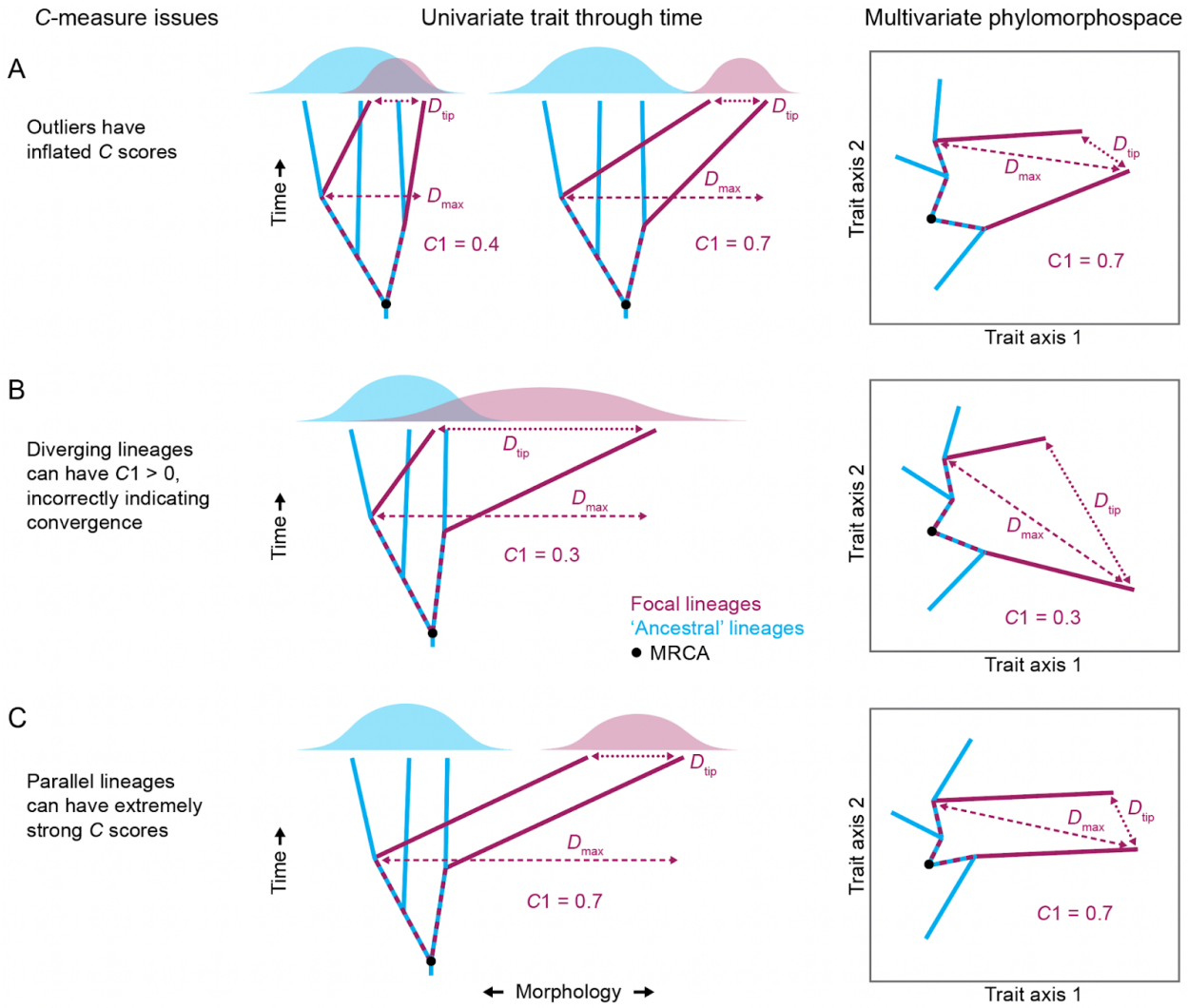


Figure 1. Conceptual illustrations of two methods for assessing phenotypic convergence of focal lineages (maroon): *A*, *C*1 of Stayton (2015A) and, *B*, multiple-regime Ornstein-Uhlenbeck (OU) model-fitting. The *C*1 score of 0.7 indicates that lineages have evolved toward each other to cover 70% of 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.



1072 **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
1074 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
1076 distribution curves in univariate illustrations could represent adaptive peaks, and OU model-fitting analyses are more likely to identify unique adaptive peaks when a peak is farther from the
1078 ancestral morphology (Table 1).

1080

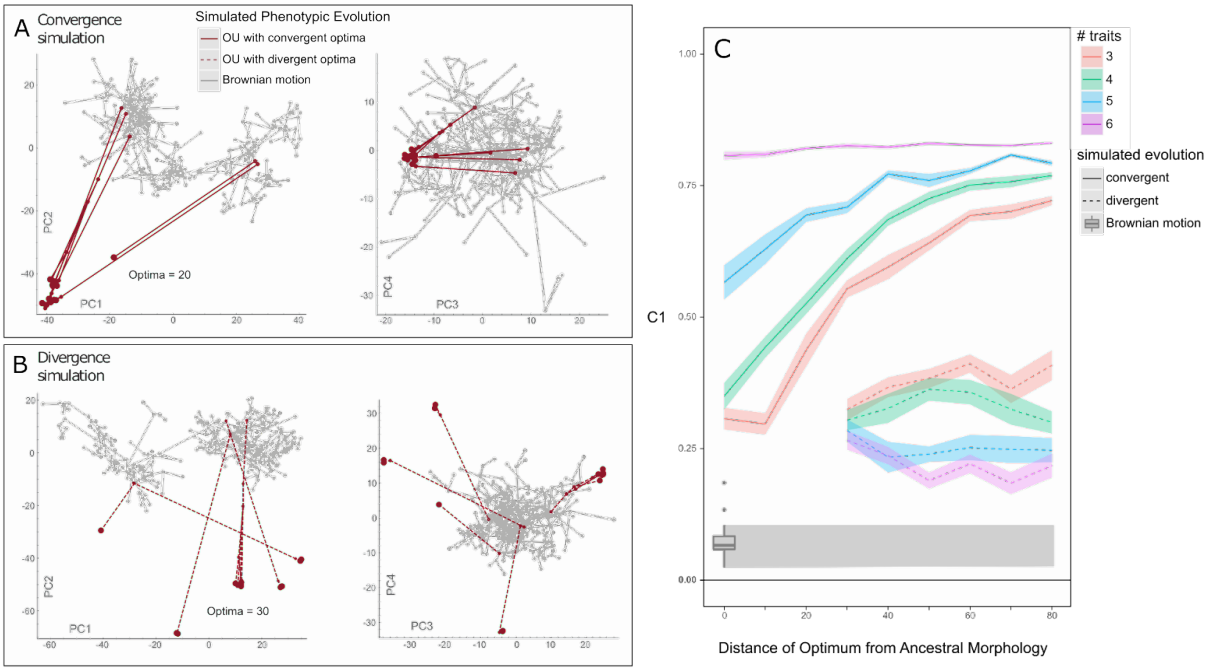


Figure 3. PCA phylomorphospaces for example datasets that simulate convergence (*A*) and divergence (*B*) of all six traits of five focal clades ('gliders'). Traits were selected toward optima via an OU process, with all traits of convergent taxa selected toward a value of 20 and traits of divergent 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.

