Supplementary Materials and Methods

Collecting locality data

We list here additional detail on collecting times and localities. GPS coordinates are indicated in decimal degrees. We collected species in the U.S.A. in the spring and summer of 2017 and 2018 and in Mexico in June 2018.

In the U.S.A., we collected *Acris blanchardi* from April–June 2017 and May–June 2018 around Stillwater, Oklahoma, USA. Specific localities included Teal Ridge wetland (36.1005, -97.0804), Sanborn Lake (36.1561, -97.0781), and a residential pond (36.0279, -97.0483). We collected *Pseudacris fouquettei* and *P. crucifer* in March 2018 along the South Smokehouse Trail near Farmington, Arkansas (36.0411; -94.2211). We also collected *P. crucifer* on Kessler Mountain (36.0404, -94.2212) near Fayetteville, Arkansas, and *P. fouquettei* on Kalamazoo Road (35.3682, -93.6937) near Paris, Arkansas, in April 2017. We found *Hyla cinerea* and *Hyla avivoca* calling in the Duck Observation Pond (33.9533, -94.7028) in Little River National Wildlife Refuge near Idabel, Oklahoma, in May 2017 and April 2018. Finally, we collected *Hyla arenicolor* north of Fort Davis, Texas, in August 2017 and 2018 along Boy Scout Road (30.8134, -103.9281).

In Mexico, we found four species in La Sierra Juárez near La Esperanza, Municipality Santiago Comaltepec, Oaxaca. We collected *Charadrahyla nephila* on large branches overhanging three streams intersecting roads leading out of town (17.6528, -96.3882; 17.65082, -96.38905; 17.6233, -96.3658). Similarly, we collected *Exerodonta abdivita* on vegetation along a stream northwest of town (17.6495, -96.3858) and *Ptychohyla zophodes* on vegetation in an open pasture northeast of town (17.6305, -96.3653). We collected *Smilisca cyanosticta* in temporary ponds along the road near the Rio Bobo (17.65703; -96.39538). Finally, we collected two species in La Sierra Sur, Oaxaca. We found both *Smilisca baudinii* and *Tlalocohyla smithii* in a temporary breeding pond along El Zapote-Copalita highway between Pluma Hidalgo and Santa María Huatulco (15.8687, -96.3852).

Justification for rate of body-temperature change

Many methods exist for changing an amphibian's body temperature from ambient to test temperature. Such methods have included immersing animals in water and changing temperature as fast as 0.5ºC per minute (Gvoždík and Van Damme, 2006) or as slow as 4ºC per hour (Wilson, 2001). Others have placed a box with (room-temperature) frogs and some water into a thermal chamber at the test temperature for 1h prior to data collection (John-Alder et al., 1988). Still others have immediately immersed animals in water at the test temperature from 5–30 minutes (John-Alder et al., 1989; Whitehead et al., 1989). Thus, a specific method and rate of change have not been consistently used in previous studies of thermal performance curves in amphibians. On one hand, changing temperature at a slow rate (e.g. 4ºC per hour) allows an organism to gradually adjust to the test temperature, more likely avoiding thermal shock. On the other hand, changing body temperature at a fast rate (immediate immersion at the test temperature) may reduce stress in animals when tested at extreme temperatures (i.e. reducing the total exposure time to those temperatures).

In this study, we used immediate immersion of frogs in water at the test temperature, leaving them in the water until their bodies achieved that temperature, usually for 5–30 minutes. The timing depended on the size of the frog, as larger frogs took more time to reach the experimental temperature (verified with an infrared thermometer prior to data collection; see below). To achieve body temperatures at or below 20°C, we placed frogs in a circulating water bath in a makeshift thermal chamber consisting of a chest freezer controlled by an Inkbird itc-308 thermocouple. The thermocouple and immersion circulator ensured that the temperature of the water bath was within 0.5°C of the desired temperature (see below; Wilson and Franklin, 2000). To achieve temperatures above 20°C, frogs were placed in a water bath whose temperature was only controlled by an immersion circulator, which maintained water within 0.1°C of the intended temperature. We determined body temperature with an infrared thermometer, following previous work on small anurans (Navas et al., 2007). A key advantage of infrared thermometers is that they avoid body-temperature change by reducing handling time (Navas and Araujo, 2000). They also reduce handling stress relative to inserting cloacal thermometers, given that we took temperatures of individuals up to 15 times on days of experiments. Because the exact experimental temperature was difficult to achieve for each frog during every trial, we collected data when body temperatures were ± 1°C of the test temperature. We later used the exact measured temperature for statistical analyses.

Before fully implementing this procedure, we wanted to ensure our results were robust to the rate of temperature change. So we conducted a preliminary test to determine whether the rate of body temperature change (i.e. gradual or rapid) to a given temperature has an effect on an organism's performance at that temperature. We adjusted the body temperature of five *Acris blanchardi* to two extreme temperatures (8°C and 32°C) using two different methods. For gradual change, we changed water temperature 3°C per hour. For rapid change, we put frogs directly in water of the desired temperature for 5–10 minutes, as this was among the smallest of our species and individuals rapidly changed temperature. We used the same individuals with both methods, and for both experimental temperatures, then collected jumping data. We compared peak jumping velocity at a given temperature reached via the two methods of temperature change.

We found no significant difference between peak jumping velocities after gradually or rapidly changing body temperature, neither to 8°C (paired T-test, t_s = 0.46, df = 4, P = 0.476), nor to 32°C (paired T-test, t_s = –0.75, df = 4, P = 0.656). However, we did anecdotally observe increased stress at the slow ramping of high temperature, as two individuals died shortly after the high temperature treatment. Thus, we used the rapid procedure to facilitate data collection, reduce mortality, and increase quality of our resulting data. Note that we did not have additional high-temperature associated mortality after implementing the rapid procedure for data collection across all species. Moreover, our adopted procedure did not result in reduced performance over time (next section), further suggesting that our experimental animals did not experience thermal shock associated with drastic changes in temperature.

