

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

4 Thomas A. Püschel

5 *Institute of Human Sciences, School of Anthropology and Museum Ethnography, University*
6 *of Oxford, Oxford, United Kingdom*

7 Correspondence: Thomas A. Püschel, thomas.puschel@anthro.ox.ac.uk

Abstract

Phylogenetic comparative methods for continuous traits, such as phylogenetic generalised least squares, Brownian motion and Ornstein–Uhlenbeck models, Pagel’s tree transformations, early-burst and multi-rate models, phylogenetic mixed models and ancestral-state reconstruction, 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 the present study, I use it to build, check and compare models of trait evolution. Most classical models share a single branch-wise form of the kernel, 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. A generic implementation in Stan reproduced standard software to numerical precision on 99.6% of simulated data sets, on ultrametric trees, trees with fossil tips and trees with polytomies. To check models, I collect the conditions under which kernels defined on patristic distances are valid. Any non-increasing function is valid 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 and Matérn kernels can fail. I also provide a numerical test of whether two kernels imply the same distribution for the tips, which recovers the known equivalence between single-optimum Ornstein–Uhlenbeck and accelerating-rate models on ultrametric trees. To build models, I write four extensions as kernels (i.e. autocorrelated rates, smooth rate-through-time functions, joint phylogenetic and spatial structure, and phylogenetically varying regression slopes), all of which have close precedents in the literature, and I assess their bias, interval coverage, power and false-positive rates 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, where kernel extensions were compared with simpler explanations (e.g. heavy-tailed residuals and a curved allometric relationship) and validated by leaving out whole orders. All methods are implemented in the R package *phylogp*. Overall, treating comparative methods as Gaussian processes offers a common and extensible basis both for developing new models of trait

35 evolution and for testing them.

36 **Keywords:** Gaussian process; phylogenetic comparative methods; kernel; Brownian
37 motion; Ornstein–Uhlenbeck; brain size

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 non-independence 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 and Hansen (1997)], Brownian motion (BM) and Ornstein–Uhlenbeck (OU) models (Felsenstein 1973; Hansen 1997; Butler and King 2004), Pagel’s λ , κ and δ transformations (Pagel 1999), early-burst and accelerating–decelerating (ACDC) models (Blomberg et al. 2003; Harmon et al. 2010), multi-rate BM (O’Meara et al. 2006; Eastman et al. 2011; Venditti et al. 2011), phylogenetic mixed models (Lynch 1991; Housworth et al. 2004; Hadfield and Nakagawa 2010; Ives and Helmus 2011) and ancestral-state reconstruction (Schluter et al. 1997; Garland and Ives 2000); the stochastic models that underlie them have recently been reviewed by Jhwueng (2025). Each of these methods comes with its own software (e.g. Paradis and Schliep 2019; Revell 2012; Pennell et al. 2014; Ho and Ané 2014a; Clavel et al. 2015), parameterisation and folklore. Consequently, it is easy to lose sight of how closely related these models actually are, and of which of their assumptions really matter (Uyeda et al. 2018), even though many of the connections between them are known (e.g. Blomberg et al. 2012).

In statistics and machine learning, a Gaussian process (GP) is a probability distribution over functions, which is fully specified by a mean function and a covariance function, or *kernel* (Rasmussen and Williams 2006). GPs underlie kriging in geostatistics (Cressie 1993; Diggle and Ribeiro 2007), Gaussian random fields in spatial statistics and ecology (Lindgren et al. 2011), and a large share of modern nonparametric regression (Rasmussen and Williams 2006). Their main appeal is their modularity. Once the data are modelled as $\mathbf{y} \sim \mathcal{N}(\mathbf{m}, \mathbf{K})$, the scientific assumptions sit entirely in the kernel $\mathbf{K}$. In addition, kernels can be added, multiplied and composed (Rasmussen and Williams 2006, ch. 4; Duvenaud et al. 2013), and inference, prediction and model comparison follow general rules that do not depend on the particular kernel being used.

The connection between PCMs and GPs is not new and has been mentioned before. The multivariate normal likelihood of BM on a tree is in fact the starting point of PGLS, and Hansen and Martins (1996) derived the interspecific correlation structures implied by a range of microevolutionary processes, which are kernels in all but name. General Gaussian models on trees with linear-time likelihoods have been formalised by Ho and Ané (2014a), as the GLInv family of Mitov et al. (2020), and for Bayesian inference by Bastide et al. (2021), whose linear-time gradients also make Bayesian relaxed random-walk models with branch-specific rates practical on trees with thousands of tips (Fisher et al. 2021). In addition, Manceau et al. (2017) proposed a unifying framework in which a wide family of models, including traits that coevolve across interacting lineages, yield Gaussian distributions of trait values at the tips. Likewise, predicting unobserved tips from a fitted PGLS is already a form of kriging (Garland and Ives 2000). Jones and Moriarty (2013) and Hadjipantelis et al. (2013) introduced “phylogenetic Gaussian process regression” for function-valued traits, using BM and OU covariance functions on the tree, an approach later applied to the evolution of bat echolocation calls (Meagher et al. 2018), while McElreath (2020) fitted primate brain size with an OU-type GP kernel on phylogenetic distance as a teaching example. In linguistics, OU-type phylogenetic kernels have been combined with geographic kernels in distance-based GP models (Guzmán Naranjo and Becker 2022; Hartmann and Jäger 2023), and in conservation biology phylogenetic GPs with random intercepts and slopes have been scaled to more than a thousand endangered plant taxa by nearest-neighbour approximation (Matsuba et al. 2024). More recently, mixed Gaussian phylogenetic models have been described as combinations of Gaussian sub-processes (BM, OU) on a tree (Brahmantio et al. 2026), and general-purpose tools for GPs on graphs are now available in Stan (Hoffmann 2023). Therefore, most of the pieces already exist, but they are scattered across different disciplines and are rarely described in the language of kernels.

The present study brings this material together and extends it. In particular, I (i) write the classical continuous-trait PCMs as GP 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 on ultrametric matrices and on metrics of negative type, which I apply rather than discover); and (iv) extend the toolbox with additional biologically motivated models, each with close precedents in the literature, 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). This is a matter of algebra and numerics, and I tested it by fitting identical models with different code. 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). Equivalence does not establish validity, since two implementations can agree perfectly and still behave badly. The statistical properties of the classical models have been documented elsewhere (e.g. Ho and Ané 2014b; Cooper et al. 2016); therefore, I only evaluated the kernel extensions, by replicated simulation including null scenarios in which the targeted process was absent.

I start with the basic BM model and then build up to multi-rate, multivariate and non-Gaussian models, in each case comparing the estimates obtained from standard R packages with those from a GP implementation in Stan (Carpenter et al. 2017). Subsequently, I introduce four kernel extensions, namely autocorrelated evolutionary rates, smooth rate-through-time functions, joint phylogenetic and spatial structure, and phylogenetically varying regression slopes, and test each of them by simulation. Finally, I apply all four to brain–body allometry across mammals (Burger et al. 2019a), using a time-calibrated molecular phylogeny (Upham et al. 2019), as an illustration of the workflow rather than as a test of hypotheses about brain evolution. All the methods are implemented in the R package `phylogp`, which writes the Stan program for any sum or

117 product of the kernels described here, and all the analysis code is archived (Data availability).

118 **Theory**

119 **Gaussian processes**

120 A Gaussian process $f \sim \mathcal{GP}(m, k)$ on an index set $\mathcal{X}$ is a collection of random variables
 121 $\{f(x) : x \in \mathcal{X}\}$, any finite subset of which is jointly Gaussian with mean $m(x)$ and covariance
 122 $\text{Cov}[f(x), f(x')] = k(x, x')$ (Rasmussen and Williams 2006). The kernel k must be positive
 123 semi-definite: for every finite set of points the matrix $K_{ij} = k(x_i, x_j)$ must have non-negative
 124 eigenvalues. Given observations $\mathbf{y} = f(\mathbf{x}) + \varepsilon$ with $\varepsilon \sim \mathcal{N}(\mathbf{0}, \Sigma_\varepsilon)$, the log marginal likelihood is

$$\log p(\mathbf{y} \mid \theta) = -\frac{1}{2}(\mathbf{y} - \mathbf{m})^\top \mathbf{V}^{-1}(\mathbf{y} - \mathbf{m}) - \frac{1}{2} \log |\mathbf{V}| - \frac{N}{2} \log 2\pi, \quad \mathbf{V} = \mathbf{K}_\theta + \Sigma_\varepsilon, \quad (1)$$

125 and predictions at new points x_* follow from the Gaussian conditional

$$\mathbb{E}[f(x_*) \mid \mathbf{y}] = m(x_*) + \mathbf{k}_*^\top \mathbf{V}^{-1}(\mathbf{y} - \mathbf{m}), \quad \text{Var}[f(x_*) \mid \mathbf{y}] = k(x_*, x_*) - \mathbf{k}_*^\top \mathbf{V}^{-1} \mathbf{k}_*. \quad (2)$$

126 Kernel hyperparameters θ are estimated by maximising (1) or by placing priors on them and
 127 sampling from the posterior.