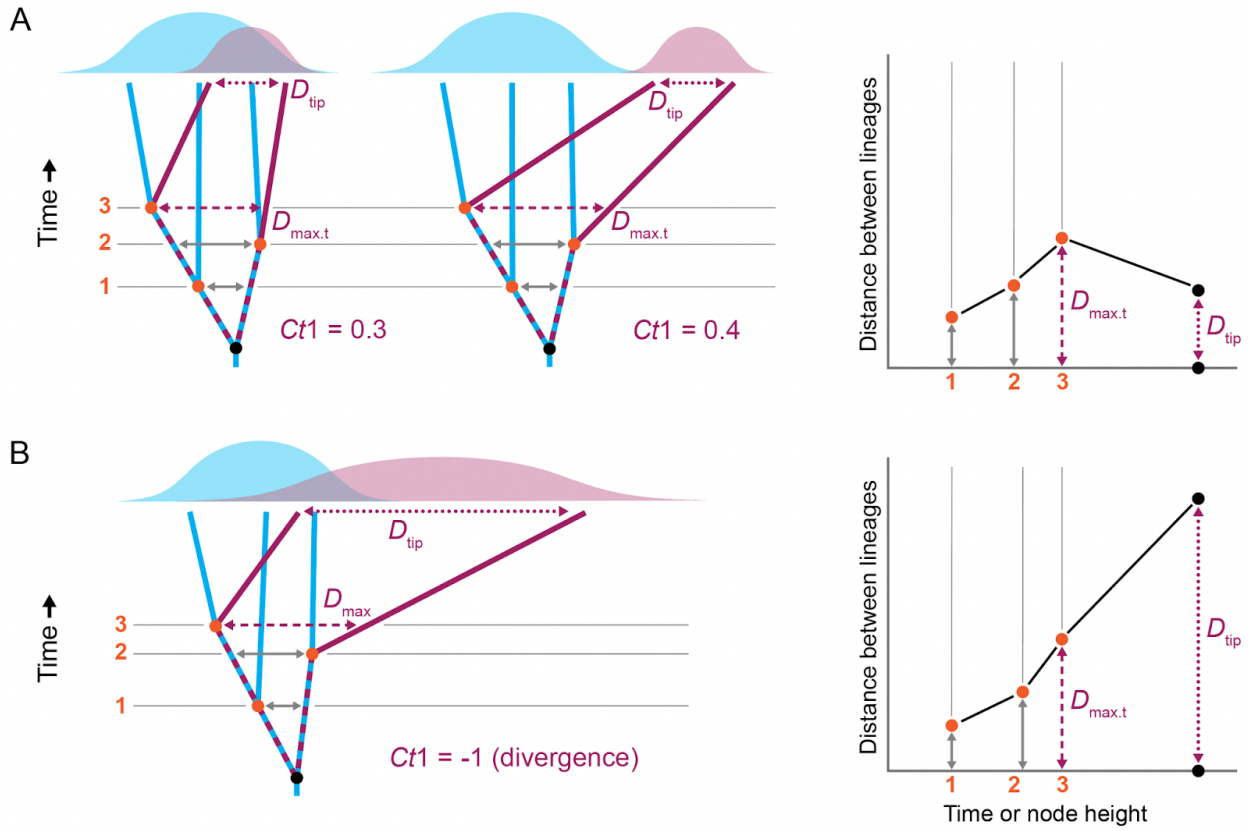


Figure 4. Conceptual illustration of our new *Ct*1 convergence measure, which is calculated like *C*1 of Stayton (2015A) but candidate $D_{max.t}$ measurements are limited to 'time slices' at internal phylogenetic nodes. The plots on the right show the three candidate $D_{max.t}$ measurements and the 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.

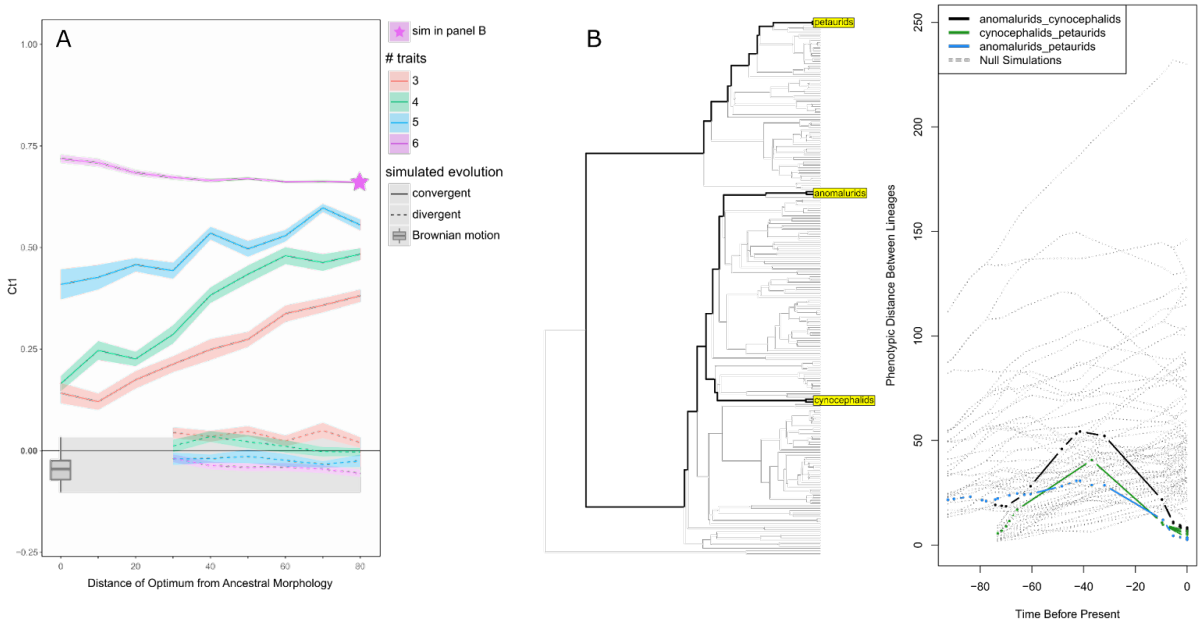


Figure 5. (*A*) *Ct*1 scores for simulated convergent lineages (top results in plot) and divergent lineages (bottom) under varying evolutionary scenarios. See the Methods and Figure 3 caption for more information. Although in some cases the divergence *Ct*1 results are greater than zero (indicating convergence), 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.

TABLES

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 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 vectors) suggest greater convergence, whereas relatively larger Wheatsheaf index, *C*1, and *Ct*1 values indicate greater convergence. Statistical significance (*, $p \leq 0.05$; **, $p \leq 0.01$; ***, $p \leq 0.001$) 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. 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). 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' 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

Table 2. Tests of convergence among lineages of the simulated datasets using evolutionary model-fitting analyses. Model-fitting results for each trait optimum are the mean AICcWs of 15 simulated datasets. Model support for the two-regime model (mvOU2) represents support for convergence because this model reflects evolution of focal lineages toward a shared adaptive peak. The 'constrained divergence' simulations limited the simulated traits of focal taxa ('gliders') to be positive values, thus constraining them to one region of morphospace (see Methods). See the Supplemental Results for 'unconstrained divergence' results. Abbreviations: AICcW, small-sample corrected Akaike weights; mvBM, multivariate Brownian motion model; mvOU, multivariate Ornstein-Uhlenbeck model.

Simulations	**Model**	**Trait optimum of focal taxa**				
		0	**20**	**40**	**60**	**80**
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

SUPPORTING INFORMATION

A cautionary note on using quantitative measures of phenotypic convergence

David M. Grossnickle, William H. Brightly, Lucas N. Weaver, Kathryn E. Stanchak, Rachel A. Roston, Spencer K. Pevsner, C. Tristan Stayton, P. David Polly, Chris J. Law

SUPPLEMENTAL METHODS

Simulations

We produced a graphical summary of our simulations to help clarify the various types and iterations of simulations used in this study (Fig. S1). The four types of simulations include 1) convergence of focal taxa via selection (OU-evolved traits), 2) divergence of focal taxa via drift (BM-evolved traits), 3) 'constrained' divergence of focal taxa 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). Each set of simulations used six traits per lineage, but for the convergence and 'constrained divergence' simulations the number of traits simulated to be convergent (or divergent) for focal taxa ('gliders') was systematically altered between three and six. Non-convergent traits were evolved via Brownian motion (BM). Varying the number of convergent traits helps to toggle the strength of convergence, with simulations of three convergent traits representing relatively weak convergence. Further, systematically altering the trait optimum values allows us to examine the performance of convergence measures when focal taxa are evolved to different distances from ancestral morphologies (i.e., varying positions in morphospace). See the Methods for more information on each of these types of simulations.