Accounting for potentially reduced performance over time

Performance may gradually decline when individuals are measured over the course of a week (Zug, 1985). Thus, we tested each frog at 20°C at the beginning of trials and then again at 20°C after all other experimental temperatures (Wilson, 2001). Some individuals performed more than 10% better at the end of trials, suggesting that they could also randomly perform more than 10% worse at the end. To account for this possibility, for each individual we subtracted its peak performance at 20°C at the end of trials from its peak performance at 20°C at the beginning of trials. We then calculated a 95% confidence interval around the mean of this metric across individuals of each species. Individuals that fell below the lower limit of the interval (i.e. significantly lower performance at the end than the beginning) were excluded from further analysis. Those that fell above the 95% CI were not excluded because they showed substantially higher performance at the end of trials, meaning their performance did not decline over time. Under this approach, we retained data from all but one individual of *Smilisca cyanosticta*.

Potential effects of body size on jumping performance

Our taxa varied in body size, which ranged across species from *Tlalocohyla smithii* (0.84g; *n* = 1) to *Smilisca baudinii* (mean ± sd = 18.12 ± 2.00 g). Size also varied within species (e.g. 3.96–12.90 g in *Hyla cinerea*). Previous comparative studies of anuran jumping have shown that absolute jumping performance (e.g. velocity in meters per second; distance in meters) can increase with body size (Emerson, 1978; Zug, 1978; Gomes et al., 2009). This pattern is stronger across than within species (James et al., 2007). Thus, size variation in our dataset may have affected differences among species and individuals in absolute jumping velocity. However, our goal in this paper was to examine species differences in the temperature at which they performed best, and not the actual value of jumping velocity attained at those temperatures. Thus, we followed previous studies and standardized jumping velocity data to the peak absolute velocity for each individual (John-Alder et al., 1988; Navas et al., 2008; Herrel and Bonneaud, 2012). This ensured that we could calculate thermal performance curves across individuals within species. It also allowed us to compare temperatures of peak performance across species.

Considering alternative methods for characterizing thermal performance curves

Many methods have been proposed for estimating species' thermal performance curves (TPCs). Some methods have been criticized as biologically unrealistic (e.g. Bulté and Blouin-Demers, 2006), and the exact optimal parameterization is still contested (e.g. Woods et al., 2018; Rezende and Bozinovic, 2019). For our approach, we followed Angilletta (2006), who suggested comparing phenomenological regression models using AICc. While more mechanistic models (e.g. Adams et al., 2017; Rezende and Bozinovic, 2019) have some desirable properties, the overall functional form of such curves is similar to those we used (i.e. a single-peaked curve crossing 0 at low and high temperatures). Moreover, a key analytical advantage of models we did not consider is that they allow for curve asymmetry with relatively few parameters. Given that we sampled the cold side of our species' TPCs, including asymmetry above peak performance temperature is unlikely to be strongly supported by our data. Thus, we do not expect our results would greatly differ with this approach or others that are similar.

Additional performance thresholds

In this paper we determined thermal performance curves in jumping, then used the lower temperature at which species drop to 80% of peak performance (L80) to quantify the lowest temperature at which they can still perform well. We used this threshold due to precedent in previous studies (John-Alder et al., 1988; Wilson, 2001). However, we also tested different performance thresholds to determine the sensitivity of our results to arbitrarily choosing 80% as "high performance." We considered thresholds at 70 and 90% of peak performance, as well as the temperature at which performance peaked. In most cases, the 70 and 80% thresholds occurred within our data (Figs. 3, 4). In *Smilisca cyanosticta*, however, all jumping data were above 80% of peak performance. Thus, for this species we extended the regression line beyond our data to reach an estimated L80. For five additional species, we extended regression lines to reach the estimated L70. We found that all thresholds generated quantitatively similar results in our evolutionary analyses (Table S2). Thus, we only present results based on L80 in the main manuscript.

We also note that in our evolutionary analyses, we only used the lower bounds of peak performance instead of the full thermal performance breadth, whereas the latter is often used in studies of thermal performance curves. We concentrated on the lower bounds for two reasons. First, the lower end of the thermal performance curve was most relevant to colonization of the temperate zone. Comparative studies of terrestrial ectotherms show that the upper (i.e. hot) bounds of both tolerances and thermal performance curves vary little with latitude (Snyder and

Weathers, 1975; Sunday et al., 2011; Sunday et al., 2019) or when comparing temperate and tropical species (John-Alder et al., 1988; van Berkum, 1988). Second, the highest temperature at which we tested frogs was 35°C, which we chose based on previous studies in anurans (John-Alder et al., 1988; Whitehead et al., 1989; Wilson, 2001; Herrel and Bonneaud, 2012). While this temperature was higher than the peak performance temperature in most species (Figs. 3, 4), it only marked a decline to 80% of peak performance in one of 12 species (Figs. 3, 4). This limited our analysis of the full breadth of performance (Huey and Stevenson, 1979), given that extrapolating curves beyond observed data can lead to large prediction errors (Sokal and Rohlf, 1995). Nonetheless, we do not expect our focus on L80 (rather than full breadth) to impact our results, as we found that the cold portion of performance breadth (i.e. the difference between the temperature of peak performance and L80) also gave nearly identical results as L80 alone in our evolutionary analyses (Table S2).

