Gaussian processes on phylogenies: a kernel framework for building, checking and comparing models of trait evolution

Thomas A. Püschel
Institute of Human Sciences, School of Anthropology and Museum Ethnography,
University of Oxford, Oxford, United Kingdom
Correspondence: Thomas A. Püschel, thomas.puschel@anthro.ox.ac.uk

Abstract

1. Phylogenetic comparative methods for continuous traits (phylogenetic generalised least squares, Brownian motion and Ornstein–Uhlenbeck models, Pagel's transformations, early-burst and multi-rate models, phylogenetic mixed models) can all be regarded as Gaussian processes on a phylogeny that differ only in their covariance function, or kernel. This link has been noted before, but it has not been developed into a working framework in which models are built, checked and compared in the same way.
2. In the present study, I write the classical models as kernels in one notation and state where each representation holds. Most of them share a branch-wise form, which makes it possible to compute likelihoods, leave-one-out and leave-clade-out predictive densities, and predictions at tips and internal nodes in a time that grows linearly with the number of species. I also collect the conditions under which kernels defined on patristic distances are valid covariances, provide a numerical test of whether two kernels imply the same distribution for the tips, and write four extensions as kernels (i.e. autocorrelated rates, smooth rate-through-time functions, joint phylogenetic and spatial structure, and phylogenetically varying slopes). I assessed

implementation equivalence against standard software and the statistical behaviour of the extensions by simulation on pure-birth trees and on subtrees of an empirical phylogeny. As an empirical example, I applied the framework to brain–body allometry in 1,349 mammal species. All methods are implemented in the R package `phylogp`.

3. A generic implementation in Stan reproduced standard software to numerical precision in 698 of 700 simulated data sets, on ultrametric trees, trees with fossil tips and trees with polytomies. The remaining two were differences between optimisers at a parameter boundary. Any non-increasing function of patristic distance is a valid kernel on ultrametric trees. However, when fossil tips are included only completely monotone functions (e.g. the exponential) are guaranteed to be valid, and the squared-exponential kernel failed at 73% of the length-scales examined. The extended kernels were never preferred over simpler models when the process was absent. In addition, power was limited for weak effects and abrupt rate shifts were shrunk. In the mammal example, most of the apparent predictive gain of the extended kernels came from heavy-tailed residuals and curvature, and leaving out whole orders reversed the advantage of varying slopes.
4. Overall, treating comparative methods as Gaussian processes offers a common and extensible basis both for developing new models of trait evolution and for checking them before they are used. Therefore, predictive comparisons should be made against simpler explanations that leave the same statistical signature.

Keywords: Gaussian process; kernel; phylogenetic comparative methods; Brownian motion; Ornstein–Uhlenbeck; R package

Introduction

Species are not independent data points. Closely related species tend to resemble each other because they share a common ancestry and, since Felsenstein (1985) and Grafen (1989), a large toolbox of phylogenetic comparative methods (PCMs) has been developed to account for this and to test hypotheses about the tempo and mode of trait evolution. For continuous traits, this toolbox includes phylogenetic generalised least squares [PGLS; Grafen (1989); Martins & Hansen (1997)], Brownian motion (BM) and Ornstein–Uhlenbeck (OU) models (Butler & King, 2004; Felsenstein, 1973; Hansen, 1997), Pagel's λ, κ and δ transformations (M. Pagel, 1999), early-burst and accelerating–decelerating (ACDC) models (Blomberg et al., 2003; Harmon et al., 2010), multi-rate BM (Eastman et al., 2011; O'Meara et al., 2006; Venditti et al., 2011), phylogenetic mixed models (Hadfield & Nakagawa, 2010; Housworth et al., 2004; Ives & Helmus, 2011; Lynch, 1991) and ancestral-state reconstruction (Garland & Ives, 2000; Schluter et al., 1997). Each of them comes with its own software (e.g. Paradis & Schliep, 2019; Clavel et al., 2015; Ho & Ané, 2014a; Pennell et al., 2014; Revell, 2012) and parameterisation, and consequently it is easy to lose sight of how closely these models are related and of which of their assumptions actually matter (Uyeda et al., 2018).

In statistics and machine learning, a Gaussian process (GP) is a distribution over functions fully specified by a mean function and a covariance function, or *kernel* (Rasmussen & Williams, 2006). GPs underlie kriging (Cressie, 1993; Diggle & Ribeiro, 2007), Gaussian random fields in spatial statistics and ecology (Lindgren et al., 2011), and a large share of modern nonparametric regression. One of their main advantages is modularity, since once the data are modelled as $\mathbf{y} \sim \mathcal{N}(\mathbf{m}, \mathbf{K})$ all the scientific assumptions are contained in the kernel $\mathbf{K}$, kernels can be added, multiplied and composed (Duvenaud et al., 2013), and inference, prediction and model comparison then follow rules that do not depend on the kernel.

The connection between PCMs and GPs is not new and has been mentioned before. The multivariate normal likelihood of BM on a tree is the starting point of PGLS, and Hansen & Martins (1996) derived the interspecific correlation structures implied by several microevolutionary processes,

which are in practice kernels. General Gaussian models on trees with linear-time likelihoods have been formalised by Ho & Ané (2014a), Mitov et al. (2020) and Bastide et al. (2021), whose linear-time gradients also make Bayesian relaxed random walks practical on large trees (Fisher et al., 2021), and Manceau et al. (2017) proposed a unifying framework in which a wide family of models yields Gaussian distributions at the tips. "Phylogenetic Gaussian process regression" has been used for function-valued traits (Hadjipantelis et al., 2013; N. S. Jones & Moriarty, 2013; Meagher et al., 2018), OU-type kernels for primate brain size (McElreath, 2020), and phylogenetic and geographic kernels in linguistics (Guzmán Naranjo & Becker, 2022; Hartmann & Jäger, 2023) and conservation biology (Matsuba et al., 2024). Hence, most of the pieces already exist, although they are scattered across different disciplines and are rarely described in terms of kernels.

In the present study, I bring this material together and extend it. I (i) write the classical continuous-trait PCMs as kernels on trees in one notation, and state where each representation holds; (ii) check numerically that a generic GP implementation reproduces the standard software; (iii) use the GP view to make the relationships among models explicit, and collect the conditions under which a distance-based kernel is valid on a tree (these follow from classical results that I apply rather than discover); and (iv) extend the toolbox with four additional biologically motivated models, each with close precedents in the literature, which are built by kernel design and kept linear in the number of species wherever the tree structure allows it. To my knowledge, this is the first time that these models have been organised and extended systematically as kernels on trees.

It is important to keep in mind that these claims rest on two different kinds of evidence, which I keep separate throughout. The first one is *implementation equivalence* (i.e. whether a GP with the right kernel reproduces existing software), which is a matter of algebra and numerics. The second one is *statistical validity* (i.e. whether a model recovers the process it is meant to detect, with calibrated uncertainty and acceptable false-positive rates). It is important to note that equivalence does not establish validity, as two implementations can agree perfectly and yet still behave badly. The statistical properties of the classical models have been documented elsewhere (e.g. Ho & Ané,

2014b; Cooper et al., 2016); therefore, I only evaluated the extensions, by replicated simulation including null scenarios. Finally, I illustrate the workflow on brain–body allometry in mammals (Burger et al., 2019; Upham et al., 2019), and all methods are implemented in the R package `phylogp`.

Materials and Methods

Models as kernels on a tree

A GP $f \sim \mathcal{GP}(m, k)$ on an index set is a collection of random variables, any finite subset of which is jointly Gaussian with mean m and covariance k. The kernel must be positive semi-definite. For observations $\mathbf{y} = f(\mathbf{x}) + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \Sigma_\varepsilon)$, the log marginal likelihood is $-\frac{1}{2}(\mathbf{y}-\mathbf{m})^\top \mathbf{V}^{-1}(\mathbf{y}-\mathbf{m}) - \frac{1}{2}\log|\mathbf{V}| - \frac{N}{2}\log 2\pi$ with $\mathbf{V} = \mathbf{K}_\theta + \Sigma_\varepsilon$, and predictions at new points are Gaussian conditionals. Viewing the phylogeny $\mathcal{T}$ as a continuous object in which every point is an ancestral lineage, BM with rate σ^2 is the GP with kernel $k_{\mathrm{BM}}(a, b) = \sigma^2 h(a \wedge b)$, where h is the distance from the root and $a \wedge b$ the most recent common ancestor. At the tips this is the familiar $\sigma^2\mathbf{C}$. At internal nodes it gives ancestral-state reconstruction and, at tips with missing data, imputation, both of which are kriging on a tree. BM is also Markov along the tree. Hence the likelihood can be evaluated in $O(N)$ operations by Gaussian pruning (Bastide et al., 2021; Felsenstein, 1973; Freckleton, 2012; Ho & Ané, 2014a; Mitov et al., 2020).