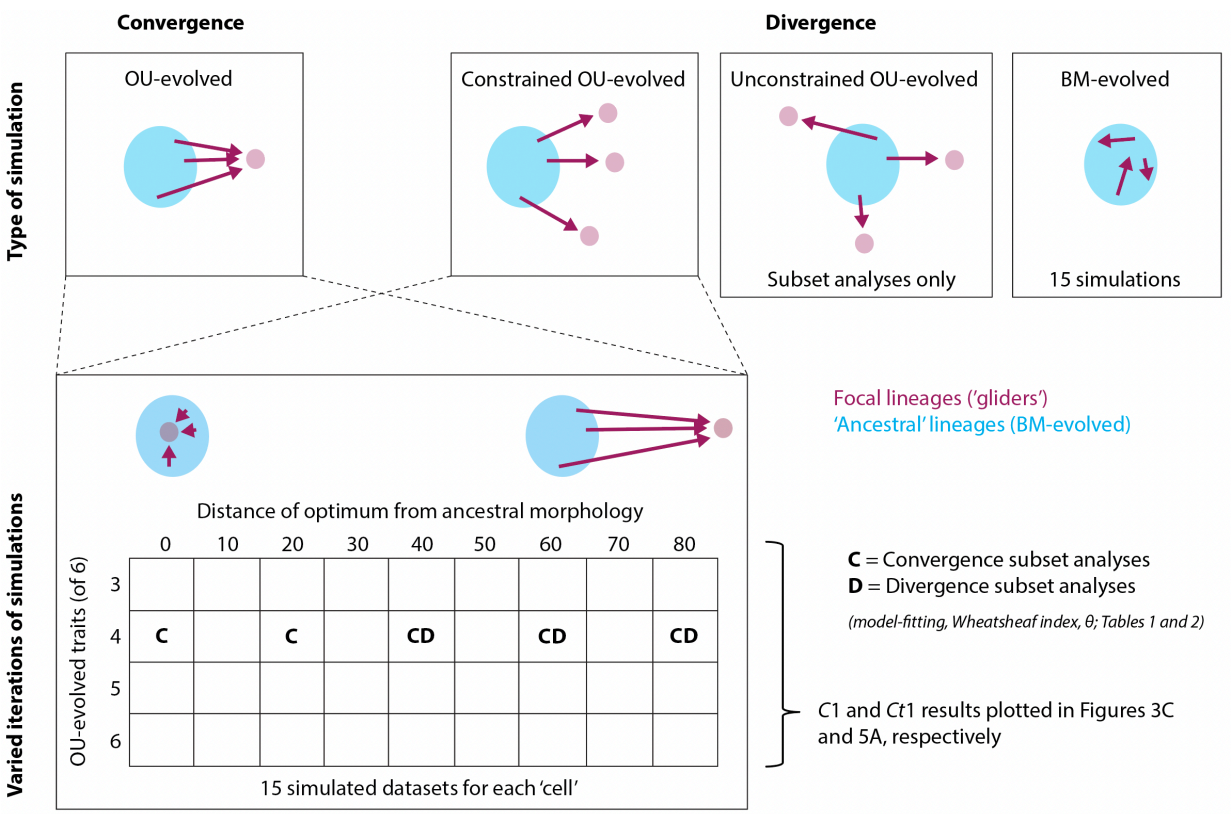


Figure S1. A graphical summary of the simulated datasets used in this study. See the Methods for full descriptions. Abbreviations: BM, Brownian motion; OU, Ornstein-Uhlenbeck.

Univariate model-fitting analyses

One limitation of the *mvMORPH* multivariate models, which are used for our primary model-fitting analyses, is that they do not permit the evolutionary rate (σ) or strength of attraction to optima (α) to vary between the two selective regimes ('gliders' and 'non-gliders'). This likely results in poor model performance because the datasets were simulated such that 'gliders' and 'non-gliders' should have different rates and attraction strengths. For example, the 'non-gliders' are evolved by BM, and thus they are not expected to exhibit attraction to a trait optimum, whereas the convergent 'glider' lineages are expected to exhibit strong attraction due to being simulated by an OU process. Further, the phylogenetic half-life (ln(2)/α) of the 'glider' regime

cannot be calculated independent of the 'non-glider' regime if the α parameter is uniform across both regimes, which is the case with the multivariate models.

Thus, we also fit seven univariate evolutionary models to the subset of simulated datasets, including several multiple-regime OU models that permit σ and α to vary between regimes. Using functions in the *OUwie R* package (Beaulieu et al. 2012), we fit these models to the first principal component (PC1) scores of a principal components analysis of the six simulated traits. The univariate models include uniform (or single-regime) BM and OU models, as well as a suite of multiple-regime OU models (i.e., 'OUM' models of Beaulieu et al. 2012). The OU2 model keeps α and σ constant for both regimes, the OU2A model allows α (but not σ) to vary between regimes, the OU2V model allows σ^2 (but not α) to vary between regimes, and the OU2VA model allows both σ and α to vary between regimes. As with the multivariate analyses, all models were fitted across 10 'simmaps' for each of the 30 datasets and relative support for models was measured using AICcW.

We recognize that fitting models to PC scores can lead to biased results (Uyeda et al. 2015, Adams and Collyer 2018), and thus our univariate results should be considered with caution. However, we feel that using PC1 scores here is justified for two reasons. First, the alternative option is to fit models to each of the six simulated traits individually, but four of the traits are evolved via a strong OU process and two traits are evolved via BM (in our subset of datasets used in model-fitting analyses; see Methods), and thus the model-fitting results are expected to vary considerably between those two types of traits. PC1 provides a single value for which results can be more easily interpreted compared to results for the six traits. Second, our conclusions concerning the use of model-fitting analyses for testing for convergence are based entirely on the multivariate model-fitting analyses (see Results & Discussion), and thus the results of the univariate model-fitting analyses (which are congruent with the multivariate results; Tables 2 and S1) do not influence the broad conclusions of this study. The univariate model-fitting analyses are simply a supplemental analysis that provide a fitted α value and phylogenetic half-life for the 'glider' regime.

***Ct*-measures**

1268 We used the *R* script from Zelditch et al. (2017) as a foundation for the updated functions for calculating *Ct*1–*Ct*4 and simulation-based *p*-values because they are computationally faster
1270 than the original *R* functions in the *convevol R* package (Stayton 2015, Stayton 2018). Note that the relevant *R* functions are titled *calcConv* (*C* calculations) and *convSig* (significance testing) in
1272 the *R* code of Zelditch et al. (2017), *convrat* and *convratsig* in the original *convevol R* package, and *convrat.t* and *convratsig.t* for our updated measures.

1274 *$D_{max.t}$ measurement*. The primary change made by the *Ct*-measures in comparison to Stayton's (2015) original *C*-measures is the way in which D_{max} is defined. *Ct*-measures were
1276 designed to ensure D_{max} (now referred to as $D_{max.t}$) was obtained from comparisons of synchronous time points along the evolutionary paths leading to the putatively convergent taxa
1278 of interest. In this way it prevents the inflation of $D_{max.t}$ that resulted from comparison of asynchronous nodes (e.g., tips and internal nodes) which often occurred when using the
1280 original metrics on lineages with outlying morphologies (Figs. 3C and 4). Several modifications to the source *R* code were made to facilitate this change. Candidate $D_{max.t}$ measurements for
1282 putatively convergent lineages are now measured at each internal node along the branch paths from the most recent common ancestor (MRCA) of the lineages (e.g., see Figures 4 and 5B). At
1284 each of these points we extracted the phenotypic distance between lineages as the euclidean distance between the ancestral reconstruction at the focal node and the coincident
1286 reconstruction along the branch path of the other lineage. Where this corresponds to a point along a branch (which is most cases) the ancestral state is estimated using formula [2] from
1288 Felsenstein (1985), which allows ancestral states to be interpolated at any point along a given branch from reconstructions at the branch's ancestral and descendant nodes. The code for this
1290 was largely repurposed from the *contMap* function of the *phytools R* package (Revell 2012). If no contemporaneous point exists on the opposite path for a given internal node (e.g., when
1292 comparing extinct and extant taxa), then a measurement is not taken at that node. All distances measured between paths are stored for each pair of user defined tips. $D_{max.t}$ is the maximum of
1294 these distance values, but it is restricted to predate either focal tip (i.e., $D_{max.t}$ cannot equal D_{tip}).