Internal node states for OU model-fitting

The multiple-optima OU models (OU2 and OU3) required specifying states of internal nodes. However, here we studied only 12 of 197 species of the Middle American clade (Faivovich et al., 2018; AmphibiaWeb, 2021), and ancestral-state estimates can be highly inaccurate with such sparse taxon sampling (Salisbury and Kim, 2001). Thus, for OU2 we used previously estimated ancestral areas for the entire Middle American clade of hylid frogs (Moen et al., 2009), which are consistent with biogeographic analyses on all anurans (Pyron, 2014). For OU3, we overlaid the ancestral-area estimates of Moen et al. (2009) with the ancestral elevation estimates of Smith et al. (2007), both estimated using the same phylogeny. Smith et al. (2007) found that most of the internal nodes in our tree were above 1000m, which we considered high elevation. The key exceptions were the nodes of three clades, all recovered as <1000m: the clade including *Hyla* (temperate), the clade including *Smilisca* and *Tlalocohyla* (tropical), and the common ancestor of these two clades (tropical). Moreover, both colonizations of temperate North America by hylids are inferred to have happened at low elevations along the coasts (Smith et al., 2007; Moen et al., 2009), and the sister group to the Middle American clade is Lophiohylini, a lowland tropical group (Wiens et al., 2010; Wiens et al., 2011; Pyron, 2014). Thus, we designated those nodes as lowland tropical (Fig. 2).

Phylogenetic comparative methods with and without standard errors

Recent studies have highlighted the potential importance of explicitly accounting for intraspecific variation in species means when doing phylogenetic comparative analyses (Ives et al., 2007; Revell and Reynolds, 2012; Silvestro et al., 2015). However, because we derived our L80 values from curves, we were unable to calculate typical standard errors of the L80 species' values. Thus, we tested models of CT$_{min}$ evolution both with and without standard errors to roughly examine the potential effect of excluding intraspecific variation on our results for L80. We found nearly identical results when including and excluding standard errors (Table S4). Thus, we expect that our L80 results were largely robust to our exclusion of measures of intraspecific variation.

Testing power and Type-I error with parametric bootstrapping

We tested the power of our data to distinguish the models we compared in this paper. Moreover, we examined the potential for Type-I error rates (i.e. falsely rejecting a simpler model when it is true). In testing OU models, power and error rates are influenced by both species number (sample size) and empirical parameter estimates (effect size; Beaulieu et al., 2012; Boettiger et al., 2012; Ho and Ané, 2013, 2014; Cressler et al., 2015). To estimate both error and power, we used parametric bootstrapping to simulate data and compare pairs of models (Boettiger et al., 2012).

Parametric bootstrapping for estimating 95% confidence intervals of parameters is well developed in the R packages *ouch* (version 2.14-1; Butler and King, 2004, King and Butler, 2009) and *OUwie* (version 2.6; Beaulieu et al., 2012). For model comparison, Boettiger et al. (2012) developed the package *pmc* to directly compare models with parametric bootstrapping. *pmc* (version 1.0.4) allows users to implement just a single function to test models with *ouch*. However, no analogous package or function yet exists for *OUwie*. To implement a similar analysis, users must simulate trait evolution in *OUwie*, then write their own code to process the simulations, fit models, and calculate statistical properties. The advantage of *OUwie* over *ouch* is that the former gives users more flexibility in model-fitting and assumptions are more explicit (see discussion in *OUwie*'s vignette, "New additions as of OUwie 2.1", found on the package's CRAN website at https://CRAN.R-project.org/package=OUwie). However, simulation in *OUwie* is simpler for some models (complex OU models) than others (Brownian motion, single-optimum OU). Therefore, we wrote an original function in R to mimic that of *pmc.* With this tool we hope to encourage more users to implement these simulations into their analyses. We use the function in our parametric bootstrapping analyses and more thoroughly detail its use in our second R tutorial (Appendix S14). We provide the function in Appendix S5. Both files are available on the Dryad Digital Repository (Moen et al., 2021).

As described by Boettiger et al. (2012), testing comparative models with parametric bootstrapping starts with simulating many replicates of trait evolution along a phylogeny. These simulations use empirically estimated parameter values for both a focal and an alternative model. One then fits both models to both simulated datasets, producing four total model fits for each simulation replicate (i.e. focal and alternative models fit to data simulated under the focal model, and both fits to data simulated under the alternative). By calculating a test statistic comparing the two models (e.g. a likelihood ratio test statistic) for each pair of model fits, two distributions of expected test statistics are generated. One reflects the expected test statistics if the data were truly generated by the focal model, and the other reflects the expected distribution of test statistics if the alternative model generated the data.

One use of these bootstrapped distributions is to probe the reliability of results by determining which scenario is more consistent with the observed test statistic from the data. The distributions can also be used to calculate Type-I error rates and statistical power of an empirical model comparison. In terms of Type-I error rates, "error" means falsely rejecting the simpler model (Sokal and Rohlf, 1995). For example, imagine comparing Brownian motion (BM) with a single-optimum OU model (OU1). With parametric bootstrapping, the data are simulated under the simpler model (BM), then those data are fit to both the simulating model (BM) and the alternative model (OU1). For each simulated dataset, one can statistically compare the models with a likelihood ratio test, a common method for comparing the relative support of nested models (Edwards, 1972; Posada and Buckley, 2004). The proportion of tests (across simulation replicates) that reject the simpler model (BM) is the Type-I error rate, since the data were simulated under the simpler model that should not be rejected. Next, to estimate the power of the data to reject the simpler model, the data are simulated under the more complex model (OU1) and then one similarly compares the support for the two models. The proportion of those model comparisons that (correctly) reject the simpler model is the statistical power of the dataset. When comparing models of phenotypic evolution, this power reflects both number of species (i.e. sample size) and signal in the data for different models (i.e. effect size; Boettiger et al., 2012; Ho and Ané, 2013, 2014).

In this paper, we focus on AICc-based model comparison and parameter estimation, rather than hypothesis testing to reject models (Butler and King, 2004; Beaulieu et al., 2012). However, here we use parametric bootstrapping to demonstrate how the method can be used to estimate the statistical properties of empirical datasets. For CT_{min}, we compared the temperate-tropical OU model (OU2) to the next most supported simpler (Brownian motion) and more complex (temperate vs. tropical lowland vs. tropical highland; OU3) models (Table 2). For L80, we compared the most supported model (Brownian motion), which was the simplest possible model, to the next two more complex models (OU1 and OU2; Table 2). For all simulations, we used maximum-likelihood parameter estimates from *OUwie* to simulate 1000 replicate datasets. For each model comparison, we then conducted likelihood ratio tests to calculate Type-I error (when data were simulated under the simpler model) and power (when data were simulated under the more complex model), as described above. We conducted all simulations and subsequent model fits with options for *OUwie* as described in the main text. The only exception was that we fit models without standard errors for species means of CT_{min}, since *OUwie* does not currently allow input of those errors when simulating trait evolution.