Two distinctions matter for what I do and do not claim. First, a GP on $\mathcal{T}$ is defined at every point of the tree, whereas data live at the tips. BM, early burst, multi-rate BM and OU are processes on the whole tree, and therefore reconstruction at internal points is meaningful for them. Pagel's λ, on the other hand, is specified only through the tip covariance. It fixes a distribution for the tips but not for internal points, which means that two models can have identical tip distributions and still be different processes. Second, many models can be written in an *edge form*,

$$\mathbf{K}(\theta) = \mathbf{B}\,\mathrm{diag}\{\mathbf{w}(\theta)\}\,\mathbf{B}^\top + v_0\mathbf{1}\mathbf{1}^\top + \mathrm{diag}(\tau^2), \tag{1}$$

where **B** is the tip-by-edge incidence matrix, $w_e \geq 0$ is the variance accumulated along edge e, v_0 is the variance of the root state and τ_i^2 is tip-specific variance (measurement error or within-species variation). The pruning likelihood applies to every model of this form. OU with $\alpha > 0$ has this form only on ultrametric trees. On other trees its stationary kernel is still valid and the likelihood is still linear, but pruning needs edge-specific coefficients (Ho & Ané, 2014a). The software provided in this study implements the OU edge form for ultrametric trees only.

Table 1 lists the classical models. Pagel's κ raises each branch length to the power κ, and δ transforms node heights. Pagel's λ on an ultrametric tree is BM plus white noise, i.e. the phylogenetic mixed model (Housworth et al., 2004; Lynch, 1991) with λ as the phylogenetic heritability. OU with a stationary root is the exponential (Matérn-1/2) kernel on the patristic distance d_{ij}. PGLS replaces the mean μ by $\mathbf{X}\beta$, multivariate BM is the separable GP $\mathrm{vec}(\mathbf{Y}) \sim \mathcal{N}(\cdot, \mathbf{R} \otimes \mathbf{C})$, and non-Gaussian traits enter as latent GPs, with the threshold model (Felsenstein, 2005) as the probit special case.

Table 1: Table 1. Classical continuous-trait models as kernels on a tree. t_{ij}: shared path length from the root to the common ancestor of tips i and j; d_{ij}: patristic distance; T: tree height; ℓ_e, h_e^-, h_e^+: length and start and end heights of edge e. All rows have the edge form (1); the OU rows do so on ultrametric trees only, and for λ and the OU edge form the rows give the distribution at the tips, not at internal points.

Model	Kernel $k(i,j)$	Edge variance w_e
BM	$\sigma^2 t_{ij}$	$\sigma^2 \ell_e$
Pagel's λ	$\sigma^2[\lambda t_{ij} + (1-\lambda)T\,\mathbb{1}_{i=j}]$	$\sigma^2\lambda\ell_e$ ($+(1-\lambda)h_e^+$ on tip edges)
Pagel's κ	$\sigma^2 \sum_{e\,\text{shared}} \ell_e^\kappa$	$\sigma^2 \ell_e^\kappa$
Pagel's δ	$\sigma^2 T^{1-\delta} t_{ij}^\delta$	$\sigma^2 T^{1-\delta}[(h_e^+)^\delta - (h_e^-)^\delta]$
Early burst / ACDC	$\sigma^2(e^{r t_{ij}} - 1)/r$	$\sigma^2(e^{r h_e^+} - e^{r h_e^-})/r$
OU, fixed root	$\frac{\sigma^2}{2\alpha} e^{-\alpha d_{ij}}(1 - e^{-2\alpha t_{ij}})$	$\frac{\sigma^2}{2\alpha} e^{-2\alpha T}(e^{2\alpha h_e^+} - e^{2\alpha h_e^-})$
OU, random root	$\frac{\sigma^2}{2\alpha} e^{-\alpha d_{ij}}$	as above, with $v_0 = \frac{\sigma^2}{2\alpha} e^{-2\alpha T}$
Multi-rate BM	$\sum_{e\,\text{shared}} \sum_r \sigma_r^2 \ell_{e,r}$	$\sum_r \sigma_r^2 \ell_{e,r}$

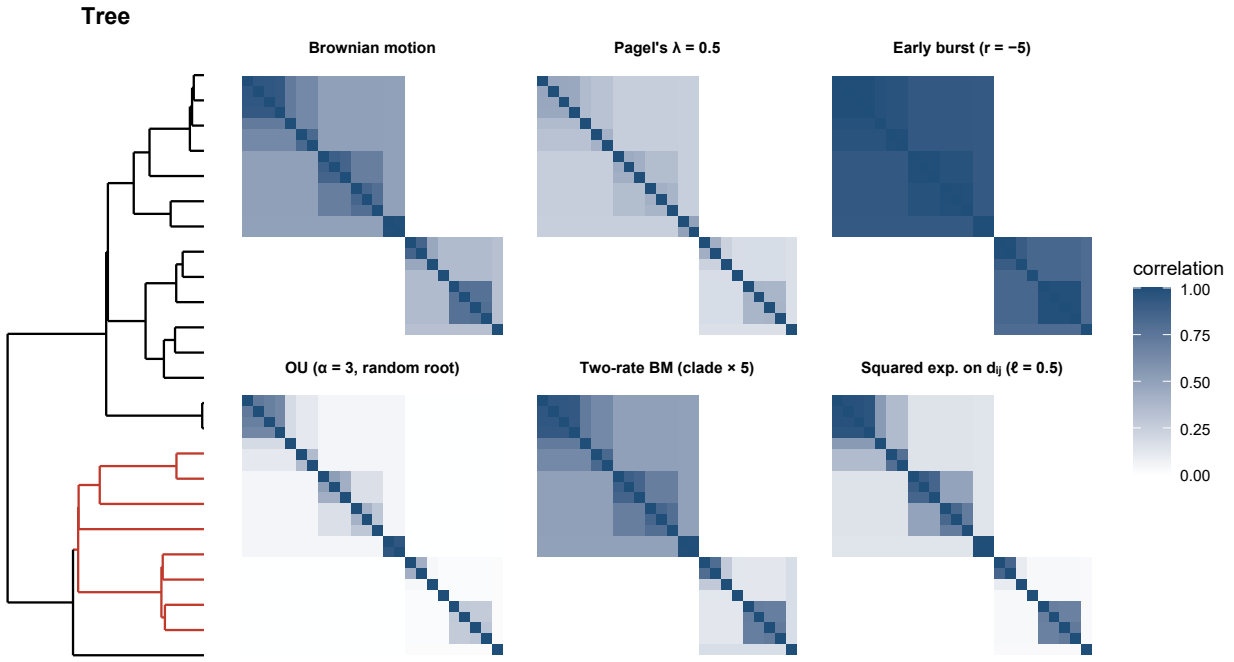


Figure 1. One 24-species tree (left; red: the clade given an elevated rate in the two-rate model) seen through six kernels. Tips are in the same order in the tree and in the correlation matrices, so clades appear as blocks. Each matrix shows the implied correlation between species. The kernel alone determines the model: BM, Pagel's λ, early burst, OU, multi-rate BM and a squared-exponential kernel on patristic distance are all Gaussian processes on the same tree.

OU and accelerating rates. The edge form makes a known identifiability result (Ho & Ané, 2014b; Murrell, 2018; Uyeda et al., 2015) transparent, and makes it possible to state exactly when it holds. On an ultrametric tree of height T, $d_{ij} = 2(T - t_{ij})$. With a fixed root, single-optimum OU has covariance $\frac{\sigma^2}{2\alpha} e^{-2\alpha T}(e^{2\alpha t_{ij}} - 1)$ and a common mean, which is exactly the ACDC kernel with $r = 2\alpha$, so the two models have identical likelihoods. With a stationary root the covariance gains a constant (i.e. a random intercept). Once the intercept is profiled out, the OU log-likelihood equals the ACDC one minus $\frac{1}{2}\log(1 + c\,\mathbf{1}^\top \mathbf{V}^{-1}\mathbf{1})$, so its maximum cannot exceed the ACDC maximum, and the REML likelihoods are identical (Appendix S1). Please note that this is an equivalence of the tip distributions only, because as processes on the whole tree the two models differ, and so do their reconstructions of internal states, and that it does not extend to trees with fossil tips or to OU models with several optima.

Kernel design and validity

One of the advantages of the GP view of PCMs is that model building becomes kernel design (Duvenaud et al., 2013; Rasmussen & Williams, 2006). Sums of kernels describe independent processes (e.g. phylogenetic, spatial and residual variation; Freckleton & Jetz (2009)), products describe interactions (e.g. $x_i x_j k_{\text{BM}}(i,j)$ is a regression slope that evolves by BM), and kernel parameters can have their own GP priors, giving hierarchical, non-stationary models such as log-rates that evolve along the tree (Thorne & Kishino, 2002).