Restriction of $D_{max.t}$ to predate the focal tips means the minimum *Ct*1 value is no longer set to zero as in the original *C1*-measure. This allows for some degree of divergence to be captured (i.e., relatively more negative *Ct*1 values may represent greater divergence). However, users are cautioned from using this to test the magnitude of divergence between clades. This is because in divergent clades $D_{max.t}$ will almost always be the last time point before the oldest focal tip. The method will thus reflect only a small portion of the period when lineages were undergoing divergent evolution. Degree of divergence will then be a function of both phenotypic rates of evolution and of subtending branch length. The latter will in many practical situations be a function of sampling, with long subtending branches due to poor sampling likely to inflate divergence measures substantially since they will provide the best scenario for a large time difference between D_{max} and D_{tip} (and thus capture the greatest proportion of divergent evolution).

The changes to D_{max} were the most consequential of those made to modify the original *C*-measures. However, a number of other new options were also included. These are briefly described below. Full documentation of these options will be available as part of the next update to the *convevol R* package (Stayton, 2018).

User-defined groups. The first new option is for users to provide grouping assignments to the tips being tested, thus allowing comparisons of clades with multiple lineages, whereas the original *C*-measures are limited to comparisons of individual lineages. This option removes pairwise comparison between tips within the same group (e.g., two flying squirrels would not be compared if all flying squirrels are defined as one group) and returns results for each unique comparison between groups in addition to overall results. This option is useful if it is hypothesized that two (or more) clades converged, and relieves the user from needing to average tip values of a clade or manually define all of the desired comparisons. When using this option, the overall (for all pairwise comparisons) and comparison-specific *Ct* and *p* values are returned. Overall results are provided as both raw values (means of all pairwise comparisons, excluding within-group comparisons) and weighted values. The latter allows each inter-group comparison to impact the overall average equally, so that larger within group sample sizes don't skew overall results. For instance, if there are three putatively convergent groups (Group

A, Group B, and Group C), and Groups A and B both include a single lineage and Group C
1326 includes 10 lineages, then there would be 21 total pairwise comparisons among groups (one for A-B, 10 for A-C, and 10 for B-C). Although constituting one third of the unique inter-group
1328 comparisons, *Ct* measurements taken from comparison of Groups A and B constitute less than 5% of those used to compute overall (average) *Ct* values. Thus, Groups A and B have a relatively
1330 smaller impact than Group C on the overall *Ct* scores and *p*-values. The weighted output scales the *Ct* results (and associated *p*-values) so that each unique inter-group comparison contributes
1332 equally to the overall results, whereas the raw overall result simply reports the mean value for all 21 pairwise comparisons. Both weighted and unweighted values are reported in the default
1334 output printed by the updated *convratsig.t* function, but we recommend the weighted result be used by default when comparing groups. Nevertheless, the raw result may be preferable in
1336 cases in which researchers believe that the more heavily sampled group(s) should have a larger impact on overall results.

1338 Note that it is possible to define groups even when those consist of a single tip. While doing so will not change which pairwise comparisons the model considers, it will provide the
1340 user with unique *Ct* scores and *p*-values for each comparison. This can be especially useful when the degree of convergence varies across the lineages of interest (e.g., see the pairwise
1342 results for anole species in Figure S6 and Table S2).

Conservative $D_{max.t}$ option. When providing user-defined groups, a conservative $D_{\mathrm{max.t}}$
1344 option is available that limits candidate $D_{\mathrm{max.t}}$ measurements to a time point predating the origination of both focal groups (i.e., the nodes of the MRCAs of each group). This is to prevent
1346 $D_{\mathrm{max.t}}$ being skewed by an early transition of one lineage toward a shared adaptive optimum that is outlying in morphospace, which can result in inflated *Ct* scores, especially when the
1348 origins of the clades are very different in age. This issue is discussed in the Supplemental Results and illustrated in Figure S4. Note that this option is only meaningful when user defined
1350 groups are provided. When one of those groups consist of a single lineage the node immediately ancestral to the tip is used. Using this method, long branches can substantially
1352 alter inferred $D_{\mathrm{max.t}}$ values. We have provided the option to print relevant information about the restrictions put on $D_{\mathrm{max.t}}$ when using this method (by setting VERBOSE = TRUE in *convrat.t*).

1354 We strongly suggest that users investigate the impact of using the conservative $D_{max.t}$ option before committing to significance tests.

1356 *Updated Ct4 computation.* In addition to changes to $D_{max.t}$, we also altered the way in which the *C*4-measure is computed. The new version (called *Ct*4) redefines $L_{tot.clade}$, which is the

1358 value used to standardize the *C*2 value (D_{max} subtracted by D_{tip}) to obtain *C*4. $L_{tot.clade}$ is described by Stayton (2015) as reflecting the total amount of morphological evolution which

1360 occurs in the clade originating with the MRCA of two putatively convergent tips. In the original *C*-measures, $L_{tot.clade}$ values were obtained as a sum of the phenotypic distances from all

1362 pairwise comparisons between nodes in the clade, but this does not fully account for phylogenetic structure and is heavily influenced by sampling intensity. We have updated this to

1364 now be the sum of the phenotypic distances accumulated along each branch in the clade of interest. This change brings *C*4 closer to the original description of the metric.

1366 *Measuring convergence of single traits.* By default, the original *C*-measures do not support investigation of convergence in a single trait (although see Spear and Williams, 2020;

1368 Law, 2022). To circumvent this limitation we have added code to the *convrat.t* function which appends an invariant trait (with value zero) to datasets consisting of a single trait. This

1370 approach was taken due to ease of integration with existing code, and although crude will provide the same phenotypic distances as would be obtained from the single trait.

1372 *Model output.* Additional changes were made to increase the amount of information returned to the user and facilitate plotting of results. This includes the addition of the novel

1374 *plot.C* function, which is described in the 'Measuring convergence through time via Ct-measures' section of the main text (with example output in Figure 5B).

1376

SUPPLEMENTAL RESULTS

1378

Univariate model-fitting analyses

1380 For univariate models fit to PC1 scores the OU2VA model, which allows varying rates and attraction strengths between regimes, is the best fitting model at all trait optimum values for

1382 both convergence and divergences datasets (Table S1). However, for convergence datasets, the

null model (BM1) is the second best-fitting model when the trait optimum is zero and 20, and the total AICcW values for the OU2VA model increases with greater optima values. This pattern is consistent with the results of the multivariate evolutionary models (Table 2), suggesting a potential bias toward better fits of multiple-regime models when focal taxa are morphological outliers. This potential bias should be considered by researchers who use fits of multiple-regime OU models to test for convergence.

Table S1. Tests of convergence among lineages of the simulated datasets using evolutionary models fit to univariate data (PC1 scores) via functions in the *OUwie R* package. Model-fitting results for each trait optimum are the mean AICcWs of 15 simulated datasets (see Methods). Model support for the two-regime models (any variation of the OU2 model) could be interpreted as support for convergence because this model reflects evolution of the putatively convergent lineages toward a shared adaptive peak (but see the Results & Discussion). Abbreviations: AICcW, small-sample corrected Akaike weights; BM, Brownian motion; OU, Ornstein-Uhlenbeck.

Simulations	**Model**	**Trait optimum of focal taxa**				
		0	**20**	**40**	**60**	**80**
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

1406

_C_1–_C_4 and _Ct_1–_Ct_4 applied to simulated data

1408 In the main text we only present results for *C*1 (Fig. 3C, Table 1) and *Ct*1 (Fig. 5A, Table 1), which
were applied to both the simulated convergence datasets and the simulated divergence
1410 datasets. However, Stayton (2015) developed four distance-based convergence measures (*C*1–
*C*4) and one frequency-based measure (*C*5), with *C*1 being the primary measure, and we altered
1412 *C*1-*C*4 to produce the *Ct*1–*Ct*4 measures. Here, we provide full results for *C*1–*C*4 (Fig. S1) and
*Ct*1–*Ct*4 (Fig. S2), which are also applied to both the convergence and divergence datasets. See
1414 the Methods and Stayton (2015) for descriptions of the four convergence measures, and see
the Methods for information on the simulated datasets. Note that the *Ct*4 measure is
1416 calculated differently than the *C*4 measure (see Supplemental Methods). For *C*1–*C*4, all results
for divergence simulations are greater than zero (Fig. S2), incorrectly indicating convergence,
1418 whereas the *Ct*1–*Ct*4 scores for divergence datasets are generally at or below zero (Fig. S3).