Parametric bootstrapping revealed strong support for CT_{min} evolving under the temperate-tropical (OU2) model as compared to Brownian motion, which had the second-highest AICc support in our main analyses (Table 2). Simulations showed that this comparison has high power (0.992) and a low Type-I error rate (0.065). Moreover, our empirical likelihood-ratio test statistic of 12.0 was higher than those from nearly all Brownian motion simulations, and it fit well within the distribution resulting from simulating OU2 evolution (Fig. S2A). In contrast, simulations showed that our data contain little information to clearly distinguish OU2 from the OU3 model (Fig. S2B), with moderate power (0.598) but high Type-I error rate (0.268). This result is somewhat surprising given the low AICc support for OU3 (Table 2). Yet its likelihood suggests that OU3 provides a reasonable improvement in model fit compared to OU2 (Table 2). Together, these results indicate that the low AICc weight for OU3 largely stems from the high AICc penalty for the additional optimum in OU3 relative to its improvement in model fit. Nevertheless, in both OU2 and OU3, lineages that colonized the temperate zone show a lower optimum for CT_{min}. What remains unclear is whether tropical highland and lowland lineages are also different.

Our results for L80 showed a low ability to distinguish models in general (Fig. S2C,D). Type-I error rates were low when comparing Brownian motion to both OU1 and OU2 (0.030 and 0.079, respectively), but power to reject the simpler model when false was poor (OU1 = 0.129; OU2 = 0.250). Moreover, the empirical likelihood-ratio test statistic fell squarely within both simulated distributions (Fig. S2C,D). These results are unsurprising, as a good fit to a simple model (e.g. BM) means that the estimates of different parameters in more complex models (e.g. different adaptive optima) will be very similar. Because those parameter estimates in turn are used for simulations to assess power, combining small differences in parameters (e.g. adaptive optima) with few species will result in poor ability to distinguish more complex models. This situation is no different than in any other type of statistical test: a small effect size will require a large sample size to show statistical significance. Overall, L80 seems clearly unassociated with colonizing the temperate zone.

References

Adams, M. P., Collier, C. J., Uthicke, S., Ow, Y. X., Langlois, L. and O'Brien, K. R. (2017). Model fit versus biological relevance: evaluating photosynthesis-temperature models for three tropical seagrass species. *Sci. Rep.* **7**, 39930. doi:10.1038/srep39930

AmphibiaWeb. (2021). AmphibiaWeb: Information on amphibian biology and conservation, vol. 2021. Accessed 28 September 2021. http://amphibiaweb.org.

Angilletta, M. J., Jr. (2006). Estimating and comparing thermal performance curves. *J. Therm. Biol.* **31**, 541–545. doi:10.1016/j.jtherbio.2006.06.002

Beaulieu, J. M., Jhwueng, D. C., Boettiger, C. and O'Meara, B. C. (2012). Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evolution* **66**, 2369–2383. doi:10.1111/j.1558-5646.2012.01619.x

Boettiger, C., Coop, G. and Ralph, P. (2012). Is your phylogeny informative? Measuring the power of comparative methods. *Evolution* **66**, 2240–2251. doi:10.1111/j.1558-5646.2011.01574.x

Bulté, G. and Blouin-Demers, G. (2006). Cautionary notes on the descriptive analysis of performance curves in reptiles. *J. Therm. Biol.* **31**, 287–291. doi:10.1016/j.jtherbio.2005.11.030

Butler, M. A. and King, A. A. (2004). Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *Am. Nat.* **164**, 683–695. doi:10.1086/426002

Cressler, C. E., Butler, M. A. and King, A. A. (2015). Detecting adaptive evolution in phylogenetic comparative analysis using the Ornstein-Uhlenbeck model. *Syst. Biol.* **64**, 953–968. doi:10.1093/sysbio/syv043

Edwards, A. W. F. (1972). *Likelihood*. Cambridge, UK: Cambridge University Press.

Emerson, S. B. (1978). Allometry and jumping in frogs: helping the twain to meet. *Evolution* **32**, 551–564. doi:10.2307/2407721

Faivovich, J., Pereyra, M. O., Luna, M. C., Hertz, A., Blotto, B. L., Vásquez-Almazán, C. R., McCranie, J. R., Sánchez, D. A., Baêta, D., Araujo-Vieira, K. et al. (2018). On the monophyly and relationships of several genera of Hylini (Anura: Hylidae: Hylinae), with comments on recent taxonomic changes in hylids. *S. Am. J. Herpetol.* **13**, 1–32. doi:10.2994/sajh-d-17-00115.1

Gomes, F. R., Rezende, E. L., Grizante, M. B. and Navas, C. A. (2009). The evolution of jumping performance in anurans: morphological correlates and ecological implications. *J. Evol. Biol.* **22**, 1088–1097. doi:10.1111/j.1420-9101.2009.01718.x

Gvoždík, L. and Van Damme, R. (2006). *Triturus* newts defy the running-swimming dilemma. *Evolution* **60**, 2110–2121. doi:10.1111/j.0014-3820.2006.tb01848.x

Herrel, A. and Bonneaud, C. (2012). Temperature dependence of locomotor performance in the tropical clawed frog, *Xenopus tropicalis*. *J. Exp. Biol.* **215**, 2465–2470. doi:10.1242/jeb.069765

Ho, L. S. T. and Ané, C. (2013). Asymptotic theory with hierarchical autocorrelation: Ornstein–Uhlenbeck tree models. *Ann. Stat.* **41**, 957–981. doi:10.1214/13-aos1105