A kernel must be positive semi-definite on the tree at hand. For $k(i,j) = \phi(d_{ij})$, validity depends on tree shape. On an ultrametric tree any non-negative, non-increasing ϕ is a non-decreasing function of shared time, so it is a tree transformation of the form (1) and always valid (Dellacherie et al., 2014). On trees with fossil tips or without a clock this argument fails. Tree metrics are of negative type (Deza & Laurent, 1997), so by Schoenberg's theorem (Schoenberg, 1938) the exponential kernel $e^{-d/\ell}$ remains valid for every length-scale, and by Bernstein's theorem so does every completely monotone ϕ, e.g. the generalised Cauchy form $(1 + d/\ell)^{-a}$ and the stretched exponential $e^{-(d/\ell)^p}$ with $0 < p \leq 1$. It is relevant to bear in mind that this is a sufficient condition and not a characterisation of all the kernels that happen to be valid on a particular tree. In contrast, the squared-exponential kernel, the default in most GP software and the "L2" kernel of McElreath (2020), is guaranteed valid only for Euclidean distances (cf. Feragen et al., 2015), and the same caveat applies to Matérn kernels with smoothness above 1/2. Proofs and a 5-tip counterexample are in Appendix S1.

Implementation and model comparison

I implemented Eq. (1) in Stan (Carpenter et al., 2017) through `cmdstanr` (Gabry et al., 2024). The program fits regressions $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{K}(\theta) + \text{diag}(\mathbf{s}^2))$ for the kernels in Table 1 with three numerically equivalent likelihoods: the $O(N)$ pruning recursion, a dense kernel built from (1), and the closed-form kernels written in t_{ij} and d_{ij}. Maximum-likelihood (ML) estimates used L-BFGS

with no priors, so that the posterior mode equals the ML estimate. Bayesian fits used four chains with weakly informative priors, and convergence was checked with $\hat{R}$, effective sample sizes and divergent transitions. The R package `phylogp` writes the Stan program for any sum or product of kernels, whereas the multivariate and non-Gaussian models are fitted with the archived Stan programs. Since version 0.2.0, the package has added a kernel for rates that depend on regimes mapped on the tree (`k_regime()`, which reproduces `phytools::brownie.lite()`), several records per species with within-species variance and known measurement error, and predictions for new species under every kernel (Appendix S5).

For model comparison I used leave-one-out (LOO) predictive densities (Vehtari et al., 2017). For a Gaussian model, the density of a left-out observation given the parameters is available in closed form (Sundararajan & Keerthi, 2001). However, this is not the quantity that LOO targets, which integrates over the posterior without observation i. I therefore approximate that integral by Pareto-smoothed importance sampling (PSIS) of the posterior draws obtained with all the data (Bürkner et al., 2021; Vehtari et al., 2017). Its reliability is indicated by the Pareto-$\hat{k}$ diagnostic, and where $\hat{k} > 0.7$ I replaced the estimate by an exact refit without that observation. For models with latent tree processes, the left-out species' own latent values must not be conditioned on, because only its own observation informs them. For the varying-slope kernel I integrated them out analytically, and for the relaxed-rate kernel I integrated out the terminal log-rate by Gauss–Hermite quadrature. For kernels of the form (1), an upward–downward pass of Gaussian messages gives all N conditional LOO densities exactly in $O(N)$ per draw (agreement with the dense calculation to 10^{-10}). I also left out whole clades [`loo_group()`] and pooled posteriors across a sample of trees.

Four kernel extensions

None of the four extensions is new as a statistical model, and each has close precedents. (a) *Relaxed rates*: the log rate evolves by BM along the tree, $\log w_e/\ell_e = \mu + \frac{1}{2}[z(a_e) + z(d_e)]$ with $z \sim \mathcal{GP}(0, \tau^2 k_{\mathrm{BM}})$, a Bayesian counterpart of autocorrelated-rate BM (Revell, 2021; Thorne & Kishino, 2002) and close to the relaxed random walk of Fisher et al. (2021). (b) *Rate through time*:

$w_e = \int \sigma^2(s)\, ds$ over the edge, with $\log \sigma^2$ piecewise constant in 20 bins under a Gaussian random-walk prior, in the spirit of Clavel & Morlon (2017). It approximates but does not contain early burst as a special case. (c) *Phylogeny plus geography*: $\mathbf{K} = \sigma_p^2\mathbf{C} + \sigma_g^2 e^{-\mathbf{D}_{\text{geo}}/\rho} + \sigma_e^2\mathbf{I}$ (Freckleton & Jetz, 2009). (d) *Varying slopes*: $y_i = \beta_0 + (\beta_1 + v_i)x_i + u_i + e_i$ with u, v independent BM processes, whose marginal kernel is that of a phylogenetic random-slope mixed model (Hadfield & Nakagawa, 2010; Matsuba et al., 2024). As such, what the kernel view provides here is that all four can be written, fitted and checked within the same framework. To keep the cost linear in N, (a), (b) and (d) sample latent node values and integrate out the other Gaussian components by pruning.

Equivalence with standard software

For each of seven models (BM, λ, κ, δ, early burst, OU with fixed and random root) I simulated 100 data sets on a 100-tip pure-birth tree under $y = 1 + 0.5x + e$, and fitted them by ML with `phylolm` and with the GP using the same parameter bounds. For one data set per model I also compared `geiger` (Pennell et al., 2014), `nlme` (Pinheiro et al., 2023) and `caper` (Orme et al., 2023). I repeated the `phylolm` comparison on four more trees: a non-clock tree (100 tips), a birth–death tree retaining extinct lineages (143 tips), an ultrametric tree with ten polytomies (100 tips), and a 1,000-tip ultrametric tree. I also compared GP kriging with `phytools::fastAnc` (Revell, 2012), the λ model with BM plus measurement error, bivariate BM and a two-regime BM with `mvMORPH` (Clavel et al., 2015), and binary and count phylogenetic GLMMs with `brms` (Bürkner, 2017).

Simulation study of the extensions

To assess bias, calibration, power and false positives, I simulated each kernel with the process present and absent on freshly simulated pure-birth trees: 50 replicates per scenario for (a), (b) and (d) (100 tips) and 30 for the dense model (c) (80 tips). Effects were an 8-fold rate increase in a ~20-tip clade (a), a mid-history rate pulse (b), $\sigma_g = 1$ with $\rho = 0.3$ (c) and $\sigma_v = 0.6$ (d); nulls were constant-rate BM (a, b), $\sigma_g = 0$ and $\sigma_v = 0$. Each extension was compared with its simpler alternative (BM for a and b, phylogeny plus white noise for c, constant-slope PGLS with λ for d) by PSIS-LOO,

and counted as detected if its expected log predictive density (elpd) exceeded the alternative's by more than two standard errors. I report detection rates, coverage of nominal 90% intervals, bias and sampler diagnostics. My first runs of the rate-through-time kernel produced divergent transitions in 92–98% of replicates, and I therefore changed the prior. Because this was done after seeing the results, I report both runs (Appendix S3). To probe limits, I added weaker effects, a 300-tip tree, two misspecified data-generating processes for the relaxed-rate kernel (OU, and BM with three jumps), random 100-species subtrees of the mammal tree of Upham et al. (2019) (empirical tree shapes), and a head-to-head with `multirateBM` (Revell, 2021) on 20 mammal subtrees of 50 species (default settings, 30-minute limit per fit). I also computed the smallest eigenvalue of exponential, Matérn and squared-exponential kernels on an ultrametric, a fossil and a non-clock tree, timed pruning against dense likelihoods on trees of 50 to 10,000 tips, and simulated fossil trees (80–300 tips) with OU or ACDC data to see whether the two models can be separated.

Case study: mammalian brain size

I used mean adult brain and body masses (Burger et al., 2019) and the DNA-only maximum clade credibility tree of Upham et al. (2019), modelling $\log_{10}$ brain mass on $\log_{10}$ body mass centred at its mean. A more complex kernel can predict better without the process it encodes having operated. Varying slopes can capture variance that grows with body size, relaxed rates on terminal branches act as heavy-tailed species-specific residuals, and both can absorb curvature, which is documented in brain–body allometry (Venditti et al., 2024). I therefore crossed three process kernels (BM, varying slopes, relaxed rates) with three residual models (Gaussian; heteroscedastic, $k_{\text{slope}}(x) \times k_{\text{white}}$; Student-$t$ with four degrees of freedom) and with a linear or quadratic mean. All models were fitted with `phylogp` (four chains, 1,000 warm-up and 1,000 sampling iterations) and compared by LOO with exact refits for every species with $\hat{k} > 0.7$. I left out each of 28 orders for four key models, and refitted those models to 50 trees from the posterior of Upham et al. (2019). For geography I used range midpoints from PanTHERIA (K. E. Jones et al., 2009) and great-circle distances, for which the exponential kernel is valid (Gneiting, 2013). The analysis is an illustration of the workflow, and

it is not intended as a test of hypotheses about brain evolution.

Results

Implementation equivalence

Across the 700 simulated data sets, ML estimates from the GP and from `phylolm` fell on the 1:1 line (Fig. 2). For all models except λ, the largest difference in log-likelihood was 1.9×10^{-7} and in the regression slope 9.0×10^{-6}, i.e. optimiser tolerance. The exceptions were λ data sets near the boundary $\lambda \approx 0$. In one of them `phylolm` stopped at the boundary and the GP found a higher likelihood (by 0.13), and in the other the GP optimiser stopped early (lower by 0.005). These therefore reflect differences between optimisers rather than between models. The default bounds of `phylolm` on the early-burst rate excluded the true value, and once both programs used the same bounds the estimates agreed. `geiger`, `nlme` and `caper` gave the same results as the GP wherever they fit the same model (Table S1). On the four additional trees the programs agreed to within 2.8×10^{-6} in 373 of 375 fits, and in the remaining two the GP fit had the higher likelihood (by up to 0.11). Both were Pagel's λ fits near the boundary, as on the original tree.