1420

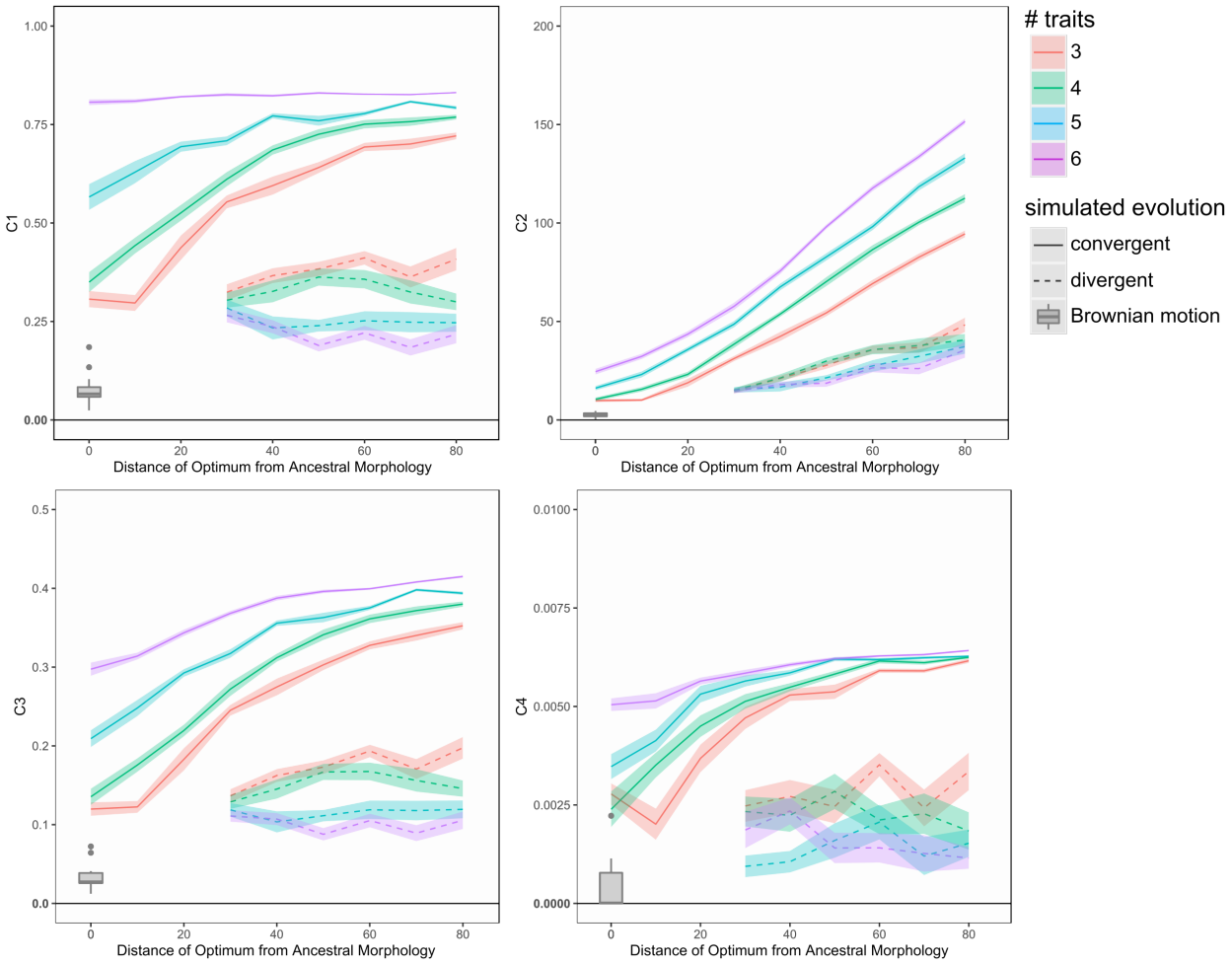


Figure S2. Plots of means and standard errors of *C*1–*C*4 scores for simulated convergent lineages (solid lines) and divergent lineages (dashed lines). Datasets varied in the number of convergent/divergent traits (represented by the different colored lines) and in the distance of trait optima from the ancestral morphology (approximated as the center of morphospace). Means and standard errors are computed from 15 simulated datasets. Greater *C*1–*C*4 values indicate greater convergence. We did not simulate divergence for trait optima of 0, 10, and 20 because at these optima our simulation methods may have inadvertently generated convergence patterns (see Methods and Figure 3). As a second means of simulating divergence, we allowed the lineages of interest ('gliders') to evolve via BM. These are provided as box-and-whisker plots, summarizing 15 simulated datasets of six traits (see Methods). Note that the divergence results are all greater than zero, incorrectly indicating convergence.