Ho, L. S. T. and Ané, C. (2014). Intrinsic inference difficulties for trait evolution with Ornstein-Uhlenbeck models. *Methods Ecol. Evol.* **5**, 1133–1146. doi:10.1111/2041-210x.12285

Huey, R. B. and Stevenson, R. D. (1979). Integrating thermal physiology and ecology of ectotherms: a discussion of approaches. *Am. Zool.* **19**, 357–366. doi:10.1093/icb/19.1.357

Ives, A. R., Midford, P. E. and Garland, T., Jr. (2007). Within-species variation and measurement error in phylogenetic comparative methods. *Syst. Biol.* **56**, 252–270. doi:10.1080/10635150701313830

James, R. S., Navas, C. A. and Herrel, A. (2007). How important are skeletal muscles in setting limits on jumping performance? *J. Exp. Biol.* **210**, 923–933. doi:10.1242/jeb.02731

Jetz, W. and Pyron, R. A. (2018). The interplay of past diversification and evolutionary isolation with present imperilment across the amphibian tree of life. *Nat. Ecol. Evol.* **2**, 850–858. doi:10.1038/s41559-018-0515-5

John-Alder, H., Morin, P. J. and Lawler, S. (1988). Thermal physiology, phenology, and distribution of tree frogs. *Am. Nat.* **132**, 506–520. doi:10.1086/284868

John-Alder, H. B., Barnhart, M. C. and Bennett, A. F. (1989). Thermal sensitivity of swimming performance and muscle contraction in northern and southern populations of tree frogs (*Hyla crucifer*). *J. Exp. Biol.* **142**, 357–372. doi:10.1242/jeb.142.1.357

King, A. A. and Butler, M. A. (2009). ouch: Ornstein-Uhlenbeck models for phylogenetic comparative hypotheses. R package version 2.14-1. https://kingaa.github.io/ouch/

Moen, D. S., Cabrera-Guzmán, E., Caviedes-Solis, I. W., González-Bernal, E., and Hanna, A. R. (2021). Supplementary datasets, data analysis code, and R tutorials for: Phylogenetic analysis of adaptation in comparative physiology and biomechanics:

overview and a case study of thermal physiology in treefrogs. Dryad Dataset, https://doi.org/10.5061/dryad.t4b8gtj2m

Moen, D. S., Smith, S. A. and Wiens, J. J. (2009). Community assembly through evolutionary diversification and dispersal in Middle American treefrogs. *Evolution* **63**, 3228–3247. doi:10.1111/j.1558-5646.2009.00810.x

Navas, C. A. and Araujo, C. (2000). The use of agar models to study amphibian thermal ecology. *J. Herpetol.* **34**, 330–334. doi:10.2307/1565438

Navas, C. A., Antoniazzi, M. M., Carvalho, J. E., Suzuki, H. and Jared, C. (2007). Physiological basis for diurnal activity in dispersing juvenile *Bufo granulosus* in the Caatinga, a Brazilian semi-arid environment. *Comp. Biochem. Physiol. A Mol. Integr. Physiol.* **147**, 647–657. doi:10.1016/j.cbpa.2006.04.035

Navas, C. A., Gomes, F. R. and Carvalho, J. E. (2008). Thermal relationship and exercise physiology in anuran amphibians: integration and evolutionary implications. *Comp. Biochem. Physiol. A Mol. Integr. Physiol.* **151**, 344–362. doi:10.1016/j.cbpa.2007.07.003

O'Meara, B. C. and Beaulieu, J. M. (2014). Modelling stabilizing selection: the attraction of Ornstein-Uhlenbeck models. In *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology* (ed. L. Z. Garamszegi), pp. 381–393. Berlin: Springer-Verlag.

Posada, D. and Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Syst. Biol.* **53**, 793–808. doi:10.1080/10635150490522304

Pyron, R. A. (2014). Biogeographic analysis reveals ancient continental vicariance and recent oceanic dispersal in amphibians. *Syst. Biol.* **63**, 779–797. doi:10.1093/sysbio/syu042

Revell, L. J. and Reynolds, R. G. (2012). A new Bayesian method for fitting evolutionary models to comparative data with intraspecific variation. *Evolution* **66**, 2697–2707. doi:10.1111/j.1558-5646.2012.01645.x

Rezende, E. L. and Bozinovic, F. (2019). Thermal performance across levels of biological organization. *Phil. Trans. Roy. Soc. Lond. B* **374**, 20180549. doi:10.1098/rstb.2018.0549

Salisbury, B. A. and Kim, J. (2001). Ancestral state estimation and taxon sampling density. *Syst. Biol.* **50**, 557–564. doi:10.1080/10635150119819

Silvestro, D., Kostikova, A., Litsios, G., Pearman, P. B., Salamin, N. and Münkemüller, T. (2015). Measurement errors should always be incorporated in phylogenetic comparative analysis. *Methods Ecol. Evol.* **6**, 340–346. doi:10.1111/2041-210x.12337

Smith, S. A., Nieto-Montes de Oca, Reeder, T. W. and Wiens, J. J. (2007). A phylogenetic perspective on elevational species richness patterns in Middle American treefrogs: why so few species in lowland tropical rainforests? *Evolution* **61**, 1188–1207. doi:10.1111/j.1558-5646.2007.00085.x

Snyder, G. K. and Weathers, W. W. (1975). Temperature adaptations in amphibians. *Am. Nat.* **109**, 93–101. doi:10.1086/282976

Sokal, R. R. and Rohlf, F. J. (1995). *Biometry*. New York, NY: W.H. Freeman.