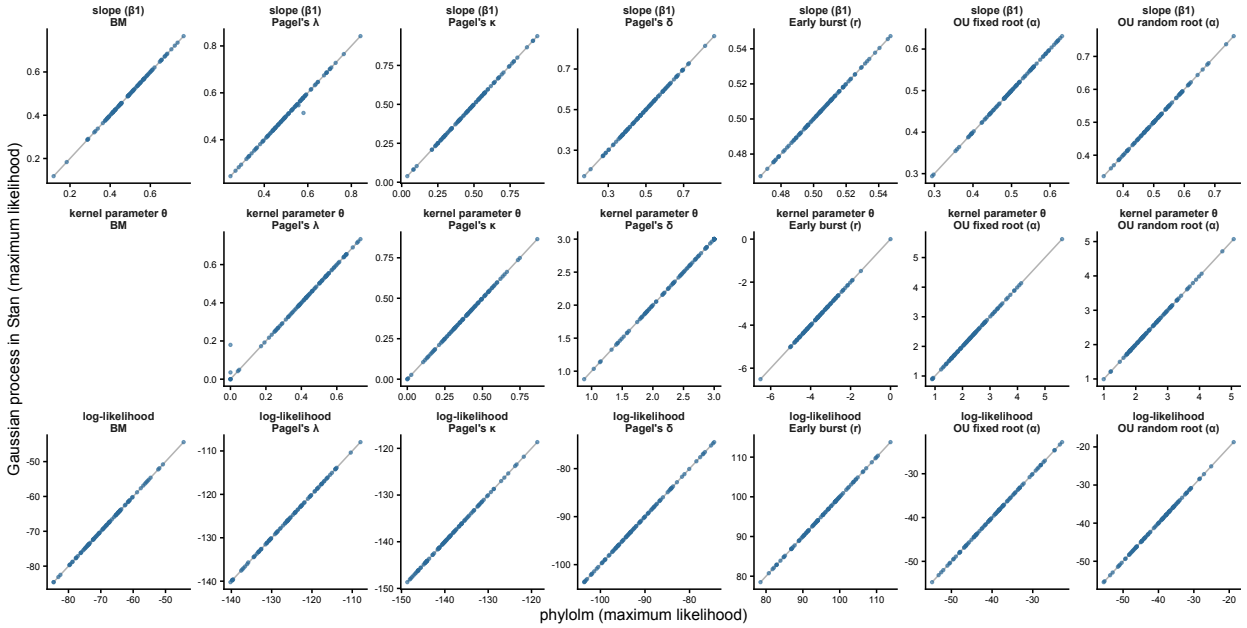


Figure 2. Maximum-likelihood estimates from `phylolm` (x-axis) and from the Gaussian-process implementation in Stan (y-axis) for 100 simulated data sets under each of seven classical models. Rows: regression slope, kernel parameter, maximised log-likelihood. BM has no kernel parameter, so that panel is empty. The off-diagonal points for Pagel's λ are the boundary cases described in the text.

GP prediction at internal nodes reproduced `fastAnc` (maximum difference 3.6×10^{-14}), and the variances matched once the root was integrated out under a flat prior. Across 1,000 replicates with 20 species held out, nominal 95%, 90% and 50% prediction intervals covered 94.5%, 89.6% and 49.9% of held-out values. The λ model and BM plus white noise gave identical maximised likelihoods, as did the Kronecker GP and `mvMORPH` bivariate BM, and the two-rate kernel and `mvMORPH` BMM. Latent-GP and `brms` posteriors for binary and count GLMMs agreed within Monte Carlo error (Appendix S4). On an ultrametric tree, a fixed-root OU fit and an ACDC fit with $r = 2\alpha$ gave identical log-likelihoods (-53.6873), whereas for the random-root OU the ML log-likelihoods differed (-54.530 vs. -54.211) and the REML log-likelihoods were identical (-55.2848), as predicted. These comparisons show that the implementations compute the same models, but they do not tell us how well those models recover evolutionary processes.

Kernel validity and identifiability

On the ultrametric tree, all four distance-based kernels were positive semi-definite at every length-scale, as predicted (the smallest eigenvalue was positive but close to zero). On the fossil and non-clock trees, the squared-exponential kernel produced eigenvalues as low as -1.1, and the Matérn-3/2 and -5/2 kernels also became indefinite at intermediate length-scales, whereas the exponential kernel remained valid throughout (smallest eigenvalue 0.00016; Fig. 3). On a fossil tree the squared-exponential kernel failed at 73% of 60 length-scales. Indefinite matrices are not valid covariance matrices, and in practice they lead to Cholesky failures or are silently 'fixed' by adding diagonal jitter, which in turn changes the model that is being fitted.

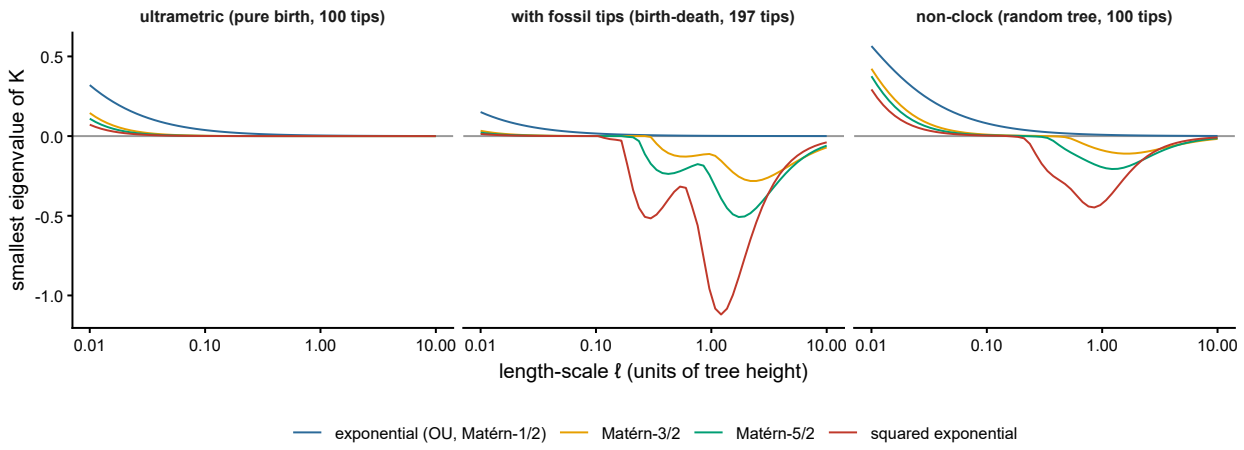


Figure 3. Smallest eigenvalue of kernel matrices $\phi(d_{ij}/\ell)$ computed from patristic distances. On an ultrametric tree every non-increasing $\phi \geq 0$ is valid. On trees with fossil tips or without a clock, values below zero mean the 'kernel' is not a valid covariance. Of the four kernels shown, only the exponential (Ornstein–Uhlenbeck) kernel, which belongs to the guaranteed-valid class of completely monotone functions, stays valid for all length-scales.

Without fossil tips, the maximised log-likelihoods of OU and ACDC differed by at most 7.8×10^{-9} in all 70 replicates, whichever process generated the data, as expected from the exact equivalence. With fossil tips, however, the difference was informative. $\log L_{\mathrm{OU}} - \log L_{\mathrm{ACDC}}$ had a median of 3.4 (90% interval 0.0 to 9.8) when OU generated the data, and -16.5 when ACDC did. A two-optimum OU model has the same kernel as the single-optimum model with a regime-dependent mean, and its

analytic mean and covariance matched 40,000 simulated tip vectors (Appendix S2).

Illustrative fits and replicated simulations of the extensions

Before the replicated study, I fitted each extension to one simulated data set with a strong, known effect (Fig. 4). The relaxed-rate GP recovered the fast clade without being told where it was (median posterior rate 6.8 times that elsewhere; truth 8; $\Delta\text{elpd} = 38.5 \pm 11.6$ over BM). The rate-through-time GP recovered a mid-history pulse that no monotone early-burst model can represent, and the additive kernel recovered the phylogenetic, geographic and residual shares of variance, which a phylogeny-only model could not (phylogenetic share 0.18 vs. true 0.25). The species slopes of the varying-slope kernel correlated with the truth ($r = 0.77$). However, a single data set cannot say anything about error rates, which are addressed by the replicated study described below.