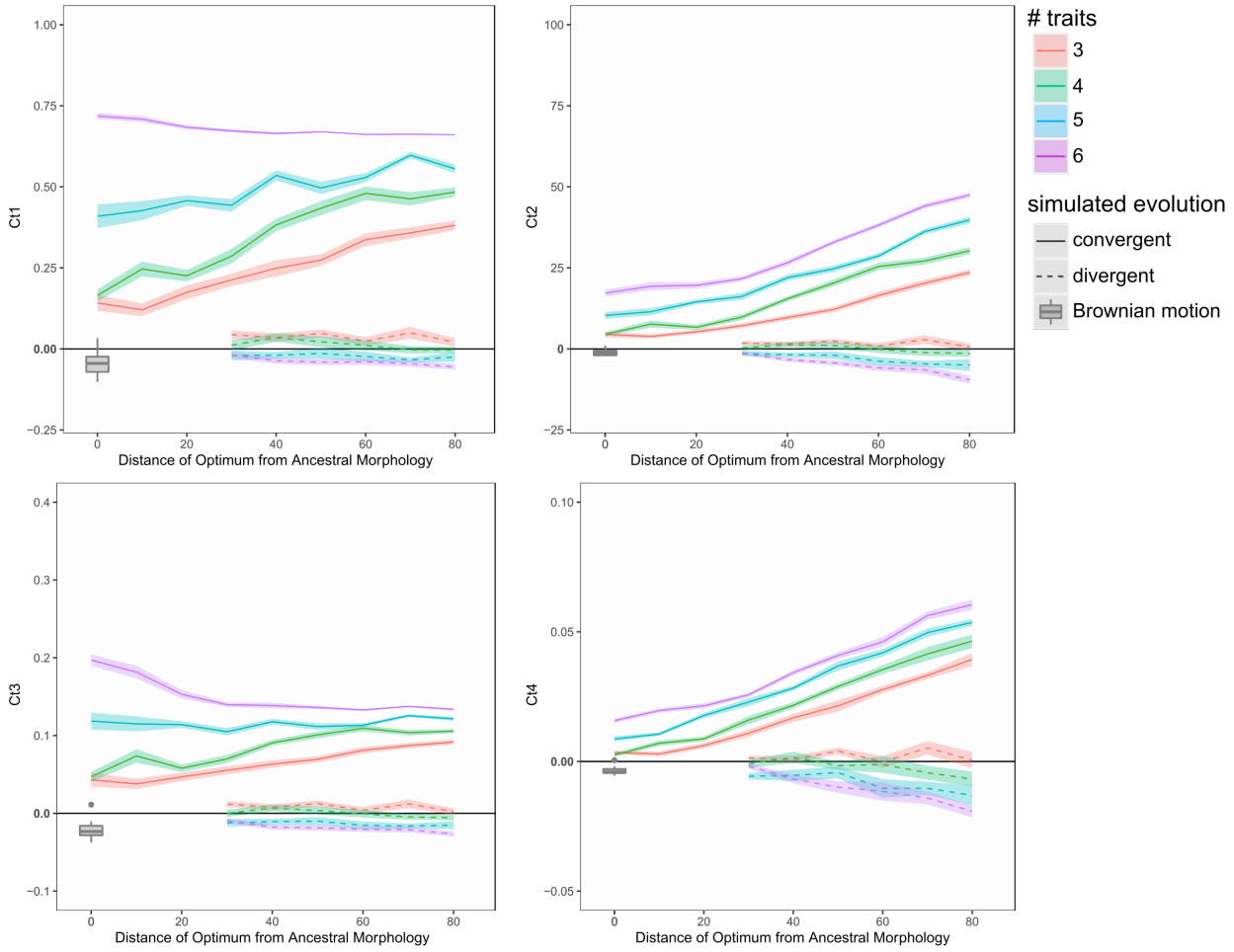


Figure S3. Plots of means and standard errors of *Ct*1–*Ct*4 scores for simulated convergent lineages (solid lines) and divergent lineages (dashed lines). Datasets varied in the number of convergent/divergent traits (represented by the different colored lines) and in the distance of trait optima from the ancestral morphology (approximated as the center of morphospace). Means and standard errors are each computed from 15 simulated datasets. Greater *Ct*1–*Ct*4 values indicate greater convergence. We did not simulate divergence for trait optima of 0, 10, and 20 because at these optima our simulation methods may have inadvertently generated convergence patterns (see Methods and Figure 3). As a second means of simulating divergence, we allowed the lineages of interest ('gliders') to evolve via BM. These are provided as box-and-whisker plots, summarizing 15 simulated datasets of six traits (see Methods). Note the differences in the scaling of the vertical axes of the *Ct*2 and *Ct*3 plots relative to the *C*2 and *C*3 plots (Fig. S1), respectively. (The scaling for *C*4 and *Ct*4 is different because these measures are calculated differently.) Also, note the different position of zero relative to results in the *Ct*1–*Ct*4 plots versus the position in *C*1–*C*4 plots (Fig. S1), as well as the overlap in the *Ct*1–*Ct*4 plots of divergence data simulated by both BM and OU processes.

1458

1460 ***Ct*-measures – the influence of origination times on results**

As discussed in the main text, the *Ct*-measures limit candidate $D_{max.t}$ measurements to specific
1462 time slices at internal nodes, and thus the timing of evolutionary change among putatively convergent lineages can influence the results of *Ct*-measures. For instance, if different lineages
1464 of interest evolve toward (or away from) a specific morphology at different points in time, then the $D_{max.t}$ measurement may not measure the morphologically farthest distances between the
1466 lineages. This issue may be magnified when convergence is expected to be linked to adaptive changes (e.g., adaptations for gliding behavior) that evolved at specific times. For instance, if
1468 colugos (i.e., Dermoptera or 'flying lemurs') evolved traits associated with gliding behavior approximately 60 Ma, and flying squirrels (Pteromyini) evolved traits associated with gliding
1470 approximately 25 Ma (e.g., Grossnickle et al. 2020), then most of the candidate $D_{max.t}$ measurements will be comparisons of dermopterans with gliding traits to stem flying squirrels
1472 without gliding traits (from 60 to 25 Ma). If the older lineage (colugos) has already undergone considerable evolutionary change by the time that the younger lineage (flying squirrels)
1474 originated, then much of the convergent evolutionary change of the older lineage is not captured by the morphological distances measured at 'time slices,' which are limited to the
1476 time period in which the lineages overlap. Ideally, most candidate $D_{max.t}$ measurements would be comparisons of non-gliding stem colugos and non-gliding stem flying squirrels that lack the
1478 adaptive traits associated with gliding. This issue might lead to candidate $D_{max.t}$ measurements being smaller than expected, or at least smaller than those calculated by measures that ignore
1480 time (e.g., *C*-measures).

Conversely, if the putatively convergent taxa evolve toward outlying regions of
1482 morphospace, then the asynchronous origins of the clades could inflate the *Ct*-measures. We illustrate this in Figure S4. Note that the orange arrows, which represent the origins of the focal
1484 lineages (and positions of candidate $D_{max.t}$ measurements), are the same morphological (x-axis) distance apart in each panel if time (the y-axis) is ignored. In the conceptual illustrations, the
1486 *Ct*1 score is consistently 0.3 when convergent lineages originate at the same time and/or when

lineages evolve toward the ancestral morphology. However, when lineages originate at different times and evolve toward an outlying region of morphospace, then the *Ct*1 score is 0.7 due to an inflated $D_{max.t}$ value (bottom right panel). Thus, researchers should be cautious when applying Ct measures to datasets with outlying taxa of various origination ages, and we offer some suggestions in the main text for mitigating this issue. It is also worth noting that this latter scenario assumes that the convergent lineages can reach adaptive zones; if the later-evolving convergent lineage is still evolving toward outlying morphospace (i.e., it has yet to reach an adaptive peak or zone) then the aforementioned issue may have less of an influence on *Ct* results.

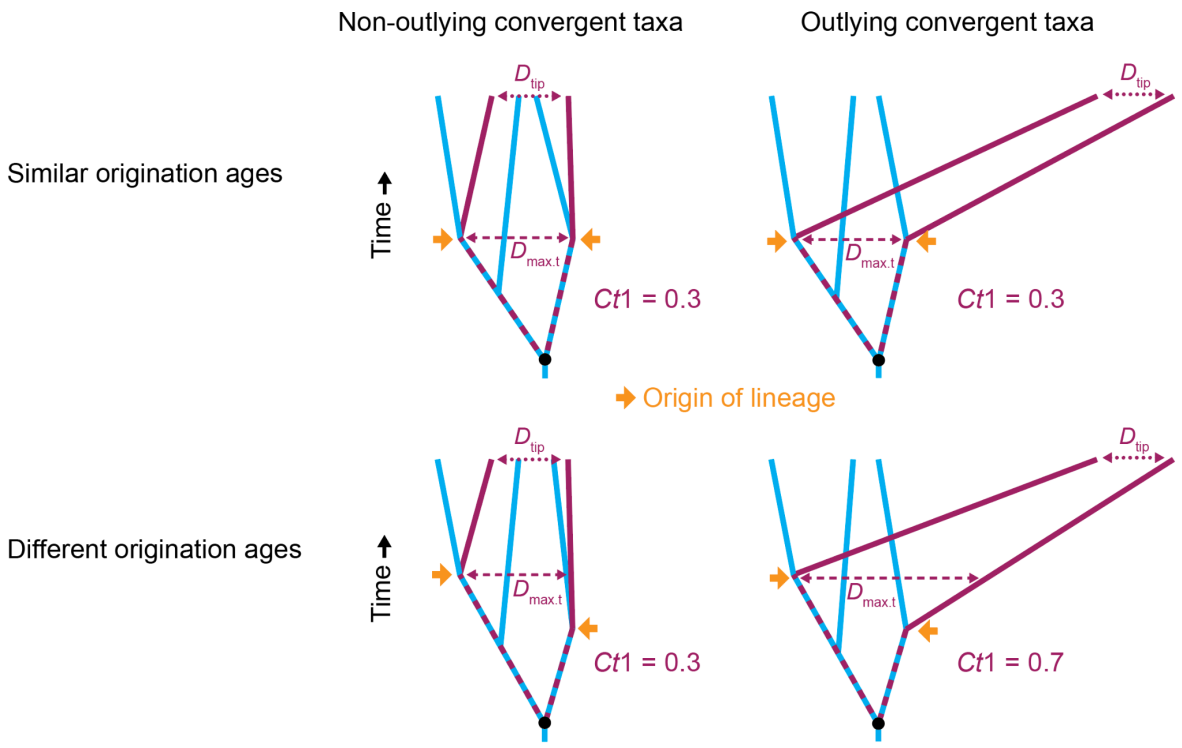


Figure S4. Conceptual illustrations demonstrating how *Ct*1 results can be influenced by a combination of outlying morphologies and varying origination times among convergent lineages. The *Ct*1 score is 0.3 in three of the scenarios but inflates to 0.7 when lineages both originate at different times and are outliers in morphospace (bottom right). To help mitigate this issue, we have included an option as part of the *convrat.t* function that allows users to

limit candidate $D_{max.t}$ measurement to the time period prior to the origination of the focal lineages (see Supplemental Methods). See the main text for descriptions of *Ct*1, $D_{max.t}$, and D_{tip}.

Influence of the number of traits on *Ct* results

As discussed in the main text (see Results & Discussion), the number of traits used in analyses (with all else equal) can bias the *Ct* scores. Inference of ancestral states via BM tends to average variation at internal nodes; thus, D_{tip} typically increases at a higher rate than $D_{max.t}$ for each non-convergent trait that is added to a dataset. (Here, we use "non-convergent traits" to refer to BM-evolved traits that are not selected to evolve toward a trait optimum via an OU process. These are often divergent, although it should be noted that BM-evolved traits could still be convergent by chance.) This is illustrated in Figure S5. The effect of this pattern is that an increased number of traits in analyses (with all else equal) could result in a relative decrease in *Ct* scores, unless those added traits are strongly convergent.