Sunday, J. M., Bates, A. E. and Dulvy, N. K. (2011). Global analysis of thermal tolerance and latitude in ectotherms. *Proc. R. Soc. B* **278**, 1823–1830. doi:10.1098/rspb.2010.1295

Sunday, J., Bennett, J. M., Calosi, P., Clusella-Trullas, S., Gravel, S., Hargreaves, A. L., Leiva, F. P., Verberk, W., Olalla-Tarraga, M. A. and Morales-Castilla, I. (2019). Thermal tolerance patterns across latitude and elevation. *Phil. Trans. Roy. Soc. Lond. B* **374**, 20190036. doi:10.1098/rstb.2019.0036

van Berkum, F. H. (1988). Latitudinal patterns of the thermal sensitivity of sprint speed in lizards. *Am. Nat.* **132**, 327–343. doi:10.1086/284856

Whitehead, P. J., Puckridge, J. T., Leigh, C. M. and Seymour, R. S. (1989). Effect of temperature on jump performance of the frog *Limnodynastes tasmaniensis*. *Physiol. Zool.* **62**, 937–949. doi:10.1086/physzool.62.4.30157938

Wiens, J. J., Kuczynski, C. A., Hua, X. and Moen, D. S. (2010). An expanded phylogeny of treefrogs (Hylidae) based on nuclear and mitochondrial sequence data. *Mol. Phylogenet. Evol.* **55**, 871–882. doi:10.1016/j.ympev.2010.03.013

Wiens, J. J., Pyron, R. A. and Moen, D. S. (2011). Phylogenetic origins of local-scale diversity patterns and the causes of Amazonian megadiversity. *Ecol. Lett.* **14**, 643–652. doi:doi:10.1111/j.1461-0248.2011.01625.x

Wilson, R. S. (2001). Geographic variation in thermal sensitivity of jumping performance in the frog *Limnodynastes peronii*. *J. Exp. Biol.* **204**, 4227–4236. doi:10.1242/jeb.204.24.4227

Wilson, R. S. and Franklin, C. E. (2000). Inability of adult *Limnodynastes peronii* (Amphibia: Anura) to thermally acclimate locomotor performance. *Comp. Biochem. Physiol. A Mol. Integr. Physiol.* **127**, 21–28. doi:10.1016/s1095-6433(00)00238-5

Woods, H. A., Kingsolver, J. G., Fey, S. B., Vasseur, D. A. and Kriticos, D. (2018). Uncertainty in geographical estimates of performance and fitness. *Methods Ecol. Evol.* **9**, 1996–2008. doi:10.1111/2041-210x.13035

Zug, G. R. (1978). Anuran locomotion – structure and function. 2. Jumping performance of semiaquatic, terrestrial, and arboreal frogs. *Sm. C. Zool.* **276**, 1–31.

Zug, G. R. (1985). Anuran locomotion: fatigue and jumping performance. *Herpetologica* **41**, 188–194.