128 **The phylogeny as an index set**

129 Let $\mathcal{T}$ be a rooted phylogeny with branch lengths, viewed as a continuous object: every point a on
 130 every branch is an ancestral lineage at a particular time. Let $h(a)$ be the distance from the root to a
 131 and $a \wedge b$ the most recent common ancestor of a and b . Brownian motion on the tree with rate σ^2
 132 starting from root state μ is the GP on $\mathcal{T}$ with

$$m(a) = \mu, \quad k_{\text{BM}}(a, b) = \sigma^2 h(a \wedge b). \quad (3)$$

Evaluated at the N tips this gives the familiar phylogenetic covariance matrix $\sigma^2 \mathbf{C}$, where $C_{ij} = t_{ij}$ is the shared path length from the root to the MRCA of species i and j . Evaluated at internal nodes, Eq. (2) gives ancestral-state reconstruction; evaluated at tips with missing data, it gives phylogenetic imputation. Both are kriging on a tree.

The BM kernel is non-stationary: it depends on the position of $a \wedge b$, not only on the distance between a and b . It is also Markov along the tree: conditional on its value at a node, the process in the subtrees below that node is independent. Hence the precision matrix of the process at all $2N - 1$ nodes is sparse, and the likelihood can be evaluated in $O(N)$ operations by Gaussian “pruning” (Felsenstein 1973; Freckleton 2012; Ho and Ané 2014a; Mitov et al. 2020; Bastide et al. 2021). This mirrors the link between GPs and Gaussian Markov random fields in spatial statistics (Lindgren et al. 2011) and is what makes GP models feasible on trees with thousands of tips.

Whole-tree processes versus tip kernels

Two distinctions matter for what the rest of the paper does and does not claim. *Index set.* A GP on $\mathcal{T}$ is defined at every point of the tree, whereas data live at the tips. Models defined by a generative process along branches (BM, early burst/ACDC, multi-rate BM, OU and OU shift models) are GPs on the whole tree, so kriging at internal points (ancestral-state reconstruction) is meaningful for them. Pagel’s κ and δ define BM on a *transformed* tree, so reconstructions refer to the transformed time scale. Pagel’s λ is specified through the tip covariance matrix only (BM plus tip-specific variance), so it fixes a distribution for the tips but, without further assumptions, not for internal points. Two models can therefore have identical tip distributions and still be different processes on the tree. This is exactly the case for OU and ACDC below: they coincide at the tips of an ultrametric tree, but ACDC has independent increments along the tree and OU does not, so their reconstructions of internal states differ.

Edge form and non-ultrametric trees. The edge form of Eq. (4) (next section) describes processes whose increments on different edges are independent, with edge variances $w_e \geq 0$. OU

with $\alpha > 0$ is not of this form on a general tree: along an edge, $x_c = \theta + e^{-\alpha \ell_e}(x_p - \theta) + \varepsilon_e$ with $\varepsilon_e \sim \mathcal{N}(0, \frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha \ell_e}))$, so the child state inherits a *shrunk* copy of its parent. The edge variances for OU given in Table 1 hold only on *ultrametric* trees and describe the distribution at the tips of such a tree; they do not extend to trees with tips at different heights. For those trees the OU model remains a valid GP (its stationary kernel $\frac{\sigma^2}{2\alpha}e^{-\alpha d_{ij}}$ is positive semi-definite on every tree, see below) and the likelihood is still evaluated in $O(N)$ by pruning, but with edge-specific coefficients $e^{-\alpha \ell_e}$ (Ho and Ané 2014a) rather than with Eq. (4). The software provided in this study implements the OU edge representation for ultrametric trees only (the OU kernel on a tree with fossils is used through its closed form, with dense algebra).

Classical PCMs as kernels

Many classical continuous-trait models can be written as a branch-wise variance decomposition,

$$\mathbf{K}(\theta) = \mathbf{B} \text{diag}\{\mathbf{w}(\theta)\} \mathbf{B}^\top + v_0 \mathbf{1}\mathbf{1}^\top + \text{diag}(\tau^2), \quad (4)$$

where $\mathbf{B}$ is the $N \times E$ tip-by-edge incidence matrix ($B_{ie} = 1$ if edge e lies on the path from the root to tip i), $w_e(\theta) \geq 0$ is the variance accumulated along edge e , v_0 is the variance of the root state and τ^2 is extra tip variance (measurement error or within-species variation; Ives et al. (2007)). The kernel of every model of this form is $k(i, j) = v_0 + \sum_{e \in \text{path}(i) \cap \text{path}(j)} w_e + \tau_i^2 \mathbb{1}_{i=j}$, and the $O(N)$ pruning likelihood applies to all of them.

Table 1 lists the models. Writing ℓ_e for the length of edge e and h_e^-, h_e^+ for the heights of its start and end above the root:

- **BM**: $w_e = \sigma^2 \ell_e$.
- **Pagel's κ** raises each *branch length* to the power κ : $w_e = \sigma^2 \ell_e^\kappa$. This is not the same as raising shared path lengths t_{ij} to the power κ .
- **Pagel's δ** transforms node *heights*: $w_e = \sigma^2 T^{1-\delta}[(h_e^+)^\delta - (h_e^-)^\delta]$. On any tree, with T the maximum tip height, this gives $k(i, j) = \sigma^2 T^{1-\delta} t_{ij}^\delta$.

- **Early burst / ACDC** (Blomberg et al. 2003; Harmon et al. 2010), with rate $\sigma^2(t) = \sigma^2 e^{rt}$:
 $w_e = \sigma^2(e^{rh_e^+} - e^{rh_e^-})/r$, so $k(i, j) = \sigma^2(e^{rt_{ij}} - 1)/r$.
- **Pagel's λ** multiplies internal branches by λ and lengthens terminal branches to keep tip heights fixed. On an ultrametric tree, $k(i, j) = \sigma^2[\lambda t_{ij} + (1 - \lambda)T \mathbb{1}_{i=j}]$. That is BM plus an independent white-noise kernel, i.e. the phylogenetic mixed model of Lynch (1991) and Housworth et al. (2004) in a different parameterisation; λ is the phylogenetic heritability.
- **Multi-rate BM** (O'Meara et al. 2006): $w_e = \sum_r \sigma_r^2 \ell_{e,r}$, where $\ell_{e,r}$ is the time edge e spends in regime r . The kernel is non-stationary in a way that depends on the regime painting.
- **OU** with selection strength α and a stationary (random) root: $k(i, j) = \frac{\sigma^2}{2\alpha} e^{-\alpha d_{ij}}$, where d_{ij} is the patristic distance (Hansen 1997; Martins and Hansen 1997). This is the exponential, or Matérn- $\frac{1}{2}$, kernel of spatial statistics applied to the tree metric. With a fixed root the kernel is multiplied by $(1 - e^{-2\alpha t_{ij}})$. The edge-variance form of this model used below assumes an ultrametric tree (see the previous subsection).