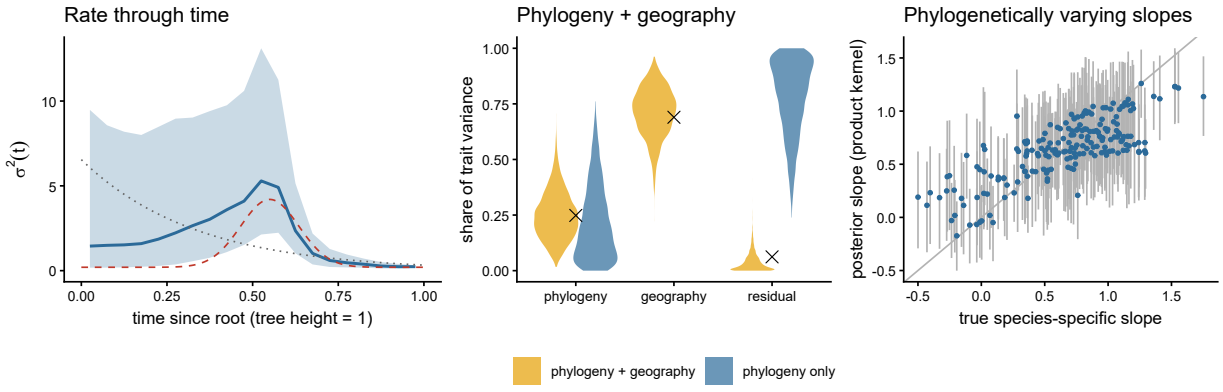


Figure 4. Kernel extensions recover known processes. (a) Rate-through-time GP: posterior median and 90% band of $\sigma^2(t)$ (blue), truth (red dashed) and the ML early-burst fit (grey dotted). (b) Share of trait variance attributed to phylogeny, geography and residual by a phylogeny-only model and by the additive phylogeny + geography kernel (crosses: truth). (c) Species-specific allometric slopes from the product (varying-slope) kernel versus truth (bars: 90% intervals).

Table 2 summarises the replicated study. Under the null scenarios, the extended kernels were preferred over their simpler alternatives in 0% of replicates for all four, and under the effect scenarios in 82% (relaxed rates), 74% (rate through time), 34% (varying slopes) and 87% (geography). Calibration was also uneven. Under the null scenarios, 90% intervals over-covered (98% for the

relaxed-rate ratio, 100% per bin for rate through time and 98% for species slopes). Under the discrete 8-fold shift, the relaxed-rate kernel shrank the clade-to-background ratio to a median of 4.5 and its intervals covered the true ratio in only 56% of replicates, because an autocorrelated-rate prior smooths an abrupt change. In other words, it locates shifts well (median branch rank correlation 0.69) but underestimates their size.

Table 2: Table 2. Replicated simulations of the kernel extensions. "Preferred" means an elpd more than two standard errors above the alternative (BM for relaxed rates and rate through time, constant-slope PGLS for varying slopes, phylogeny plus white noise for geography). Under nulls, "extension preferred" is the false-positive rate. Coverage and all other scenarios are in Appendix S4.

Kernel	Scenario	Reps	Extension preferred (%)	Simpler preferred (%)	Key estimate (median)
Relaxed rates	8× clade shift	50	82	0	ratio 4.5 (true 8)
Relaxed rates	constant BM	50	0	28	ratio 1.00
Rate through time	mid-history pulse	50	74	0	peak within 2 bins 86%
Rate through time	constant BM	50	0	24	RMSE log rate 0.15
Varying slopes	$\sigma_v = 0.6$	50	34	0	σ_v 0.56 (true 0.6)
Varying slopes	$\sigma_v = 0$	50	0	12	σ_v 0.12
Phylogeny + geography	$\sigma_g = 1$	30	87	0	geography share 0.72 (true 0.69)
Phylogeny + geography	$\sigma_g = 0$	30	0	40	geography share 0.03

The largest $\hat{R}$ across all fits was 1.048. However, this maximum was taken over hundreds of latent node values from only 2,000 draws. To check, I reran 40 replicates with 4 × 1,000 draws, which gave $\hat{R} \leq 1.009$ for every hyperparameter and latent value, so the maxima reflect short chains for a few latent values. The largest Pareto-$\hat{k}$ was 1.33, and 36% of the model comparisons had a $\hat{k}$ above 0.7. To check whether this matters, I recomputed the elpd exactly for 20 effect and 20 null relaxed-rate replicates. The model-selection decision changed in 1 of 40 replicates (detection 75% with PSIS vs. 70% exact, and 0% false positives in both). For the rate-through-time kernel, the original prior produced divergent transitions in most replicates, and a tighter prior removed them, at the cost of lower per-bin coverage (0.85 vs. 0.91).

With weaker effects, power fell sharply. The relaxed-rate kernel detected a 3-fold shift in 30% of replicates, the rate-through-time kernel a pulse of amplitude 1 in 10%, the varying-slope kernel $\sigma_v = 0.3$ in 10%, and the geographic kernel $\sigma_g = 0.5$ in 60%. On 300-tip trees the relaxed-rate kernel detected the 8-fold shift in every replicate but still shrank the ratio (median 4.4). Under misspecification the relaxed-rate kernel was conservative. On OU data it was never preferred over BM, and with three large jumps it was preferred in 13% of replicates. On random 100-species subtrees of the mammal tree, the results were close to those on pure-birth trees. The relaxed-rate kernel detected the shift in 77% of effect replicates and was preferred in 1 of 30 null replicates, whereas the varying-slope kernel did so in 20% and 0 of 30, respectively. Against `multirateBM`, which finished in 20 of 20 replicates, the kernel recovered branch rates equally well (median Spearman correlation 0.69 for both) and shrank the shift similarly (2.7 vs. 2.8; true 7.8), In addition, a four-chain HMC fit took a median of 8 s against 480 s for `multirateBM`.

Computational scaling

Pruning time grew with exponent 1.08 for the log-likelihood and 0.97 for all N LOO densities, and dense time with exponents 2.95 and 2.75. At 2,000 tips one pruning likelihood took 3 ms against 750 ms for the dense likelihood, and all 2,000 LOO densities took 4 ms against 2.0 s. Pruning needed 25 ms at 10,000 tips, where the dense matrix was not attempted. These timings are for unoptimised

R code. In Stan, one evaluation of the log density and its gradient took 2.9 ms at 5,000 tips with pruning (exponent 1.36 between 1,000 and 5,000 tips), against 923 ms for the dense kernel at 1,000 tips. Full HMC for the relaxed-rate kernel (which has $2N - 2$ latent log-rate increments) took 3, 11 and 40 s at 100, 300 and 1,000 tips without divergent transitions. These numbers support linear cost per gradient, but not necessarily for whole analyses of very large trees.

Mammalian brain size

Of 1,552 species with brain and body masses, 1,349 matched the phylogeny. For all seven classical kernels, the GP and `phylolm` log-likelihoods agreed to within 9.5×10^{-7}. Pagel's λ had the lowest AIC ($\lambda = 0.956$), and improved the fit by far more than any change in the evolutionary process among the other kernels. The fixed-root OU estimate ($\alpha = 2.88$) and an ACDC model with $r = 5.76$ gave the same maximised log-likelihood (843.856; BM 809.417). This means that, in these data, stabilising selection towards one optimum and rates of brain-size evolution increasing towards the present cannot be distinguished.

All 13 Bayesian fits converged (largest $\widehat{R}$ 1.004, no divergent transitions), and every species with $\widehat{k} > 0.7$ (0–19 per model) was refitted exactly. Our results highlight three main patterns (Fig. 5a). First, the residual model was the most influential component, as Student-t residuals improved on Gaussian residuals by 147.5 ± 55.6 elpd for the BM kernel, whereas residual variance growing with body size added much less (31.2 ± 36.4) and nothing once slopes varied. Second, the brain–body relationship was curved, and the quadratic coefficient was -0.019 (90% credible interval -0.023 to -0.014), which is concave on the log scale, as reported by Venditti et al. (2024) (Fig. 5b). Third, once heavy tails and curvature were modelled, varying slopes (23.1 ± 10.4) and relaxed rates (38.1 ± 19.6) still predicted left-out species better than the curved BM + t model, by about two standard errors, but could not be separated from each other (14.9 ± 18.1). The slope variability ($\sigma_v = 0.076$) was smaller than without the controls (0.107). A comparison with BM and Gaussian residuals alone would have credited varying slopes with 79.7 ± 39.8 elpd, most of which came from heavy tails and curvature.