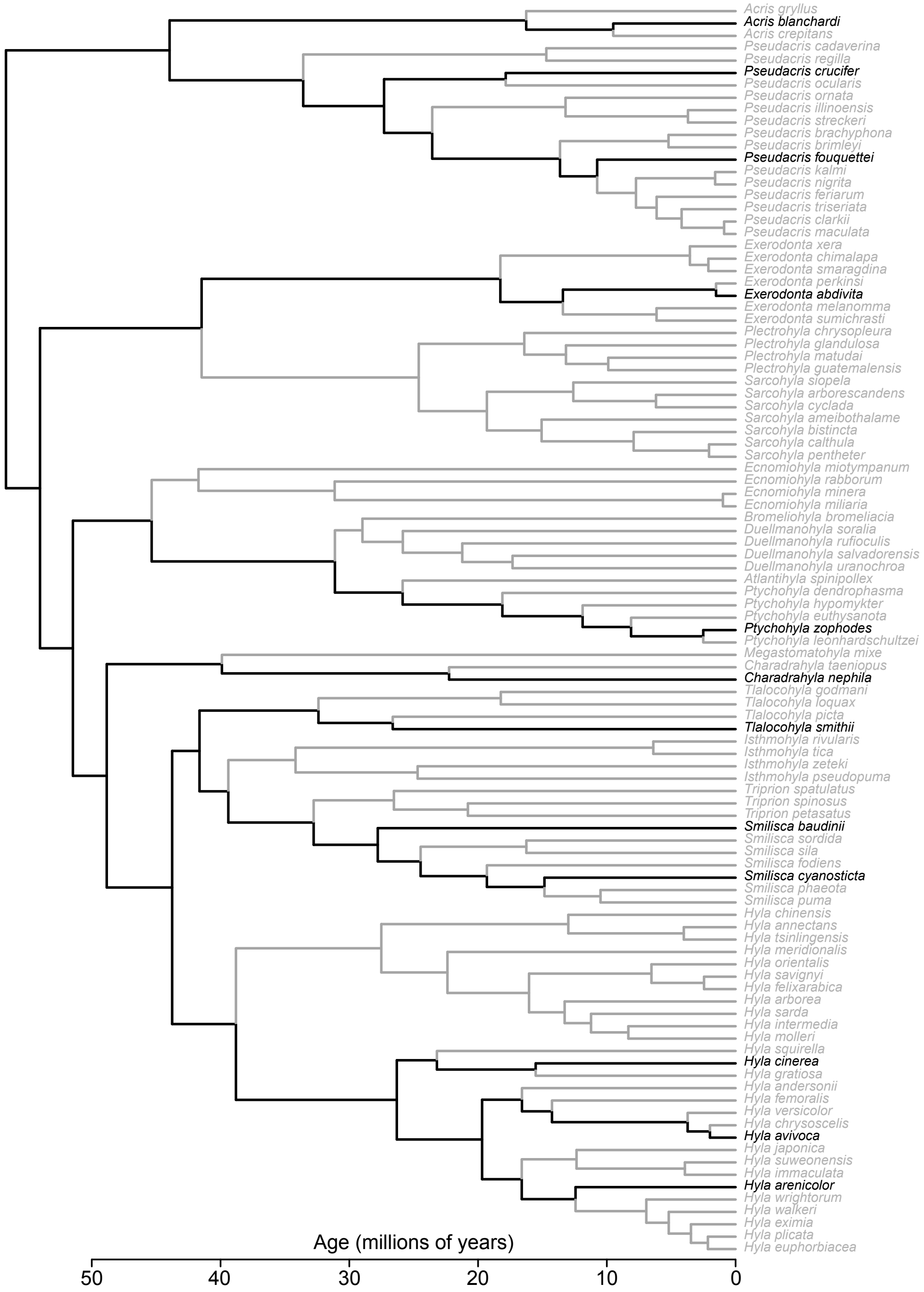

Fig. S1. Phylogeny of the Middle American clade (MAC), showing how our sampling fit within the larger group. As for Fig. 2, we summarized the posterior distribution of Jetz and Pyron (2018) for all species for which they had genetic data. Here, that total was 101 of 197 species within the MAC. We generated this summary tree in the same way as we described in the main text for Fig. 2; the latter is effectively a pruned version of the phylogeny in this figure, in which all unsampled taxa were removed. We indicate the species we sampled for this study with black text and black branches. Species we did not sample are indicated in gray. All branches are supported by Bayesian posterior probabilities of 1.0. Taxonomy follows Faivovich et al. (2018) and AmphibiaWeb (2021).

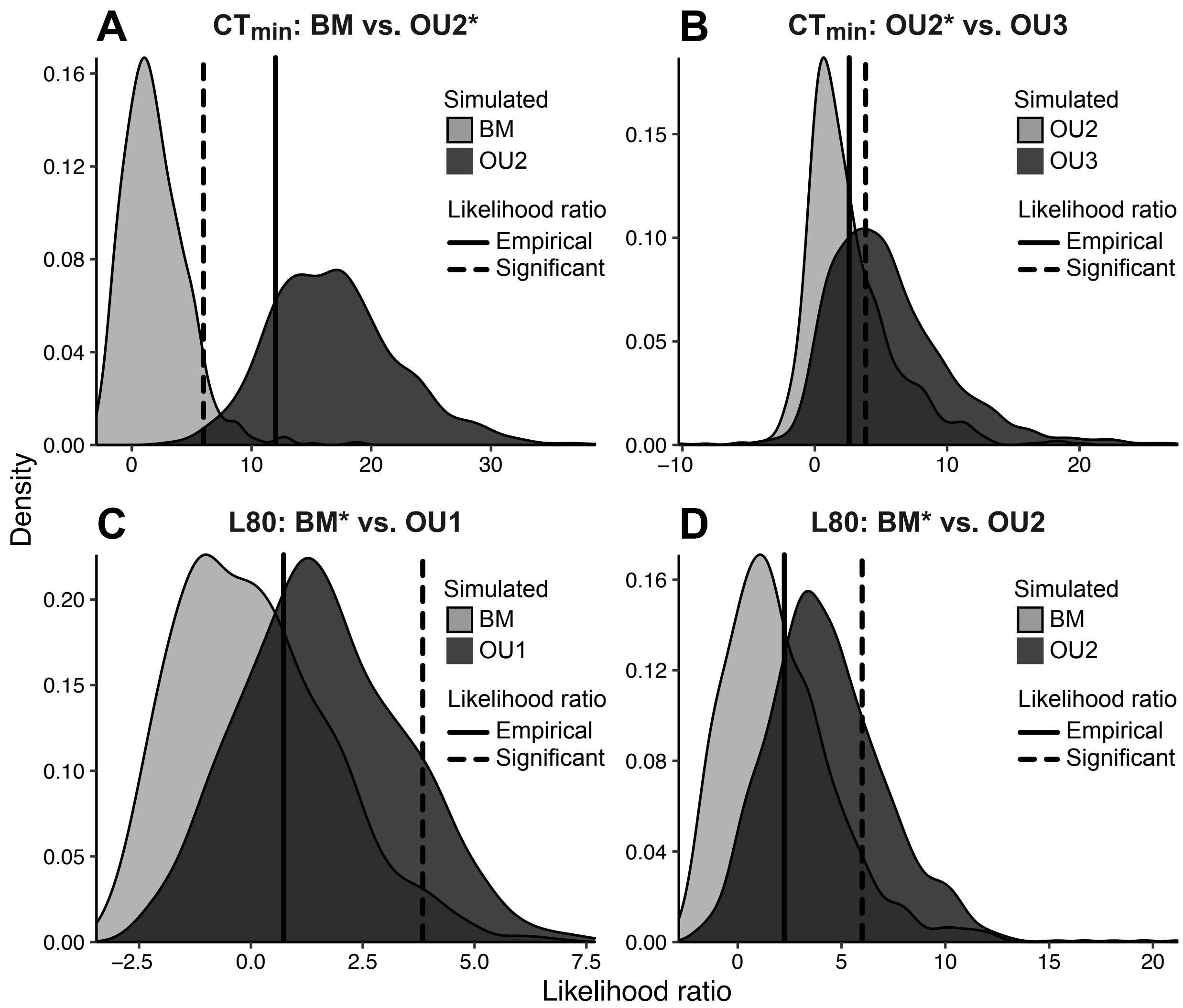

Fig. S2. Parametric bootstrapping results. Each panel shows the results of parametric bootstrapping of each trait and model comparison, with the most highly supported model by AICc in each comparison indicated with an asterisk. In all cases, the first, simplest model (BM or OU2) was used to simulate trait data for the null distributions (light gray) and the second, more complex model (OU1, 2, or 3) was used to simulate trait data for the test distribution (dark gray). Each distribution represents the likelihood ratios resulting from fitting both models on the simulated datasets; distribution overlap is indicated by the darkest gray color. Dashed vertical lines indicate the likelihood ratio that would result in rejecting the simpler model in favor of the more complex model. Thus, the proportion of the null distributions to the right of these lines indicate the probability of incorrectly rejecting the simpler model (i.e. Type-I error rates). The proportion of the test distributions to the right of the dashed line indicate statistical power to reject the simpler model. The solid vertical lines indicate the empirical likelihood ratios calculated from our data. This ratio strongly favors the more complex model when it occurs (1) to the right of the dotted line, (2) outside the null (light gray) distribution, and (3) within the test distribution (dark gray). This only occurs in (A), given that the simpler model was favored in all of our other model comparisons (Table 2).