OU and accelerating rates. The edge representation makes a known identifiability result (Ho and Ané 2014b; Uyeda et al. 2015; Murrell 2018) transparent, and makes it possible to state exactly when it holds. Consider a single-optimum OU process on an ultrametric tree of height T , so that $d_{ij} = 2(T - t_{ij})$. (i) *Fixed root.* The tip means are all equal to $\theta + (x_0 - \theta)e^{-\alpha T}$, whatever the root state x_0 and optimum θ , and the covariance is $\frac{\sigma^2}{2\alpha} e^{-2\alpha T} (e^{2\alpha t_{ij}} - 1)$. This is *exactly* the ACDC/early-burst kernel with positive rate $r = 2\alpha$ and $\sigma_{\text{ACDC}}^2 = \sigma_{\text{OU}}^2 e^{-2\alpha T}$, with a common mean. The two models therefore have identical likelihoods under ML, REML or Bayesian inference with corresponding priors. Note that this is an equivalence of the *tip distributions*: as processes on the whole tree the two models differ (ACDC has independent increments, OU does not), so they imply different reconstructions of internal states. (ii) *Stationary (random) root.* The covariance gains the constant $\frac{\sigma^2}{2\alpha} e^{-2\alpha T}$, i.e. a random intercept. Under ML the two likelihoods differ. Once the intercept (or, in a regression, the coefficients) is profiled out, the random-root OU log-likelihood at corresponding kernel parameters equals the ACDC one minus

208 $\frac{1}{2} \log(1 + c \mathbf{1}^\top \mathbf{V}^{-1} \mathbf{1})$, where c is this constant and $\mathbf{V}$ the ACDC covariance, so its maximum can
 209 never exceed the ACDC maximum (Appendix S1). For a fixed mean $\mathbf{m}$ the difference contains, in
 210 addition, the quadratic term $\frac{cb^2}{2(1+ca)}$ with $a = \mathbf{1}^\top \mathbf{V}^{-1} \mathbf{1}$ and $b = \mathbf{1}^\top \mathbf{V}^{-1}(\mathbf{y} - \mathbf{m})$, which vanishes at
 211 the profiled intercept. However, because this constant lies in the column space of the intercept, the
 212 REML likelihood (which depends only on contrasts orthogonal to the intercept) is identical. I
 213 confirmed both statements numerically (Results). (iii) The equivalence does *not* extend to
 214 non-ultrametric trees, where tips differ in height, or to OU models with several optima, whose mean
 215 functions differ among tips. Within its scope, however, it implies that extant-only data cannot
 216 distinguish “stabilising selection towards one optimum” from “rates of evolution increasing
 217 towards the present”, which may explain part of the well-known difficulty of interpreting OU fits
 218 (Ho and Ané 2014b; Cooper et al. 2016; Uyeda et al. 2018). Regression models (PGLS) simply
 219 replace μ by $\mathbf{X}\beta$, and multivariate BM with evolutionary rate matrix $\mathbf{R}$ is the separable multi-output
 220 GP $\text{vec}(\mathbf{Y}) \sim \mathcal{N}(\text{vec}(\mathbf{1}\mu^\top), \mathbf{R} \otimes \mathbf{C})$ (“intrinsic coregionalisation”).

Table 1: Classical continuous-trait models written as Gaussian-process kernels on a tree. t_{ij} : shared path length from the root to the MRCA of tips i and j ; d_{ij} : patristic distance; T : tree height (ultrametric); ℓ_e , h_e^- , h_e^+ : length and start/end heights of edge e . All rows except the last two have the edge form of Eq. (4), so the $O(N)$ pruning likelihood applies; the OU rows have this form on ultrametric trees only (on other trees the stationary OU kernel is still valid, but pruning needs edge-specific coefficients). Rows give the distribution at the tips: for Pagel’s λ , and for the OU edge form, they do not determine the process at internal points of the tree (OU ancestral states must be computed from the OU covariance itself; see Discussion).

Model	Kernel $k(i, j)$	Edge variance w_e	Interpretation
BM	$\sigma^2 t_{ij}$	$\sigma^2 \ell_e$	drift at constant rate
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)	BM + white noise; phylogenetic heritability
Pagel’s κ	$\sigma^2 \sum_{e \text{ shared}} \ell_e^\kappa$	$\sigma^2 \ell_e^\kappa$	change concentrated at speciation ($\kappa \rightarrow 0$)

Model	Kernel $k(i, j)$	Edge variance w_e	Interpretation
Pagel's δ	$\sigma^2 T^{1-\delta} t_{ij}^\delta$	$\sigma^2 T^{1-\delta} [(h_e^+)^{\delta} - (h_e^-)^{\delta}]$	early ($\delta < 1$) or late ($\delta > 1$) change
Early burst / ACDC	$\sigma^2 (e^{rt_{ij}} - 1)/r$	$\sigma^2 (e^{rh_e^+} - e^{rh_e^-})/r$	rate $\sigma^2 e^{rt}$ declining ($r < 0$) or increasing
OU, fixed root	$\frac{\sigma^2}{2\alpha} e^{-\alpha d_{ij}} (1 - e^{-2\alpha t_{ij}})$	$\frac{\sigma^2}{2\alpha} e^{-2\alpha T} (e^{2\alpha h_e^+} - e^{2\alpha h_e^-})$	= ACDC with $r = 2\alpha$ (ultrametric, single optimum)
OU, random root	$\frac{\sigma^2}{2\alpha} e^{-\alpha d_{ij}}$ (Matérn- $\frac{1}{2}$)	as above, $v_0 = \frac{\sigma^2}{2\alpha} e^{-2\alpha T}$	stationary; = ACDC under REML (ultrametric)
Multi-rate BM	$\sum_{e \text{ shared}} \sum_r \sigma_r^2 \ell_{e,r}$	$\sum_r \sigma_r^2 \ell_{e,r}$	regime-specific rates (non-stationary)
Measurement error	$k(i, j) + s_i^2 \mathbb{1}_{i=j}$	$\tau_i^2 = s_i^2$	known within-species error
Multivariate BM	$\mathbf{R} \otimes \mathbf{C}$	—	correlated evolution (coregionalisation)
PGLMM	latent $f \sim \mathcal{GP}(0, \sigma^2 \mathbf{C})$, $y_i \sim p(\cdot $ $g^{-1}(\mathbf{x}_i^\top \beta + f_i))$	—	non-Gaussian traits; threshold model (probit)

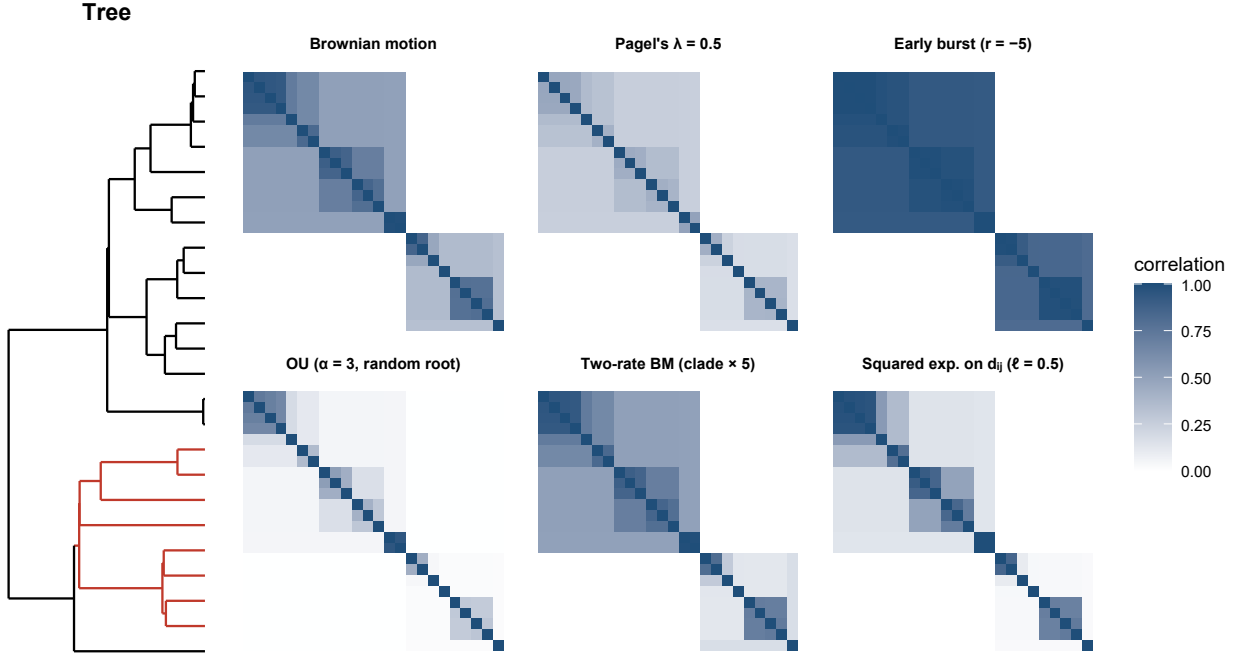


Figure 1. One 24-species tree (left; red: the clade given an elevated rate in the two-rate model) seen through six kernels. Tips are in the same order in the tree and in the correlation matrices, so clades appear as blocks. Each matrix shows the implied correlation between species. The kernel alone determines 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.