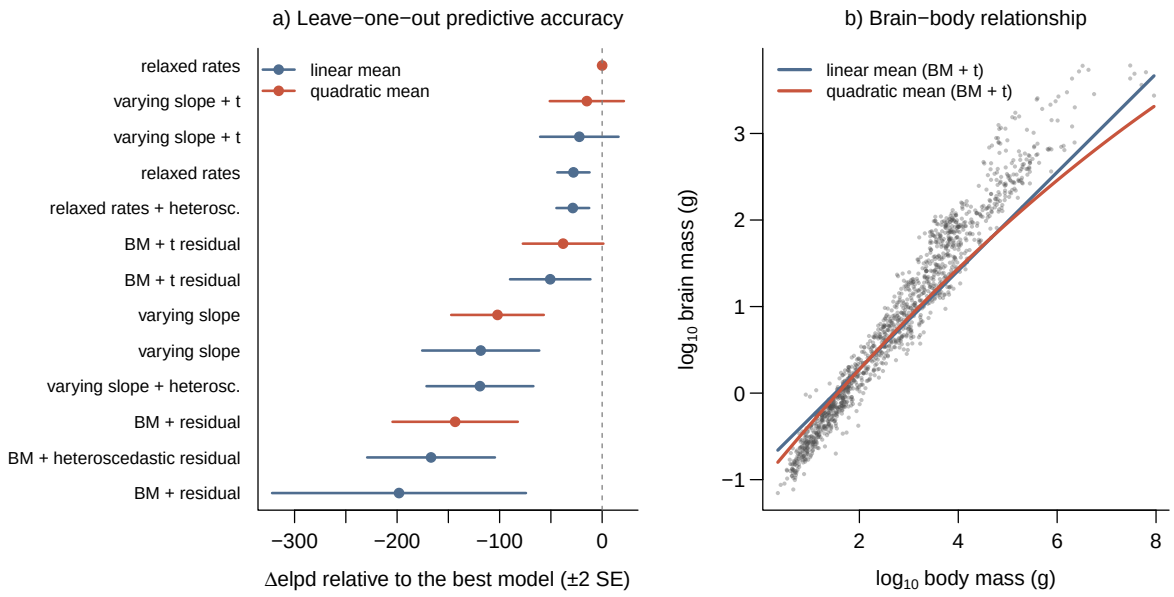


Figure 5. Mammalian brain size: kernels compared with simpler alternative explanations. (a) Expected log predictive density relative to the best model (exact leave-one-out: every species with Pareto-$\hat{k} > 0.7$ refitted; bars $\pm 2SE$) $formodelscrossingtheprocesskernel(BM, phylogeneticallyvaryingslope, relaxedrates), the$ and the mean (blue, linear; red, quadratic in log body mass). (b) Brain–body relationship with the linear and quadratic means of the BM + t model.

Leaving out whole orders, however, changed the results obtained for varying slopes. The curved varying-slope model predicted left-out orders worse than the curved BM + t model (-41.3 ± 25.5 summed over 28 orders), as expected for a slope that evolves locally and so can only be extrapolated, not estimated, for an unseen clade. Relaxed rates predicted left-out orders best (67.0 ± 58.9), but with only 28 groups the uncertainty was large, and most of the gain came from Cetacea ($+49.0$) and Chiroptera ($+27.5$). Across 50 posterior trees the ranking of the four key models was the same as on the maximum clade credibility tree, and estimates varied little (slope variability 0.069–0.087; quadratic coefficient -0.020 to -0.018). Adding a geographic kernel barely changed the fit. The ML estimate attributed 0.3% of the variance in relative brain size to geography against 95% to phylogeny, and the leave-one-out gain was less than two standard errors (5.5 ± 2.9). When marine mammals were removed (74 species), the share rose to 0.7% and the gain to 11.0 ± 4.5. Therefore, any geographic structure seems to be weak at this scale, and whether it can be detected depends on

the species that are included.

Discussion

As shown here, standard continuous-trait PCMs are Gaussian processes on a phylogeny (or on its tips) that differ only in their kernels and mean functions. A systematic treatment therefore provides a single notation, explicit conditions under which each representation holds, as well as a route to new models. A generic GP implementation reproduced the estimates from `phylolm`, `geiger`, `nlme`, `caper`, `mvMORPH`, `phytools` and `brms` to numerical precision (or within Monte Carlo error) across univariate, multivariate, multi-rate, non-Gaussian and ancestral-state problems. The few disagreements were caused by optimiser settings or default bounds in the specialised software, rather than by differences between models. It is important to note, however, that this is evidence of implementation equivalence only. In other words, it shows that the GP code computes the same likelihoods as established software and, therefore, that it inherits the known statistical properties (and pathologies) of those models.

I consider that the GP view is useful for three reasons. First, it clarifies what models assume and how they relate to one another. For instance, on an ultrametric tree Pagel's λ is BM plus white noise, and ancestral-state reconstruction under a process defined on the whole tree is kriging at internal points. For OU, this kriging has to use the OU covariance between node and tip states and not the ultrametric edge variances of Table 1, because the latter only reproduce the tip distribution and would give ACDC reconstructions instead. This equivalence has been noted before (Ho & Ané, 2014b; Murrell, 2018; Uyeda et al., 2015), but the edge form turns it into a one-line consequence of Eq. (1) and pins down when it holds. Under those conditions, extant species alone cannot distinguish "stabilising selection towards an optimum" from "rates of evolution increasing towards the present", regardless of how much data are available. The distinction can only come from information beyond the extant tip distribution, such as non-ultrametric trees (i.e. fossils), multiple optima with independent information about where the shifts occurred, or priors.

Second, the kernel framework is generative, as new evolutionary hypotheses become new kernels, which are built from a small set of operations (i.e. sums, products and hierarchical priors on kernel parameters) with well-understood validity conditions. However, my simulations only characterise the four extensions within the scenarios I simulated, namely pure-birth trees of 80–300 tips, random subtrees of the mammal tree, two effect sizes per kernel, and a few misspecified processes for the relaxed-rate kernel. Within these scenarios, the power to detect weak effects was low, false positives were rare, and the relaxed-rate kernel was conservative under the misspecifications that I tried. How the kernels behave on other tree shapes and under other kinds of misspecification (and, for the other three kernels, under misspecification at all) still needs to be studied before they are used routinely. LOO preferences, in particular, are statements about predictive performance in a given data set, not proof that a process actually operated. In addition, they rest on importance-sampling approximations whose reliability varies. The largest Pareto-$\hat{k}$ exceeded 0.7 in many replicates, so the rates in Table 2 should be regarded as approximate, although exact refitting left the relaxed-rate conclusions unchanged. Finally, the two-standard-error rule never selected an extension under a null scenario, but it selected the true, simpler model in only 12–40% of null replicates. It therefore behaves as a conservative test of the extension rather than a symmetric model-selection rule, and standard errors of elpd differences are themselves unreliable for small samples (Sivula et al., 2020).

The varying-slope kernel provides a model-based treatment of the long-standing "taxon-level problem" in allometry (Lande, 1979; M. D. Pagel & Harvey, 1989). Instead of fitting separate regressions within arbitrarily chosen taxonomic ranks, it allows the slope itself to evolve along the tree. Phylogenetic random-slope models with the same marginal covariance can also be fitted in mixed-model software (Hadfield & Nakagawa, 2010; Matsuba et al., 2024), and the kernel form makes this connection explicit. The relaxed-rate kernel, on the other hand, is a Bayesian alternative to reversible-jump approaches (Eastman et al., 2011; Venditti et al., 2011), to penalised-likelihood rates (Revell, 2021) and to relaxed random walks (Gill et al., 2017). Its closest precedent is Fisher et al. (2021), and the main differences are the autocorrelated prior on log rates and that it is one component that can be combined with others in the same model. It recovered branch rates as well

as `multirateBM` while running much faster, but a comparison with BEAST (Suchard et al., 2018) and RevBayes (Höhna et al., 2016) is still needed. Other extensions follow the same recipe, such as Lévy-type jumps (Landis et al., 2013) as heavy-tailed increments, OU shift models (Bastide et al., 2018; Butler & King, 2004; Uyeda & Harmon, 2014) as regime-specific means (as I verified for models with shared α and σ^2 in Appendix S2), and phylogenetic uncertainty (Villemereuil et al., 2012) by marginalising over trees.

Third, the GP framework comes with general tools for inference and model checking. Leave-one-out cross-validation (Bürkner et al., 2021; Sundararajan & Keerthi, 2001) compares models with different numbers and kinds of parameters based on their predictive performance, without relying on the asymptotic justification of AIC, which is unreliable near parameter boundaries. For Gaussian models the conditional LOO densities are exact and, for tree kernels, they can be computed in linear time. The Pareto-$\hat{k}$ diagnostics should always be checked, and exact refits for influential species are cheap enough to be routine. The mammal example illustrates why these checks matter. When compared with BM and Gaussian residuals, both varying slopes and relaxed rates appeared to improve prediction considerably, but most of that came from heavy-tailed residuals and curvature, and leaving out whole orders reversed the advantage of varying slopes. Therefore, kernel extensions should be compared with simpler explanations that leave the same statistical signature, and assessed with more than one form of cross-validation, before their predictive gains are interpreted as evidence for a particular process. Whilst this example does not provide definitive answers about brain evolution (and these results are specific to this data set and set of models), it does show how the workflow can be used to question an apparently strong result.

The GP view is complementary to existing general frameworks for Gaussian models on trees (Bastide et al., 2021; Manceau et al., 2017; Mitov et al., 2020). These are organised around the generating process (i.e. how a trait changes along a branch), which gives fast likelihoods for multivariate processes with shifts. By contrast, the GP view is organised around the covariance function. This makes sums, products, hierarchical priors and validity conditions the natural operations, and it

connects the models to generic GP software and diagnostics. For the models in Table 1 both descriptions coincide. What has been missing until now is a framework organised around the covariance function, in which (i) kernels can be combined freely and fitted with the same code; (ii) a kernel can be checked for validity on the tree at hand, and two kernels can be checked for whether they imply the same tip distribution; and (iii) leave-one-out and leave-clade-out predictive densities, and predictions at tips and internal nodes, are computed in linear time for any branch-wise kernel. The `phylogp` package provides these. Specialised packages remain the better choice for models whose likelihood is not a GP on the tree (e.g. discrete-state shift models) and for fast, well-tested ML fits of the classical models. My claim is that these models can be regarded as special cases of a common framework, rather than that existing software should be replaced.