Table S1. Comparison of two models of how jumping performance varies with respect to temperature in each species.

	AICc		w_i	
Species	Polynomial	Gaussian	Polynomial	Gaussian
Acris blanchardi	**-84.842**	-84.619	**0.528**	0.472
Pseudacris crucifer	**-104.903**	-104.008	**0.610**	0.390
Pseudacris fouquettei	**-70.711**	-70.555	**0.519**	0.481
Hyla arenicolor	**-92.791**	-92.150	**0.579**	0.421
Hyla avivoca	-109.360	**-110.505**	0.361	**0.639**
Hyla cinerea	**-95.131**	-93.927	**0.646**	0.354
Charadrahyla nephila	-65.132	**-66.708**	0.313	**0.687**
Exerodonta abdivita	**-83.788**	-83.628	**0.520**	0.480
Ptychohyla zophodes	**-66.022**	-65.026	**0.622**	0.378
Smilisca baudinii	**-62.666**	-58.781	**0.875**	0.125
Smilisca cyanosticta	**-81.882**	-81.761	**0.515**	0.485
Tlalocohyla smithii	**33.527**	33.541	**0.502**	0.498

AICc = small-sample-size corrected Akaike information criterion. w_i = AICc weight of each model (within species). For each species, we indicate the optimal model (i.e. lowest AICc and highest w_i) in bold.

Table S2. Evolutionary model comparisons with alternative temperature thresholds for peak performance, showing highly similar results across variables.

Variable: model	ln *L*	AICc	w_i
L70:			
Brownian motion	**-36.376**	**78.086**	**0.780**
OU single optimum	-35.949	80.897	0.191
OU2	-35.521	84.756	0.028
OU3	-35.473	90.946	0.001
L80:			
Brownian motion	**-35.342**	**76.018**	**0.778**
OU single optimum	-34.977	78.953	0.179
OU2	-34.220	82.154	0.036
OU3	-32.838	85.677	0.006
L90:			
Brownian motion	**-35.242**	**75.817**	**0.699**
OU single optimum	-34.433	77.866	0.251
OU2	-34.206	82.126	0.030
OU3	-31.429	82.857	0.021
Peak temperature:			
Brownian motion	**-32.688**	**70.710**	**0.822**
OU single optimum	-32.891	74.783	0.107
OU2	-32.884	79.483	0.010
OU3	-27.972	75.944	0.060
Breadth:			
Brownian motion	**-33.192**	**71.717**	**0.555**
OU single optimum	-31.991	72.983	0.295
OU2	-30.381	74.475	0.140
OU3	-29.856	79.712	0.010

Variables represent the threshold at which we considered peak performance to decline: L70 = the lower temperature at which the jumping velocity of a species reached 70% of its peak; L80 = the 80% threshold; L90 = the 90% threshold; peak temperature = the temperature of peak jumping performance; breadth = the difference between the peak temperature and L80. We compared four models, including Brownian motion and three Ornstein-Uhlenbeck (OU) models. OU single optimum = one optimal temperature for all species. OU2 = separate temperature optima for tropical and temperate lineages. OU3 = separate temperature optima for temperate, lowland tropical, and highland tropical lineages. ln *L* = log-likelihood. AICc = corrected Akaike information criterion. w_i = AICc model weight. For each variable, we indicate the optimal model (i.e. the lowest AICc and highest w_i) in bold.

Table S3. Comparison of evolutionary modeling results for L80 with and without the tropical lowland species *Tlalocohyla smithii*.

Model	With *Tlalocohyla*			Without *Tlalocohyla*		
	ln *L*	AICc	w_i	ln *L*	AICc	w_i
Brownian motion	**-35.342**	**76.018**	**0.778**	**-32.864**	**71.229**	**0.813**
OU single optimum	-34.977	78.953	0.179	-32.520	74.469	0.161
OU2	-34.220	82.154	0.036	-31.841	78.349	0.023
OU3	-32.838	85.677	0.006	-30.071	82.143	0.003

We compared four models: Brownian motion, an Ornstein-Uhlenbeck (OU) model with a single optimal temperature for all species, an OU model in which temperate and tropical species had different optimal temperatures (OU2), and an OU model in which temperate, lowland tropical, and highland tropical species all had different temperature optima (OU3). ln *L* = log-likelihood. AICc = corrected Akaike information criterion. w_i = AICc weight of a model. For each analysis, we indicate the optimal model (i.e. the lowest AICc and highest w_i) in bold.

Table S4. Comparison of evolutionary modeling results for CT_{min} both with and without using standard errors in the analyses.

Model	With standard error			Without standard error		
	ln *L*	AICc	w_i	ln *L*	AICc	w_i
Brownian motion	-24.720	54.941	0.209	-24.745	54.990	0.180
OU single optimum	-25.870	61.170	0.009	-26.016	61.460	0.007
OU2	**-18.888**	**52.442**	**0.729**	**-18.742**	**52.150**	**0.744**
OU3	-17.843	57.686	0.053	-17.443	56.885	0.070

We compared four models: Brownian motion, an Ornstein-Uhlenbeck (OU) model with a single optimal temperature for all species, an OU model in which temperate and tropical species had different optimal temperatures (OU2), and an OU model in which temperate, lowland tropical, and highland tropical species all had different temperature optima (OU3). ln *L* = log-likelihood. AICc = corrected Akaike information criterion. w_i = AICc weight of a model. Note that Brownian motion is not nested within OU models in the way we estimated them (O'Meara and Beaulieu, 2014) and so may not necessarily have a lower log-likelihood. For each analysis, we indicate the optimal model (i.e. the lowest AICc and highest w_i) in bold.