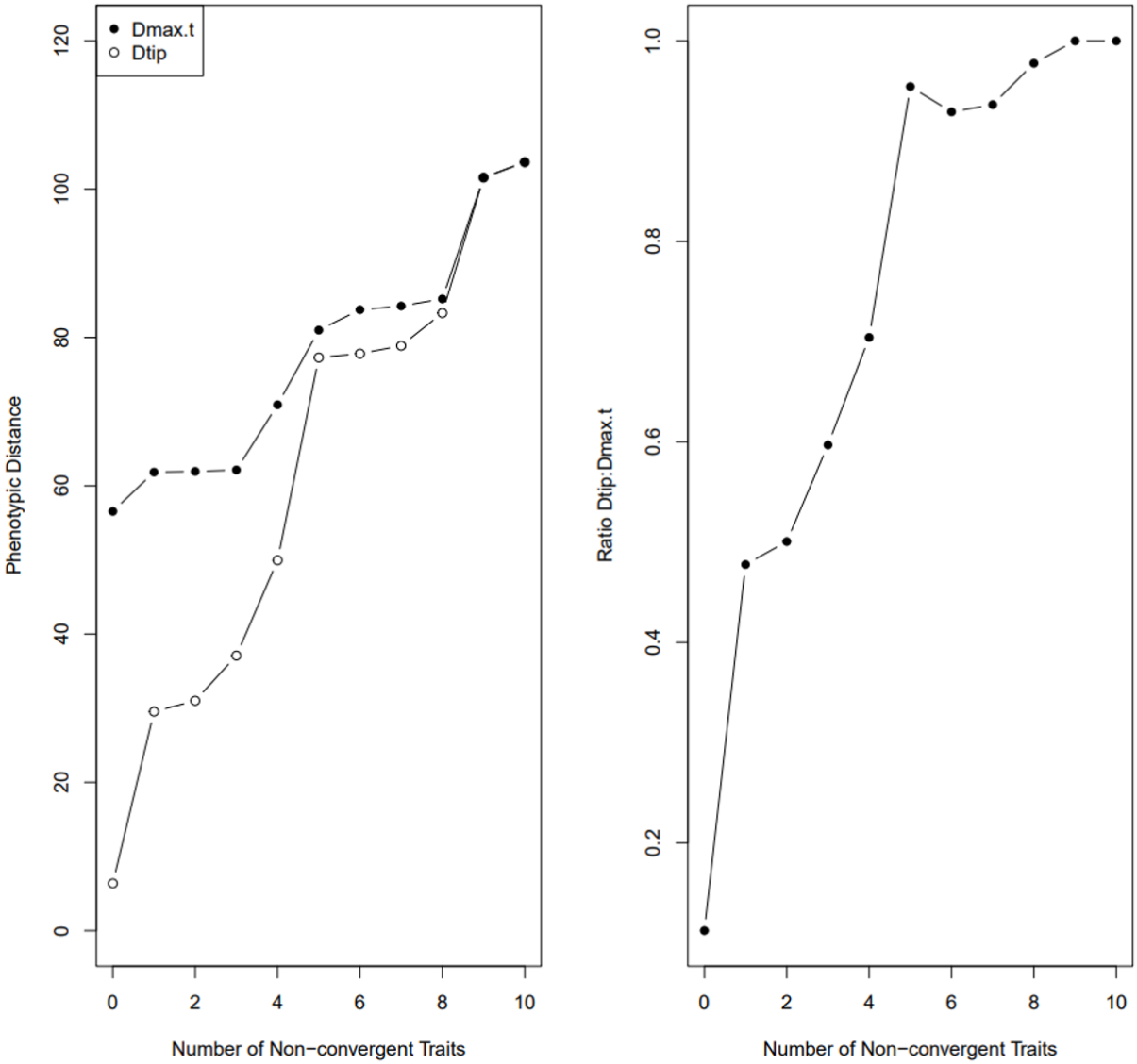


Figure S5. Illustration of how the number of traits used in analyses can influence *Ct*-measures, demonstrating the increased rate at which D_{tip} values increase relative to $D_{max.t}$ as additional non-convergent traits are included in analyses. (Here, 'non-convergent traits' refers to BM-evolved traits, which are expected to be divergent in most cases.) The left panel shows D_{tip} and $D_{max.t}$ measured between two 'glider' lineages with two simulated convergent traits (optimum = 100) and varying number of additional traits simulated via BM. The right panel shows the ratio between the D_{tip} and $D_{max.t}$ values.

Empirical example - *Anolis* 'twig' ecomorphotype

To test the novel *Ct*-measures and compare *Ct* results to those of *C*-measures (see the *Empirical examples* subsection of the Results & Discussion), we re-analyzed a classic example of convergence among *Anolis* lizards (Mahler et al. 2013), focusing specifically on five 'twig' ecomorphotype lineages. We chose this ecomorphotype because the taxa are morphological outliers that occupy a unique region of *Anolis* morphospace (Huie et al. 2021), and they have especially strong *C*-measure scores (Stayton 2015, Huie et al. 2021), although we believe that this is due in part to the lineages being morphological outliers (see Results & Discussion). Following the methods of Mahler et al. (2013), we size-corrected the traits via PGLS regression of each trait against the snout-to-vent length via PGLS. The *Ct*-measure results for this analysis are provided in Figure S6 and Table S2. Whereas the *C*1 score is 0.36 (Stayton 2015), we find the overall *Ct*1 score to be near zero for both the raw and weighted results (Table S2). This helps to highlight the inflated *C*-measure results due to the issues highlighted in the Results & Discussion. However, note that there is considerable diversity in the results among the ten pairwise comparisons; four are strongly statistically significant, whereas some (e.g., *Anolis occultus* and the *A. paternus* clade) show considerable divergence (*Ct*1 = -0.763; Table S2). To highlight the differences between convergent and non-convergent (or not significant convergence) pairwise comparisons, we separate those comparisons in Figure S6. Thus, we recommend that researchers examine and report results for pairwise comparisons whenever examining more than two putatively convergent lineages.