Kernel validity deserves particular attention, as off-the-shelf GP software often defaults to squared-exponential kernels and it is tempting to simply plug patristic distances into them. On ultrametric trees this is harmless, because any decreasing function of patristic distance is then just another tree transformation. However, as soon as the tree includes fossil tips, or is a phylogram without a clock, such matrices can fail to be positive semi-definite, whereas the exponential kernel (which is exactly OU) and other completely monotone functions remain valid. This is particularly relevant for total-evidence analyses, which are precisely the setting in which GPs on trees are most attractive. In these analyses, kernels should be built from Eq. (1) or from completely monotone functions of patristic distance, and adding "jitter" to the diagonal should not be used to rescue an invalid kernel.

In my view, two limitations matter more than computation. The first one is identifiability. As the OU and ACDC case shows, different kernels can imply the same distribution for extant species, and many more will probably be only weakly distinguishable in practice. The second one is model adequacy, since a better predictive score does not show that the process encoded by a kernel actually operated. Computation, on the other hand, is less of a constraint. Dense GP likelihoods cost $O(N^3)$, but for kernels of the form (1) the $O(N)$ pruning recursion avoids this (Felsenstein, 1973; Ho & Ané, 2014a). Kernels that break the Markov structure of the tree, such as spatial kernels, require

dense algebra or approximations (Lindgren et al., 2011; Matsuba et al., 2024; Rue et al., 2009).

In summary, treating PCMs as Gaussian processes on trees brings a diverse set of methods under a single perspective and helps to clarify long-standing problems, such as the identifiability of OU models. It also provides an extensible and computationally practical route to new models of trait evolution, whose statistical behaviour can be assessed with the same general tools. Future work would benefit from testing these kernels on a wider range of trees and empirical data sets, and from comparing them directly with the Bayesian relaxed-rate models now available in BEAST and RevBayes. Rates that depend on a discrete character mapped on the tree are already available in the package (version 0.3.0; Appendix S5), and one interesting next step would be to bring multivariate and non-Gaussian models into the package, as well as to use the same approach to build kernels for trait evolution along phylogenetic networks.

Conflict of interest

I declare no conflicts of interest.

Funding

This research received no specific funding.

Acknowledgements

I thank the Comparative Evolutionary Anthropology group at the University of Oxford for constructive discussion.

Use of AI tools

I conceived the study, developed the theoretical framework and the derivations, wrote the original Stan programs and analysis code, and wrote the first draft of the manuscript. I then used Claude (Anthropic), through Claude Code, as a programming and writing assistant. It helped to develop the R package `phylogp` from my Stan programs, to write and run the additional analyses carried out during revision, and to edit and revise the text of the manuscript and the appendices. I checked the code, results, derivations and text, and I take full responsibility for the content of the paper.

Data availability

The methods are implemented in the R package `phylogp` (https://github.com/ThomasPueschel/phylogp; version 0.3.0 archived at Zenodo, doi:10.5281/zenodo.23889024). The mammal reanalyses and additional simulations used development builds of the same repository (0.2.0.9000), whose features are included in this release (Appendix S5), and the earlier version 0.2.0 is archived separately (doi:10.5281/zenodo.23472027). All analysis scripts, simulated data and results underlying every figure and table are archived at Zenodo (doi:10.5281/zenodo.23472071). The empirical data come from published sources: brain and body masses from Burger et al. (2019) (Dryad, doi:10.5061/dryad.2r62k7s), and the maximum clade credibility tree and the posterior sample of trees from Upham et al. (2019) (https://github.com/n8upham/MamPhy_v1).

Supplementary material

Appendix S1 (proofs), Appendix S2 (multi-optimum Ornstein–Uhlenbeck models), Appendix S3 (comparison with R packages, Table S1; sampler diagnostics and prior sensitivity; mammal sensitivity analyses), Appendix S4 (full methods and results), Appendix S5 (development of the package since version 0.2.0) and Figures S1–S5 are provided as online supplementary material.

References

Bastide, P., Ané, C., Robin, S., & Mariadassou, M. (2018). Inference of adaptive shifts for multivariate correlated traits. *Systematic Biology*, *67*, 662–680.

Bastide, P., Ho, L. S. T., Baele, G., Lemey, P., & Suchard, M. A. (2021). Efficient Bayesian inference of general Gaussian models on large phylogenetic trees. *The Annals of Applied Statistics*, *15*, 971–997. https://doi.org/10.1214/20-AOAS1419

Blomberg, S. P., Garland, T., & Ives, A. R. (2003). Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. *Evolution*, *57*, 717–745.

Burger, J. R., George, M. A., Leadbetter, C., & Shaikh, F. (2019). The allometry of brain size in mammals. *Journal of Mammalogy*, *100*, 276–283.

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Bürkner, P.-C., Gabry, J., & Vehtari, A. (2021). Efficient leave-one-out cross-validation for Bayesian non-factorized normal and Student-t models. *Computational Statistics*, *36*, 1243–1261.

Butler, M. A., & King, A. A. (2004). Phylogenetic comparative analysis: A modeling approach for adaptive evolution. *The American Naturalist*, *164*, 683–695.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32.

Clavel, J., Escarguel, G., & Merceron, G. (2015). mvMORPH: An R package for fitting multivariate evolutionary models to morphometric data. *Methods in Ecology and Evolution*, *6*, 1311–1319.

Clavel, J., & Morlon, H. (2017). Accelerated body size evolution during cold climatic periods in the Cenozoic. *Proceedings of the National Academy of Sciences of the USA*, *114*, 4183–4188. https://doi.org/10.1073/pnas.1606868114

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. *Biological Journal of the Linnean Society*, *118*, 64–77.

Cressie, N. A. C. (1993). *Statistics for spatial data* (Revised). Wiley.

Dellacherie, C., Martínez, S., & San Martín, J. (2014). *Inverse M-matrices and ultrametric matrices* (Vol. 2118). Springer. https://doi.org/10.1007/978-3-319-10298-6

Deza, M. M., & Laurent, M. (1997). *Geometry of cuts and metrics*. Springer.

Diggle, P. J., & Ribeiro, P. J., Jr. (2007). *Model-based geostatistics*. Springer.

Duvenaud, D., Lloyd, J., Grosse, R., Tenenbaum, J., & Ghahramani, Z. (2013). Structure discovery in nonparametric regression through compositional kernel search. *Proceedings of the 30th International Conference on Machine Learning, PMLR*, *28*, 1166–1174.

Eastman, J. M., Alfaro, M. E., Joyce, P., Hipp, A. L., & Harmon, L. J. (2011). A novel comparative method for identifying shifts in the rate of character evolution on trees. *Evolution*, *65*, 3578–3589.

Felsenstein, J. (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics*, *25*, 471–492.

Felsenstein, J. (1985). Phylogenies and the comparative method. *The American Naturalist*, *125*, 1–15.

Felsenstein, J. (2005). Using the quantitative genetic threshold model for inferences between and within species. *Philosophical Transactions of the Royal Society B*, *360*, 1427–1434.

Feragen, A., Lauze, F., & Hauberg, S. (2015). Geodesic exponential kernels: When curvature and linearity conflict. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3032–3042.

Fisher, A. A., Ji, X., Zhang, Z., Lemey, P., & Suchard, M. A. (2021). Relaxed random walks at scale. *Systematic Biology*, *70*(2), 258–267. https://doi.org/10.1093/sysbio/syaa056

Freckleton, R. P. (2012). Fast likelihood calculations for comparative analyses. *Methods in Ecology and Evolution*, *3*, 940–947.

Freckleton, R. P., & Jetz, W. (2009). Space versus phylogeny: Disentangling phylogenetic and spatial signals in comparative data. *Proceedings of the Royal Society B*, *276*, 21–30.

Gabry, J., Češnovar, R., Johnson, A., & Bronder, S. (2024). *Cmdstanr: R interface to CmdStan*.

Garland, T., Jr., & Ives, A. R. (2000). Using the past to predict the present: Confidence intervals

for regression equations in phylogenetic comparative methods. *The American Naturalist*, *155*, 346–364. https://doi.org/10.1086/303327

Gill, M. S., Ho, L. S. T., Baele, G., Lemey, P., & Suchard, M. A. (2017). A relaxed directional random walk model for phylogenetic trait evolution. *Systematic Biology*, *66*, 299–319. https://doi.org/10.1093/sysbio/syw093

Gneiting, T. (2013). Strictly and non-strictly positive definite functions on spheres. *Bernoulli*, *19*, 1327–1349. https://doi.org/10.3150/12-BEJSP06

Grafen, A. (1989). The phylogenetic regression. *Philosophical Transactions of the Royal Society of London B*, *326*, 119–157.