Non-Gaussian traits enter as latent GPs. A phylogenetic GLMM (Lynch 1991; Hadfield and Nakagawa 2010; Ives and Helmus 2011) is $g(\mathbb{E}[y_i]) = \mathbf{x}_i^\top \beta + f(i)$ with $f \sim \mathcal{GP}(0, \sigma_{\text{phy}}^2 k_{\text{BM}})$, and the threshold model (Felsenstein 2005) is the special case with a probit link and fixed residual variance.

Kernel design

One of the advantages of the GP view of PCMs is that model building becomes kernel design (Rasmussen and Williams 2006; Duvenaud et al. 2013). Sums of kernels describe independent processes acting together, e.g. phylogenetic + spatial + residual variation (Freckleton and Jetz 2009). Products of kernels describe interactions; for instance $k((i, x_i), (j, x_j)) = x_i x_j k_{\text{BM}}(i, j)$ is the kernel of a regression slope that itself evolves by BM along the tree. Kernel parameters can

be given their own GP priors, giving hierarchical, non-stationary models, such as a log-rate that evolves along the tree [as in relaxed molecular clocks; Thorne and Kishino (2002)] or a rate that varies smoothly through time. Section *Kernel extensions* below implements each of these.

A kernel used for a GP must be positive semi-definite on the index set. For kernels written as functions of patristic distance, $k(i, j) = \phi(d_{ij})$, the answer depends on tree shape. On an *ultrametric* tree, $d_{ij} = 2(T - t_{ij})$, so any non-negative, non-increasing ϕ is a non-decreasing function $g(t_{ij})$ of shared time. It is therefore a tree transformation of the form (4) (equivalently, $[g(t_{ij})]$ is an ultrametric matrix, and such matrices are positive semi-definite; Dellacherie et al. (2014)), with $w_e = g(h_e^+) - g(h_e^-) \geq 0$ and $v_0 = g(0)$, and is always valid. On *non-ultrametric* trees (fossil tips, or phylograms not constrained to a clock), the argument fails. Tree metrics embed isometrically in L_1 and are therefore of negative type (Deza and Laurent 1997), so by Schoenberg’s theorem (Schoenberg 1938) $e^{-d/\ell}$, the OU kernel, remains positive semi-definite for every length-scale ℓ . By Bernstein’s theorem, so does every completely monotone ϕ , i.e. every mixture of exponentials $\phi(d) = \int e^{-sd} \mu(ds)$; examples are the generalised Cauchy form $(1 + d/\ell)^{-a}$ (not the usual rational-quadratic kernel, which is a function of d^2 and is not guaranteed valid on such trees) and the stretched exponential $e^{-(d/\ell)^p}$ with $0 < p \leq 1$. This is a sufficient condition, not a characterisation of all kernels that happen to be valid on a particular tree. In contrast, the squared-exponential kernel $e^{-d^2/2\ell^2}$, the default in most GP software and the “L2” kernel of McElreath (2020), is guaranteed positive definite only for Euclidean distances (cf. Feragen et al. 2015; for Matérn kernels on graphs, which are constructed differently, see Borovitskiy et al. 2021). Formal statements and proofs, including a 5-tip counterexample, are in Appendix S1. Tree metrics with unequal root-to-tip distances are generally not Euclidean, and the same caveat applies to Matérn kernels with smoothness $\nu > \frac{1}{2}$. I show numerically below that these kernels do fail on such trees.

Materials and Methods

Implementation

I implemented Eq. (4) in Stan (Carpenter et al. 2017) through `cmdstanr` (Gabry et al. 2024); the programs used here are released, in generalised form, in the R package `phylogp` (Data availability). The program `gp_phylo.stan` fits regressions $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 \mathbf{K}(\theta) + \text{diag}(\mathbf{s}^2))$ for the BM, λ , κ , δ , EB, OU-fixed-root and OU-random-root kernels. It offers three numerically equivalent likelihoods: (i) the $O(N)$ pruning recursion; (ii) a dense kernel built from Eq. (4); and (iii) the closed-form kernels written as functions of t_{ij} and d_{ij} (the “textbook” GP form). Other programs implement multi-rate BM, multivariate BM, latent-GP GLMMs and the kernel extensions described below. Maximum-likelihood (ML) estimates were obtained with Stan’s L-BFGS optimiser, with no priors and no Jacobian adjustment. The posterior mode then equals the ML estimate, and `lp__` equals the exact log-likelihood because all likelihood terms are accumulated with `target +=`. Bayesian fits used four chains of 1,000 warm-up and 1,000 sampling iterations (500 and 500 in the replicated simulation study, i.e. 2,000 post-warm-up draws per fit), with weakly informative priors (half-normal on standard deviations, uniform on bounded kernel parameters). Convergence was checked with $\hat{R}$, effective sample sizes and divergent transitions.

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* θ is available in closed form, $p(y_i | \mathbf{y}_{-i}, \theta) = \mathcal{N}(y_i | y_i - [\mathbf{V}^{-1}\mathbf{r}]_i / [\mathbf{V}^{-1}]_{ii}, 1/[\mathbf{V}^{-1}]_{ii})$ with $\mathbf{r} = \mathbf{y} - \mathbf{m}$ (Sundararajan and Keerthi 2001). This conditional density is exact, but it is not the quantity that leave-one-out cross-validation targets, which integrates over the posterior of θ given the data *without* observation i . I approximate that integral by importance-sampling the posterior draws obtained with all data (Pareto-smoothed importance sampling, PSIS (Vehtari et al. 2017; Bürkner et al. 2021)), without refitting the model N times. The resulting expected log predictive density (elpd) is therefore an importance-sampling estimate of parameter-integrated LOO, whose reliability is indicated by the Pareto- $\hat{k}$ diagnostics (summarised in the Results), not an exact LOO. Where

$\hat{k} > 0.7$, the estimate for that observation can be replaced by an exact value obtained by refitting without it; I did this for selected simulation replicates and for every influential species in the mammal analysis, so those comparisons combine PSIS with exact refits. For models with latent tree processes, the left-out species’ own latent values must not be conditioned on, because only y_i informs them. For the varying-slope kernel, the latent slopes and intercepts are integrated out analytically (the marginal kernel is Gaussian). For the relaxed-rate kernel, the rate of a terminal branch depends on the latent log-rate at the tip itself. I integrate that value out against its conditional prior given its parent node by Gauss–Hermite quadrature (30 nodes; this agreed with brute-force numerical integration to 10^{-14}), so that each conditional LOO density is $p(y_i \mid \mathbf{y}_{-i}, \theta_{-i})$ with θ_{-i} excluding species i ’s own latent rate. Computing these terms densely costs $O(N^3)$ per posterior draw. For kernels of the form (4) I instead used an upward–downward pass of Gaussian messages over the tree: the downward (“outside”) message arriving at tip i gives the distribution of the latent tip value x_i given $\mathbf{y}_{-i}$ and θ , and adding the tip variance τ_i^2 gives $p(y_i \mid \mathbf{y}_{-i}, \theta)$ exactly, so all N conditional leave-one-out densities cost $O(N)$ per draw. This agreed with the dense calculation to 10^{-10} in the tests carried out.