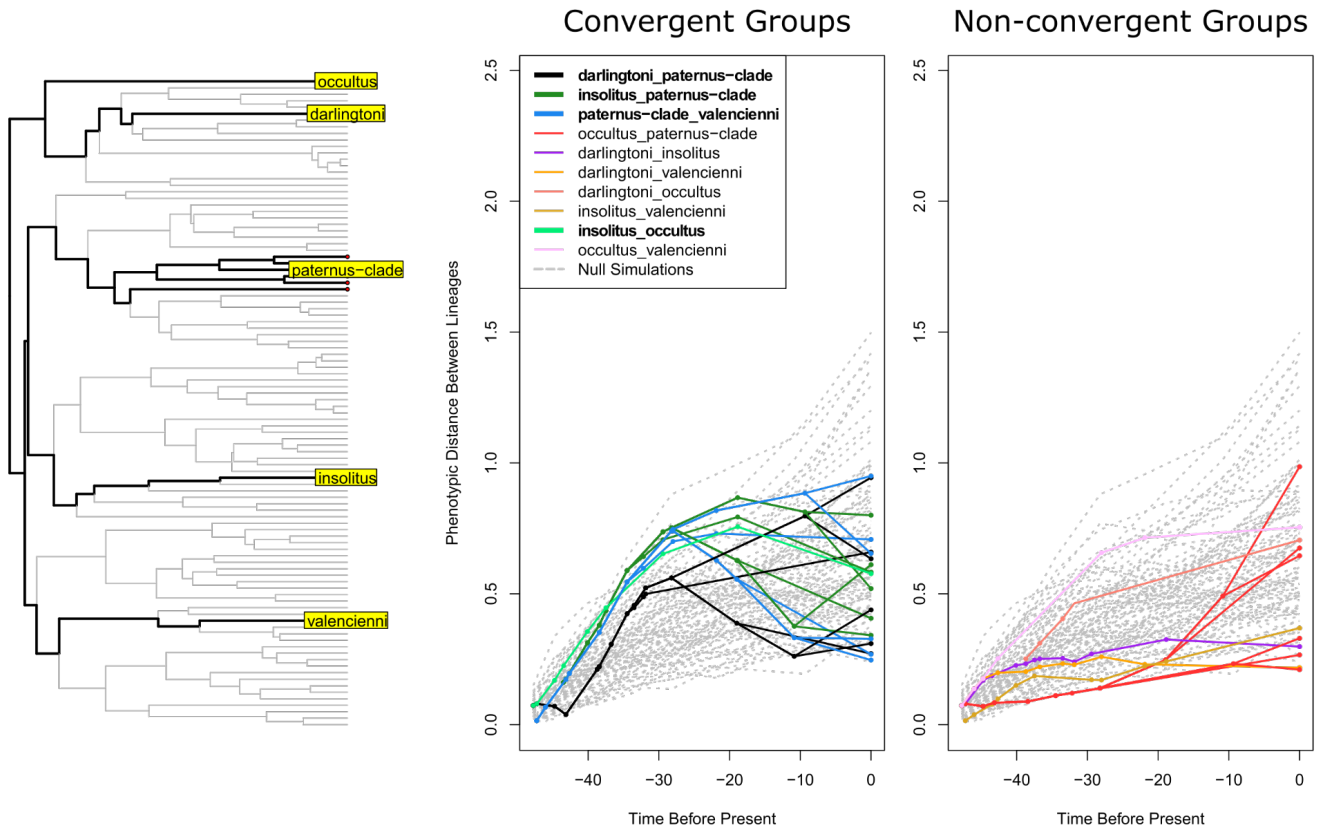


Figure S6. Summary of empirical tests of convergence in *Anolis* species belonging to the 'twig ecomorph' (Mahler et al. 2013). We size-corrected (via PGLS regression) and then analyzed the ten skeletal traits of the dataset of Mahler et al. (2013), with taxa assigned to groups based upon unique origins of the 'twig' ecomorphotype (see the *Ct-measures* section of the Supplemental Methods). The plots are the output of the *plot.C* function of the convevol R package, although the distance-through-time plot has been split to show statistically significant (left) and not significant (right) pairwise comparisons separately (see also Table S2). Significant pairwise comparisons are also indicated in bold in the key. Note that two of the 'non-convergent' comparisons in the right panel do have a positive *Ct*1 value, but they are statistically not significant (Table S2). There are 50 null simulations (light gray lines).

Table S2. *Ct*-measure values obtained for analyses run using the anole dataset of Mahler et al. (2013; ten standardized skeletal traits). Values are reported for overall comparison of ten 'twig ecomorph' species in five groups (corresponding to each independent origin of the ecomorph; Fig. S6). Pairwise comparisons of groups are also illustrated in (Fig. S6). See the Supplemental Methods for an explanation of the difference between 'overall raw' and 'overall weighted' results. Note that 'pat' refers to a five-species clade that includes *Anolis paternus* and four closely related species, whereas all other 'twig' taxa include a single lineage (Fig. S6); see the Methods for updates to the *convevol R* package that allow for comparisons among taxa with more than one lineage. Asterisks denote values returned as significantly different from null simulations (. - p < 0.1, * - p < 0.05, ** - p < 0.01). Abbreviations: *dar, Anolis darlingtoni*; *ins, Anolis insolitus*; *occ, Anolis occultus*; *pat, Anolis paternus*; *val*, *Anolis valencienni*.

	Overall		Pairwise comparisons									
	Raw	**Weighted**	***dar - pat***	***ins - pat***	***pat - val***	***occ - pat***	***dar - ins***	***dar - val***	***dar - occ***	***ins - val***	***ins - occ***	***occ - val***
***Ct* 1**	-0.01**	-0.057**	0.147**	0.323**	0.346**	-0.763	0.083 .	0.161 .	-0.521	-0.527	0.237**	-0.055
***Ct* 2**	0.072**	0.022**	0.086**	0.254**	0.261**	-0.216	0.027 .	0.042 .	-0.241	-0.127	0.179**	-0.039
***Ct* 3**	0.039**	0.012**	0.047**	0.111**	0.140**	-0.090	0.013 .	0.023 .	-0.117 .	-0.063	0.071*	-0.018
***Ct* 4**	0.002**	-0.006 .	0.003**	0.019**	0.011**	-0.007	0.001 .	0.001 .	-0.089	-0.005	0.006**	-0.001

Unconstrained divergence

For our primary selection-based divergence simulations, we 'constrained' the focal lineages ('gliders') to one region of morphospace by limiting the OU-evolved divergent trait optima to be positive values, whereas all BM-evolved traits (of both 'gliders' and 'non-gliders') could be positive or negative (see Methods and Figure S1). We believe that the 'constrained divergence' simulations help to mimic empirical datasets in which lineages exhibit some morphological similarities but are still geometrically divergent.

Nonetheless, we also simulated 'unconstrained divergence' for a smaller subset of evolutionary scenarios (see discussion on subset analyses in the Methods) in which the 'glider' trait optimum values were not limited to positive values. The results of convergence measures applied to these simulations are provided in Table S3. The results indicate that all convergence measures correctly identify the 'gliders' to be divergent. For instance, the two-regime OU model (mvOU2) is out-performed by the single-regime mvBM1 and mvOU1 models, indicating a

lack of evidence for 'glider' lineages converging on a shared adaptive peak. The mean *C*1 results are greater than zero, indicating a small amount of convergence. However, in contrast to *Ct*-measures, *C*-measures are not permitted to be less than zero (see Methods). Thus, the mean *C*1 results are biased toward being greater than zero because none of the 15 simulations can have values less than zero, and the small positive value is likely due to the occasional convergence of the 'glider' lineages by chance.

Table S3. Convergence measures applied to 'unconstrained divergence' simulations in which focal taxa ('gliders') were permitted to evolve in any direction (see Methods and Figure S1). Results are the means of 15 simulated datasets for each trait optimum value. The Akaike weights are small-sample corrected Akaike weights. None of the results for the distance-based convergence measures are statistically significant. For θ results, we report θ_{real} standardized to phylogenetic distance between clades. See Tables 1 and 2 for more information.

		Trait optimum of focal taxa		
		40	**60**	**80**
Model-ftting analsyes (Akaike weights)	mvBM1 model	1.000	1.000	1.000
	mvOU1 model	0.000	0.000	0.000
	mvOU2 model	0.000	0.000	0.000
Distance-based convergence measures	θ (*RRphylo*)	0.509	0.468	0.530
	Wheatsheaf index	0.840	0.640	0.481
	C 1	0.056	0.064	0.037
	Ct 1	-0.101	-0.081	-0.108

Influence of ancestral state reconstructions

As discussed in the main text, the assumption of a BM model of evolution for ancestral state reconstructions may often dilute *Ct* scores. We tested this assumption by comparing *Ct* scores obtained using our revised metric (with ancestral reconstructions via a Brownian motion model of evolution) to those obtained using a similar approach but employing the generated ('true') ancestral morphologies. The latter method used saved node values for each simulated

morphological trait. Comparison between these two approaches was conducted using 15 simulated datasets, with all six traits of 'gliders' converging to an optimum trait value of 50 (see Fig 5). Group-weighted *Ct* values showed consistently stronger convergence when true ancestral states were used in calculations (Table S4). However, differences were comparatively small (e.g., differences smaller than those between datasets containing five and six convergent traits; Fig 5). Nevertheless, these results indicate that the assumption that focal traits evolved via Brownian motion used to calculate *Ct* (and original *C*) scores can in some cases result in an underestimate of the degree to which lineages converge.

Table S4. We compared *Ct* values obtained using our revised metric (with ancestral reconstructions estimated via a BM model of evolution) to those obtained using the generated ('true') ancestral morphologies. Comparison between these two approaches was conducted using 15 simulated datasets, with all six 'glider' traits converging to an optimum trait value of 50. The standard errors (SEs) for the 15 datasets are also reported.

Ancestral traits	***Ct* 1**	***Ct* 2**	***Ct* 3**	***Ct* 4**
Estimated	0.67	32.63	0.14	0.04
SE	0.003	0.961	0.003	0.001
Generated	0.78	63.09	0.2	0.07
SE	0.002	1.536	0.003	0.002

LITERATURE CITED (in the Supporting Information)

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

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.

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

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

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.

Revell, L. J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. 3:217–223.

Stayton, C. T. 2015. The definition, recognition, and interpretation of convergent evolution, and two new measures for quantifying and assessing the significance of convergence. Evolution 69:2140–2153.

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.

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