Guzmán Naranjo, M., & Becker, L. (2022). Statistical bias control in typology. *Linguistic Typology*, *26*, 605–670. https://doi.org/10.1515/lingty-2021-0002

Hadfield, J. D., & Nakagawa, S. (2010). General quantitative genetic methods for comparative biology: Phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology*, *23*, 494–508.

Hadjipantelis, P. Z., Jones, N. S., Moriarty, J., Springate, D. A., & Knight, C. G. (2013). Function-valued traits in evolution. *Journal of the Royal Society Interface*, *10*, 20121032.

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

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

Harmon, L. J., Losos, J. B., Davies, T. J., Gillespie, R. G., Gittleman, J. L., Jennings, W. B., et al. (2010). Early bursts of body size and shape evolution are rare in comparative data. *Evolution*, *64*, 2385–2396.

Hartmann, F., & Jäger, G. (2023). Gaussian process models for geographic controls in phylogenetic trees. *Open Research Europe*, *3*, 57. https://doi.org/10.12688/openreseurope.15490.2

Ho, L. si T., & Ané, C. (2014a). A linear-time algorithm for Gaussian and non-Gaussian trait

588 evolution models. *Systematic Biology*, *63*, 397–408.
589 Ho, L. si T., & Ané, C. (2014b). Intrinsic inference difficulties for trait evolution with Ornstein-
590 Uhlenbeck models. *Methods in Ecology and Evolution*, *5*, 1133–1146.
591 Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck,
592 J. P., & Ronquist, F. (2016). RevBayes: Bayesian phylogenetic inference using graphical
593 models and an interactive model-specification language. *Systematic Biology*, *65*(4), 726–736.
594 https://doi.org/10.1093/sysbio/syw021
595 Housworth, E. A., Martins, E. P., & Lynch, M. (2004). The phylogenetic mixed model. *The*
596 *American Naturalist*, *163*, 84–96.
597 Ives, A. R., & Helmus, M. R. (2011). Generalized linear mixed models for phylogenetic analyses of
598 community structure. *Ecological Monographs*, *81*, 511–525.
599 Jones, K. E., Bielby, J., Cardillo, M., Fritz, S. A., O'Dell, J., Orme, C. D. L., Safi, K., Sechrest,
600 W., Boakes, E. H., Carbone, C., Connolly, C., Cutts, M. J., Foster, J. K., Grenyer, R., Habib,
601 M., Plaster, C. A., Price, S. A., Rigby, E. A., Rist, J., … Purvis, A. (2009). PanTHERIA: A
602 species-level database of life history, ecology, and geography of extant and recently extinct
603 mammals. *Ecology*, *90*, 2648. https://doi.org/10.1890/08-1494.1
604 Jones, N. S., & Moriarty, J. (2013). Evolutionary inference for function-valued traits: Gaussian
605 process regression on phylogenies. *Journal of the Royal Society Interface*, *10*, 20120616.
606 Lande, R. (1979). Quantitative genetic analysis of multivariate evolution, applied to brain:body size
607 allometry. *Evolution*, *33*, 402–416.
608 Landis, M. J., Schraiber, J. G., & Liang, M. (2013). Phylogenetic analysis using Lévy processes:
609 Finding jumps in the evolution of continuous traits. *Systematic Biology*, *62*, 193–204.
610 Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and
611 Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal*
612 *of the Royal Statistical Society B*, *73*, 423–498.
613 Lynch, M. (1991). Methods for the analysis of comparative data in evolutionary biology. *Evolution*,
614 *45*, 1065–1080.

615 Manceau, M., Lambert, A., & Morlon, H. (2017). A unifying comparative phylogenetic framework
616 including traits coevolving across interacting lineages. *Systematic Biology*, *66*(4), 551–568.
617 https://doi.org/10.1093/sysbio/syw115
618 Martins, E. P., & Hansen, T. F. (1997). Phylogenies and the comparative method: A general
619 approach to incorporating phylogenetic information into the analysis of interspecific data. *The*
620 *American Naturalist*, *149*, 646–667.
621 Matsuba, M., Fukasawa, K., Aoki, S., Akasaka, M., & Ishihama, F. (2024). Scalable phylogenetic
622 Gaussian process models improve the detectability of environmental signals on local extinctions
623 for many Red List species. *Methods in Ecology and Evolution*, *15*(4), 756–768. https://doi.org/
624 10.1111/2041-210X.14291
625 McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd
626 ed.). CRC Press.
627 Meagher, J. P., Damoulas, T., Jones, K. E., & Girolami, M. (2018). Phylogenetic Gaussian processes
628 for bat echolocation. In *Statistical data science* (pp. 111–124). World Scientific. https:
629 //doi.org/10.1142/9781786345400_0007
630 Mitov, V., Bartoszek, K., Asimomitis, G., & Stadler, T. (2020). Fast likelihood calculation for
631 multivariate Gaussian phylogenetic models with shifts. *Theoretical Population Biology*, *131*,
632 66–78.
633 Murrell, D. J. (2018). A global envelope test to detect non-random bursts of trait evolution. *Methods*
634 *in Ecology and Evolution*, *9*(7), 1739–1748. https://doi.org/10.1111/2041-210X.13006
635 O'Meara, B. C., Ané, C., Sanderson, M. J., & Wainwright, P. C. (2006). Testing for different rates
636 of continuous trait evolution using likelihood. *Evolution*, *60*, 922–933.
637 Orme, D., Freckleton, R., Thomas, G., Petzoldt, T., Fritz, S., Isaac, N., & Pearse, W. (2023). *Caper:*
638 *Comparative analyses of phylogenetics and evolution in R*.
639 Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, *401*, 877–884.
640 Pagel, M. D., & Harvey, P. H. (1989). Taxonomic differences in the scaling of brain on body weight
641 among mammals. *Science*, *244*, 1589–1593.

Paradis, E., & Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*, 526–528.

Pennell, M. W., Eastman, J. M., Slater, G. J., Brown, J. W., Uyeda, J. C., FitzJohn, R. G., Alfaro, M. E., & Harmon, L. J. (2014). Geiger v2.0: An expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics*, *30*, 2216–2218.

Pinheiro, J., Bates, D., & R Core Team. (2023). *Nlme: Linear and nonlinear mixed effects models*.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

Revell, L. J. (2012). Phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, *3*, 217–223.

Revell, L. J. (2021). A variable-rate quantitative trait evolution model using penalized-likelihood. *PeerJ*, *9*, e11997.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society B*, *71*, 319–392.

Schluter, D., Price, T., Mooers, A. Ø., & Ludwig, D. (1997). Likelihood of ancestor states in adaptive radiation. *Evolution*, *51*, 1699–1711. https://doi.org/10.1111/j.1558-5646.1997.tb05095.x

Schoenberg, I. J. (1938). Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, *44*, 522–536.

Sivula, T., Magnusson, M., Matamoros, A. A., & Vehtari, A. (2020). *Uncertainty in Bayesian leave-one-out cross-validation based model comparison*. arXiv preprint arXiv:2008.10296 (version 6, 2025).

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., & Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, *4*(1), vey016. https://doi.org/10.1093/ve/vey016

Sundararajan, S., & Keerthi, S. S. (2001). Predictive approaches for choosing hyperparameters in Gaussian processes. *Neural Computation*, *13*, 1103–1118.

669 Thorne, J. L., & Kishino, H. (2002). Divergence time and evolutionary rate estimation with
670 multilocus data. *Systematic Biology*, *51*, 689–702.
671 Upham, N. S., Esselstyn, J. A., & Jetz, W. (2019). Inferring the mammal tree: Species-level sets of
672 phylogenies for questions in ecology, evolution, and conservation. *PLoS Biology*, *17*, e3000494.
673 Uyeda, J. C., Caetano, D. S., & Pennell, M. W. (2015). Comparative analysis of principal components
674 can be misleading. *Systematic Biology*, *64*(4), 677–689. https://doi.org/10.1093/sysbio/syv019
675 Uyeda, J. C., & Harmon, L. J. (2014). A novel Bayesian method for inferring and interpreting the
676 dynamics of adaptive landscapes from phylogenetic comparative data. *Systematic Biology*, *63*,
677 902–918.
678 Uyeda, J. C., Zenil-Ferguson, R., & Pennell, M. W. (2018). Rethinking phylogenetic comparative
679 methods. *Systematic Biology*, *67*, 1091–1109.
680 Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-
681 out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432.
682 Venditti, C., Baker, J., & Barton, R. A. (2024). Co-evolutionary dynamics of mammalian brain and
683 body size. *Nature Ecology & Evolution*, *8*, 1534–1542. https://doi.org/10.1038/s41559-024-
684 02451-3
685 Venditti, C., Meade, A., & Pagel, M. (2011). Multiple routes to mammalian diversity. *Nature*, *479*,
686 393–396.
687 Villemereuil, P. de, Wells, J. A., Edwards, R. D., & Blomberg, S. P. (2012). Bayesian models for
688 comparative analysis integrating phylogenetic uncertainty. *BMC Evolutionary Biology*, *12*, 102.