Equivalence with standard software

Univariate models. I simulated a pure-birth tree with 100 tips scaled to height 1. For each of seven models I simulated 100 data sets under $y = 1 + 0.5x + e$, with $x \sim \text{BM}$ and $e \sim \mathcal{N}(\mathbf{0}, \mathbf{K}(\theta))$. True parameters were $\lambda = 0.5$, $\kappa = 0.4$, $\delta = 2$, $r = -4$ (EB) and $\alpha = 2$ (both OU variants). Each data set was fitted by ML with `phylolm` (Ho and Ané 2014a), using the same parameter bounds as the GP, and with the GP model. For one further data set per model I also compared `geiger::fitContinuous` (Pennell et al. 2014) (intercept-only models), `nlme::gls` (Pinheiro et al. 2023) with `ape` correlation structures, and `caper::pgls` (Orme et al. 2023), always with the same parameter bounds as the GP. For one data set per model I compared the Bayesian GP posterior of the kernel parameter with the profile likelihood. Because a single ultrametric tree is a narrow test, I repeated the `phylolm` comparison on four further trees: a random non-clock tree (100

tips) and a birth–death tree that retains extinct lineages (143 tips), both non-ultrametric; an ultrametric tree with ten polytomies (100 tips); and a 1,000-tip ultrametric tree. I used 20 data sets per model (5 for the large tree). OU models were fitted only on the ultrametric trees, because the OU edge representation implemented in the software assumes that all tips are at the same height (see *Whole-tree processes versus tip kernels*).

Ancestral states, multivariate and multi-rate models. I compared GP kriging at internal nodes with `phytools::fastAnc` and `ape::ace` (Revell 2012; Paradis and Schliep 2019); the λ model with BM + measurement error in `phylolm`; bivariate BM with `mvMORPH::mvBM` (Clavel et al. 2015); and a two-regime BM (a 6-fold rate increase in one clade) with `mvMORPH`’s BMM model.

Phylogenetic GLMMs. I simulated binary and count traits from latent-BM GLMMs and fitted them with a latent-GP Stan program and with `brms` (Bürkner 2017) using a phylogenetic random effect and identical priors.

Kernel extensions

I implement four kernel extensions, each aimed at a different macroevolutionary question. None is new as a statistical model, and each has close precedents: (a) is a Bayesian counterpart of the autocorrelated-rate Brownian motion that `multirateBM` fits by penalised likelihood (Revell 2021), and is related to autocorrelated relaxed clocks (Thorne and Kishino 2002); (b) generalises early burst to a flexible function of time, in the spirit of time- and environment-dependent rate models (Clavel and Morlon 2017); (c) is the phylogenetic + spatial model of Freckleton and Jetz (2009); and (d) is the marginal form of a phylogenetic random-slope mixed model, which mixed-model software such as `MCMCglmm` and `brms` can already fit (Hadfield and Nakagawa 2010; Bürkner 2017). What the kernel view adds is that all four are written, fitted and checked in one framework, with linear-time likelihoods and leave-one-out densities where the tree structure allows. All four are tested by simulation and applied, as an illustration, to the mammal data. (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 $z \sim \mathcal{GP}(0, \tau^2 k_{\text{BM}})$,

where a_e and d_e are the start and end nodes of edge e . It asks whether, and in which lineages, a trait evolved faster or slower than elsewhere, without specifying the shifts in advance (e.g. bursts of brain-size evolution in particular clades). (b) **Rate through time:** $w_e = \int_{h_e^-}^{h_e^+} \sigma^2(s) ds$ with $\log \sigma^2$ piecewise constant over 20 time bins and a Gaussian random-walk prior. A piecewise-constant rate can approximate the exponential rate trajectory of EB/ACDC, which is the limit as the bins become narrow, but EB is not an exact special case of the binned model. It asks how the pace of evolution changed through time across the whole clade, for example after a mass extinction or during a period of climatic change, without assuming a monotone trend. (c) **Phylogeny + geography:** $\mathbf{K} = \sigma_p^2 \mathbf{C} + \sigma_g^2 e^{-\mathbf{D}_{\text{geo}}/\rho} + \sigma_e^2 \mathbf{I}$. It separates resemblance due to shared ancestry from resemblance due to geographic proximity (shared environments, local adaptation or dispersal), which otherwise inflates or deflates estimates of phylogenetic signal. (d) **Varying slopes:** $y_i = \beta_0 + (\beta_1 + v_i)x_i + u_i + e_i$ with u, v independent BM processes, i.e. marginal kernel $\sigma_u^2 \mathbf{C} + \sigma_v^2 (\mathbf{xx}^\top) \circ \mathbf{C} + \sigma_e^2 \mathbf{I}$. It asks whether the relationship between two traits, such as an allometric scaling exponent, itself evolved, and where on the tree it changed. To keep cost linear in N , models (a), (b) and (d) sample the latent node values (non-centred) and integrate out the remaining Gaussian components by pruning.

Illustrative fits. For each kernel I first show one simulated data set in detail: an 8-fold rate increase in a 25-tip clade of a 150-tip tree (a), a pulse of elevated rate halfway through the history of a 200-tip tree (b), phylogenetically structured species locations on a 150-tip tree (c), and BM-evolving slopes ($\sigma_v = 0.6$) on a 200-tip tree (d). These illustrate what the models estimate; they do not establish their statistical properties.

Replicated simulation study. To assess bias, calibration, power and false positives, I simulated each kernel under an *effect* scenario (the process present) and a *null* scenario (the process absent), each on a freshly simulated pure-birth tree per replicate: 50 replicates per scenario for (a), (b) and (d) (100 tips) and 30 for the more expensive dense model (c) (80 tips). Effect scenarios were: an 8-fold rate increase in a ~20-tip clade (a); the mid-history rate pulse (b); $\sigma_g = 1, \rho = 0.3$

(c); and $\sigma_v = 0.6$ (d). Null scenarios were constant-rate BM (a, b), $\sigma_g = 0$ (c) and $\sigma_v = 0$ (d). Each extended kernel was compared with its natural simpler alternative (BM for a and b, phylogeny + white noise for c, and constant-slope PGLS with Pagel’s λ for d) by PSIS-LOO built on the conditional Gaussian predictive densities described above. I counted a “detection” when the extended kernel’s expected log predictive density exceeded the alternative’s by more than two standard errors. I report detection rates (power under effect scenarios, false-positive rate under nulls), the coverage of nominal 90% posterior intervals, point-estimate bias, and sampler diagnostics. With 50 replicates, a rate near 90% has a Monte Carlo standard error of about 4 percentage points, and zero detections in n null replicates are compatible with a true false-positive rate up to $1 - 0.05^{1/n}$ (one-sided 95% upper bound: 5.8% for $n = 50$, 9.5% for $n = 30$). For the rate-through-time model my first runs (random-walk SD $\tau \sim \text{half-normal}(0, 1)$, target acceptance 0.95) produced divergent transitions in 92–98% of replicates (median 5 of 2,000 post-warm-up draws). In a diagnostic on 8 replicates per scenario, a target acceptance of 0.99 together with $\tau \sim \text{half-normal}(0, 0.5)$ removed the divergences. I re-ran the rate-through-time replicates with these settings, and Table 2 uses them. Because the prior was changed after seeing the first run, Appendix S3 reports both runs as a sensitivity analysis (the first-run results are archived with the code).

Kernel validity. I also computed the smallest eigenvalue of exponential, Matérn-3/2, Matérn-5/2 and squared-exponential kernels on patristic distances over a range of length-scales, for three trees: an ultrametric pure-birth tree (100 tips), a birth–death tree that retains its extinct lineages as fossil tips (197 tips), and a random non-clock tree (100 tips). Formal statements and proofs of the results used here are in Appendix S1.

Computational scaling. I compared the $O(N)$ pruning likelihood with a dense Cholesky likelihood for BM on pure-birth trees of 50 to 10,000 tips. I timed one log-likelihood evaluation, the computation of all N conditional LOO densities, exact given the parameters (message passing vs. dense inversion), and a complete ML fit with `phylo1m`. Both likelihoods were implemented in R without compiled code, so the comparison concerns algorithms, and it is conservative for pruning (an interpreted loop over $2N - 2$ edges). Scaling exponents are slopes of log time on log N for

$N \geq 500$. The dense matrix was not attempted beyond 3,000 tips. For the Stan implementation I timed one evaluation of the log density and its gradient, the operation that optimisation and HMC repeat, with the compiled model methods of `cmdstanr` (BM kernel; pruning up to 5,000 tips, dense kernel up to 1,000 tips, where its automatic-differentiation memory became limiting). I also timed complete HMC runs of the relaxed-rate kernel (four chains of 500 warm-up and 500 sampling iterations) at 100, 300 and 1,000 tips, recording wall time and cost per gradient (sampling time divided by the number of leapfrog steps).

Fossil tips and identifiability. I simulated birth–death trees that retain extinct lineages (80–300 tips) and generated data from a fixed-root OU process ($\alpha = 2$) or from an ACDC process ($r = 4$). The OU root state was set equal to the optimum, so that all tips have the same expected value and the two models differ only in their covariances. This is the hardest case: with $x_0 \neq \theta$, expected values change with tip height, which gives extra information for distinguishing the models. Each data set was fitted by ML with the OU and ACDC ($r > 0$) kernels, once on the full tree and once after discarding fossil tips (an ultrametric tree), and I recorded $\log L_{\text{OU}} - \log L_{\text{ACDC}}$. *Multiple optima.* I checked that a two-optimum OU model is the same kernel with a regime-dependent mean by comparing its analytic mean and covariance with 40,000 simulated tip vectors (Appendix S2).

Additional simulations. To probe the limits of the kernel extensions I added: weaker effects (a 3-fold rate shift; a rate pulse of amplitude 1 instead of 4; $\sigma_v = 0.3$; $\sigma_g = 0.5$); a 300-tip tree for relaxed rates (8-fold shift in a ~30-tip clade); two misspecified data-generating processes for the relaxed-rate kernel (fixed-root OU on an ultrametric tree, and BM with three jumps of SD 2); Designs, priors, the LOO decision rule and diagnostics were as in the main study. *Empirical tree shapes.* Because pure-birth trees are a narrow class, I repeated the relaxed-rate and varying-slope designs (effect and null, 30 replicates each) on random 100-species subtrees of the mammal tree of Upham et al. (2019), rescaled to unit height, with the `phylogp` package and $4 \times 1,000$ post-warm-up draws. *Comparison with multirateBM.* On 20 random 50-species subtrees of the same tree with an 8-fold rate increase in a clade of about ten species, I compared the relaxed-rate kernel with

multirateBM (Revell 2021) (phytools 2.5.2, default settings: maximum likelihood, optimisers L-BFGS-B, Nelder–Mead, BFGS and CG, penalty $\lambda = 1$, one iteration; 30-minute limit per fit, never reached). multirateBM estimates rates at nodes; as for the kernel, an edge's rate was the geometric mean of the rates at its two ends. Both were scored by the Spearman correlation between estimated and true branch rates, the recovered clade/background rate ratio and the run time.

Case study: mammalian brain size

Brain size relative to body size is one of the most studied traits in comparative biology (Pagel and Harvey 1989; Smaers et al. 2021). I use it to illustrate how kernels, and the tools around them, are used to weigh an evolutionary explanation against simpler alternatives. The analysis is not intended as a comprehensive test of hypotheses about mammalian brain evolution: any biological conclusion would require further work on data quality, measurement error, taxonomic matching and a wider range of models.

I used mean adult brain and body masses compiled by Burger et al. (2019a, 2019b) and the DNA-only, node-dated maximum clade credibility tree of Upham et al. (2019). Species names were matched exactly; unmatched species were dropped, polytomies were resolved with zero-length branches and the tree was made exactly ultrametric. I modelled $\log_{10}$ brain mass as a function of $\log_{10}$ body mass centred at its mean; the centring is the reference point for slope deviations in the varying-slope kernel. For the classical kernels I first compared maximum-likelihood fits from `phylolm` and the GP implementation.

A more complex kernel can predict better without the process it encodes having operated. Three alternatives are relevant here. A varying slope makes the variance of brain size grow with distance from the centre of body size, so it can capture heteroscedasticity. A relaxed rate on terminal branches makes each species' residual a scale mixture of normal distributions, so it can capture heavy-tailed residuals, that is, outlying species. And both can absorb curvature in the brain–body relationship, which Venditti et al. (2024) showed to be log-curvilinear across mammals.

436 I therefore crossed three process kernels (BM, phylogenetically varying slopes $k_{\text{slope}}(x) \times k_{\text{BM}}$,
 437 and relaxed rates) with three residual models (Gaussian; heteroscedastic, $k_{\text{slope}}(x) \times k_{\text{white}}$,
 438 i.e. independent species-specific slope deviations whose variance is proportional to squared centred
 439 body mass; and Student- t with four degrees of freedom, a scale mixture with a latent variance
 440 multiplier per species) and with a linear or quadratic mean, in the combinations listed in Fig. 6a.
 441 Every model included a species-level residual.

442 All models were fitted with phylogp (Data availability): four chains of 1,000 warm-up and
 443 1,000 sampling iterations, target acceptance 0.95, with $\hat{R}$ and effective sample sizes computed for
 444 the model parameters. Models were compared by leave-one-out predictive density; every species
 445 with $\text{Pareto-}\hat{k} > 0.7$ was refitted without its data and its contribution replaced by the exact value
 446 (`loo_refit()`), so the comparisons do not rest on unreliable importance-sampling estimates.
 447 Because leave-one-out mostly asks how well a species is predicted from its closest relatives, I also
 448 left out each of the 28 orders in turn and scored the joint predictive density of its species
 449 (`loo_group()`) for four key models. To account for phylogenetic uncertainty, I refitted the same
 450 four models to 50 trees drawn at random from the posterior sample of Upham et al. (2019) and
 451 pooled the posterior draws (trees weighted equally) for leave-one-out. For the phylogeny +
 452 geography kernel I took the midpoint of each species' geographic range from PanTHERIA (Jones
 453 et al. 2009) and computed great-circle distances between midpoints; the exponential kernel is
 454 positive definite for great-circle distances on the sphere (Gneiting 2013), so the additive kernel is
 455 valid. Because this kernel is dense, I fitted it by maximum likelihood and compared it with the
 456 phylogeny-only model (BM plus white noise) by AIC and by Gaussian leave-one-out densities at
 457 the ML estimates, with and without marine mammals (Appendix S3).

Results

Implementation equivalence: classical PCMs are recovered exactly by their GP kernels

For the single test data sets, the log-likelihoods given by `phylo1m` and by the three GP likelihood forms (pruning, dense edge kernel, closed-form kernel) agreed to machine precision for all seven models. Across the 700 simulated data sets, ML estimates from the GP and from `phylo1m` fell on the 1:1 line (Fig. 2), with two exceptions described below. For all models except λ , the largest absolute difference in log-likelihood was 1.9×10^{-7} , and in the regression slope 9.0×10^{-6} , i.e. optimiser tolerance; over all 698 remaining data sets it was 1.9×10^{-7} . The exceptions were λ data sets near the boundary $\lambda \approx 0$: in one, `phylo1m` stopped at the boundary and the GP found a higher likelihood (by 0.13 units), and in one the GP optimiser stopped early (lower by 0.005). These are differences between optimisers, not between models. (The tolerance statement above does not apply to them.) With default settings, `phylo1m` also restricts the EB rate to $r \geq -3/T$, which excluded the true value used here; once both programs used the same bounds, the EB estimates agreed exactly. Table S1 (Appendix S3) shows that `geiger`, `nlme`, `caper` and `phylo1m` give the same log-likelihoods and estimates as the GP (to four or more significant figures) wherever they implement the same model. Bayesian posteriors for the kernel parameters closely tracked the profile likelihoods (Fig. S1). On the four additional trees (non-clock, fossil, polytomies, 1,000 tips), the two programs agreed to within 2.8×10^{-6} log-likelihood units in 373 of 375 fits. In the remaining 2, the GP fit had the higher likelihood (by up to 0.11; all were Pagel's λ fits near the boundary, as on the original tree).

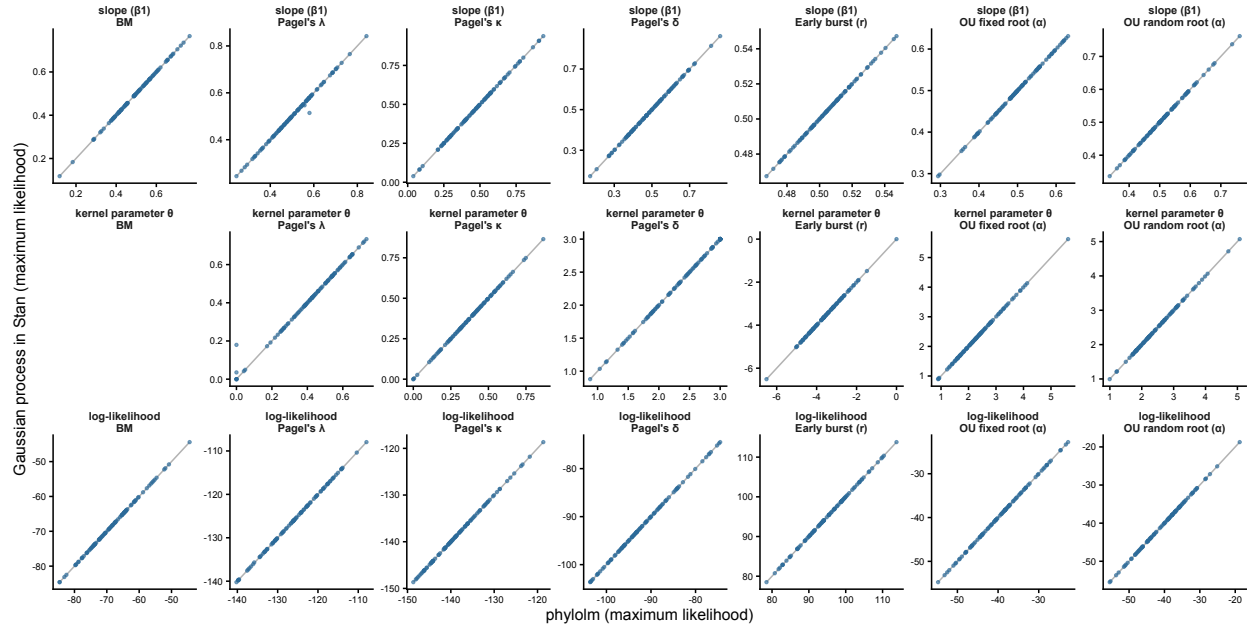


Figure 2. Maximum-likelihood estimates from phylo1m (x-axis) and from the Gaussian-process implementation in Stan (y-axis) for 100 simulated data sets under each of seven classical models. Rows: regression slope, kernel parameter, maximised log-likelihood. BM has no kernel parameter, so that panel is empty. The off-diagonal points for Pagel's λ are the boundary cases described in the text.

GP prediction at internal nodes reproduced fastAnc ancestral states (maximum difference 3.6×10^{-14}). The reported variances also matched (4.9×10^{-13}) once the root was integrated out under a flat prior, i.e. fastAnc performs *universal kriging* with the REML rate. Across 1,000 replicate data sets with 20 species held out each time, nominal 95%, 90% and 50% GP prediction intervals covered 94.5%, 89.6% and 49.9% of held-out values (one replicate is shown in Fig. S2). The λ model and BM + white noise gave identical maximised likelihoods (-116.1944 vs. -116.1944), and the implied λ values matched (0.6076 vs. 0.6076). The Kronecker GP reproduced mvMORPH bivariate BM (log-likelihood -152.1208 vs. -152.1208; evolutionary covariance 0.697 vs. 0.697). The non-stationary two-rate kernel reproduced mvMORPH BMM (log-likelihood -95.6601 vs. -95.6601; regime rates 1.04 and 5.46, true values 1 and 6). Latent-GP and brms posteriors for Bernoulli and Poisson phylogenetic GLMMs agreed within Monte Carlo error (Fig. S3). Finally, on an ultrametric tree a fixed-root OU fit and an ACDC fit with $r = 2\alpha$

gave identical maximised log-likelihoods (-53.6873 vs. -53.6873), rates ($r = 3.593$ vs. 3.593) and scale parameters. For the random-root OU, the ML log-likelihoods differed (-54.530 vs. -54.211) but the REML log-likelihoods were identical (-55.2848 vs. -55.2848), as predicted. These comparisons establish that the implementations compute the same models; they say nothing about how well those models recover evolutionary processes.

Kernel extensions: validity, illustrative fits and replicated simulations

Kernel validity. On the ultrametric tree, all four distance-based kernels were positive semi-definite at every length-scale, as the theory predicts (the smallest eigenvalue was positive, $7.4\text{e-}08$, but close to zero, as expected for the near-singular matrices that arise at long length-scales; they need no correction other than standard numerical care). On the tree with fossil tips and on the non-clock tree, 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. The exponential (OU) kernel, which is completely monotone, remained valid throughout (smallest eigenvalue 0.00016; Fig. 3). Indefinite matrices are not covariance matrices. In practice they cause Cholesky failures, or are silently ‘fixed’ by adding diagonal jitter, which changes the model being fitted.

Illustrative fits. Before the replicated study, I fitted each of the four kernels to one simulated data set with a strong, known effect. These fits show what each kernel estimates and how its output can be read, but a single data set says nothing about error rates, which the replicated study below addresses. *Relaxed rates.* Without being told where the shift lay, the autocorrelated-rate GP recovered the clade with elevated rate: the median posterior rate on branches inside the clade was 6.8 times that elsewhere (true ratio 8), and the elevated rates were confined to the correct clade (Fig. 4). The rate-variability parameter was estimated well away from zero ($\tau = 1.08$, 90% CrI 0.77–1.52). PSIS-LOO favoured the relaxed kernel over constant-rate BM ($\Delta\text{elpd} = 38.5 \pm 11.6$).

Rate through time. The rate-through-time GP recovered the mid-history pulse in rate, which no monotone EB/ACDC model can represent (the ML early-burst fit reached its bound; Fig. 5a).

516 Uncertainty was large early in the tree’s history, when few lineages existed. *Phylogeny* +
 517 *geography*. When species locations were phylogenetically structured, a phylogeny-only model
 518 absorbed the spatially generated variance into the residual term and underestimated the
 519 phylogenetic share (median 0.18 vs. true 0.25). The additive kernel recovered all three components
 520 (phylogeny 0.24, geography 0.72, residual 0.02; truth 0.25, 0.69, 0.06), and LOO strongly preferred
 521 it ($\Delta\text{elpd} = 62.0 \pm 8.4$; Fig. 5b). *Varying slopes*. The product kernel’s species-specific slopes
 522 correlated with the true slopes ($r = 0.77$), and the slope-variability parameter was estimated at 0.41
 523 (true 0.6). The model outperformed a constant-slope PGLS by LOO ($\Delta\text{elpd} = 11.9 \pm 5.8$; Fig.
 524 5c).

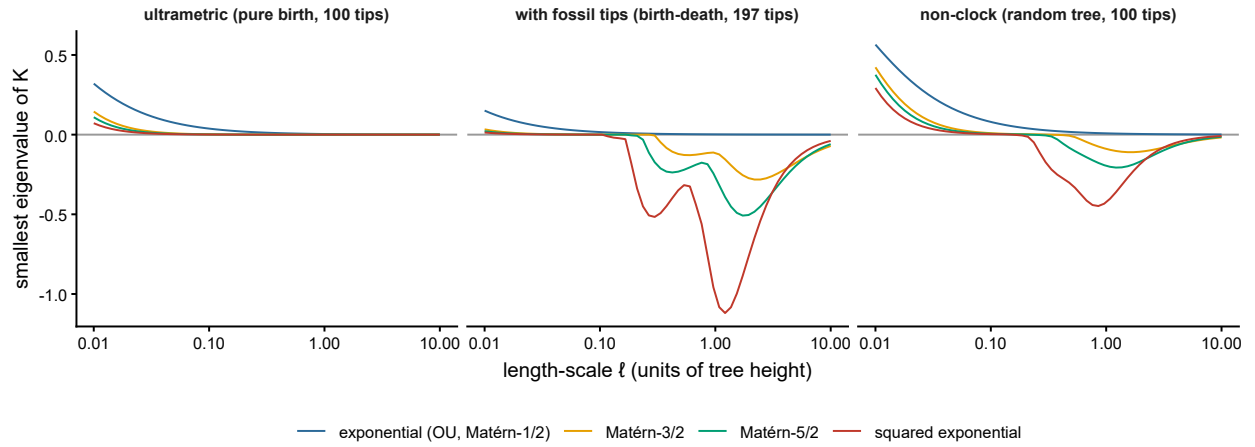


Figure 3. Smallest eigenvalue of kernel matrices $\phi(d_{ij}/\ell)$ computed from patristic distances. On an ultrametric tree every non-increasing $\phi \geq 0$ is valid. On trees with fossil tips or without a clock, values below zero mean the ‘kernel’ is not a valid covariance. Of the four kernels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

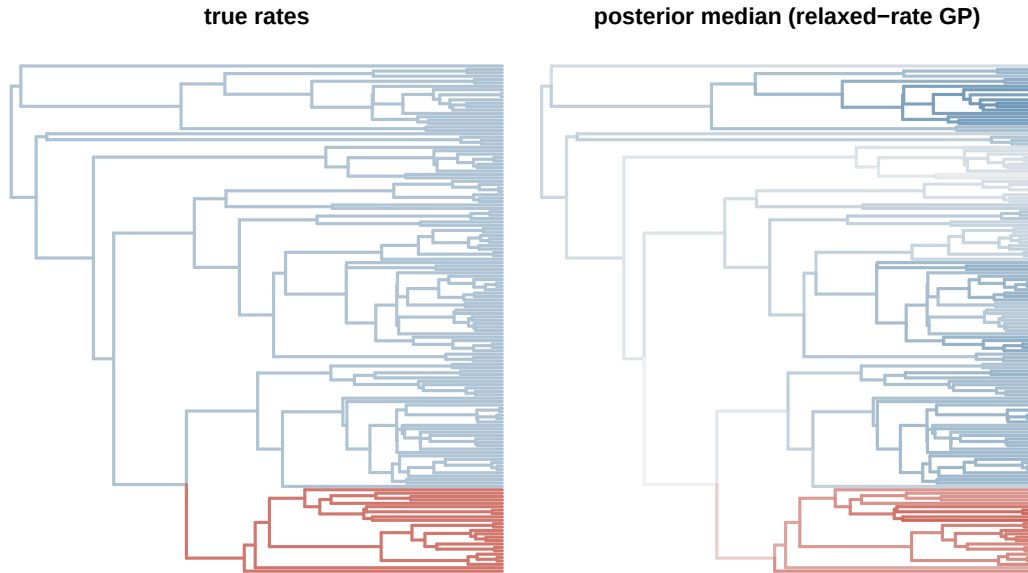


Figure 4. Relaxed (autocorrelated) rates as a hierarchical GP. Left: simulated branch rates (8-fold increase in one clade). Right: posterior median branch rates from the relaxed-rate kernel, which was not told the location of the shift. Blue, slow; red, fast (log scale).

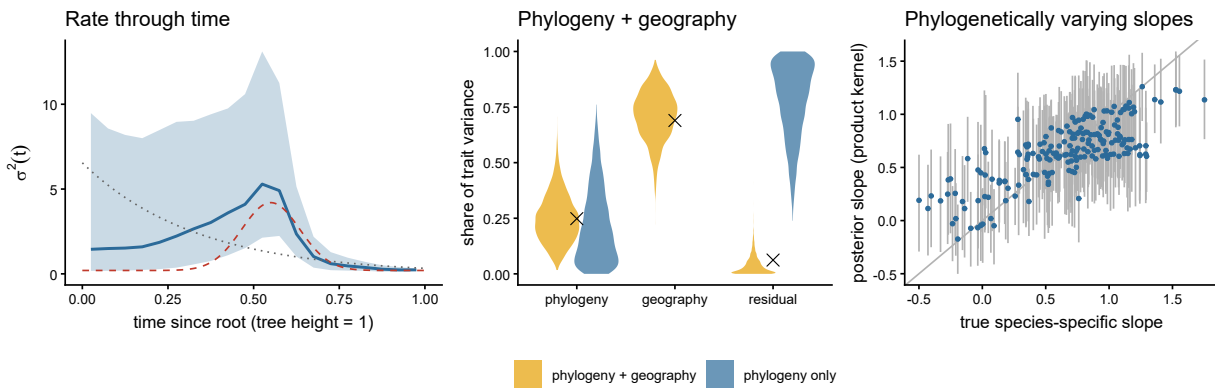


Figure 5. Kernel extensions recover known processes. (a) Rate-through-time GP: posterior median and 90% band of $\sigma^2(t)$ (blue), truth (red dashed) and the ML early-burst fit (grey dotted). (b) Share of trait variance attributed to phylogeny, geography and residual by a phylogeny-only model and by the additive phylogeny + geography kernel (crosses: truth). (c) Species-specific allometric slopes from the product (varying-slope) kernel versus truth (bars: 90% intervals).

525 *Replicated simulations.* Table 2 summarises the replicated study of all four kernels. Under
 526 the null scenarios, the kernel extensions were preferred over their simpler alternatives in 0%
 527 (relaxed rates), 0% (rate through time), 0% (varying slopes) and 0% (geography) of replicates.

Under the effect scenarios, the corresponding detection rates were 82%, 74%, 34% and 87%. The largest $\hat{R}$ across all fits was 1.048; the largest $\hat{R}$ exceeded 1.01, the current threshold for reliable inference, in 92% of relaxed rates replicates, 11% of rate through time replicates, 89% of varying slopes replicates, 37% of geography replicates, and the maximum occurred in the varying slopes study. These maxima are taken over every quantity in a fit, including hundreds of latent node values and all pointwise log-likelihood terms, from 2,000 post-warm-up draws, so values slightly above 1.01 are expected by chance even when the main parameters have converged. To check, I reran 40 fresh replicates of the relaxed-rate and varying-slope designs with $4 \times 1,000$ post-warm-up draws and computed $\hat{R}$ separately for the hyperparameters and the latent node values. With the original 4×500 draws, the hyperparameters had $\hat{R}$ at most 1.019 (10% of replicates above 1.01), while the latent values exceeded 1.01 in 95% of replicates but only for 3.3% of the latent values at most. With $4 \times 1,000$ draws every hyperparameter and latent value had $\hat{R} \leq 1.009$, and the smallest bulk effective sample size of a hyperparameter was 568. The reported maxima therefore reflect short chains for a few latent values, not unconverged model parameters. The proportions of replicates with any divergent transition were relaxed 10%, ratetime 3%, varslope 9%, space 0% (median number of divergent transitions per replicate, out of 2,000 post-warm-up draws: relaxed 0, ratetime 0, varslope 0, space 0). The largest Pareto- $\hat{k}$ in any PSIS-LOO computation was 1.33, and 36% of the replicate-level model comparisons had a largest $\hat{k}$ above 0.7, where importance-sampling estimates of LOO are considered unreliable. To check whether this matters, I recomputed the elpd exactly, by refitting without the observation, for every species with $\hat{k} > 0.7$ in 20 effect and 20 null replicates of the relaxed-rate study (on average 4.6 and 2.0 refits per replicate). The model-selection decision changed in 1 of 40 replicates (detection 75% with PSIS vs. 70% exact; false positives 0% vs. 0%), and the median elpd difference was 16.8 vs. 16.9 in the effect scenario. So the high $\hat{k}$ values flag individually influential species (mostly in the fast clade), but did not distort the comparisons reported here.

Sampler diagnostics. For the rate-through-time kernel, the original prior and sampler settings produced divergent transitions in most replicates. A higher target acceptance alone reduced

but did not remove them; they disappeared with a tighter prior on the random-walk SD (Appendix S3). With that prior, detection in the effect scenario was 74% (first run 72%), and mean per-bin coverage fell from 0.91 to 0.85, because the tighter prior shrinks sharp changes in rate.

Table 2: Replicated simulation study of the kernel extensions. 'Preferred' means an expected log predictive density (PSIS-LOO from conditional Gaussian predictive densities) more than two standard errors above the alternative: BM for relaxed rates and rate through time, constant-slope Pagel's λ PGLS for varying slopes, and phylogeny + white noise for geography. Under null scenarios, 'extension preferred' is a false-positive rate. Coverage refers to nominal 90% posterior intervals. For relaxed rates the model is misspecified for a discrete shift (it assumes autocorrelated, gradually changing rates), so coverage of the true ratio is not expected to be nominal.

Kernel	Scenario	Reps	Extension preferred (%)	Simpler preferred (%)	Key estimate (median over replicates)	90% interval coverage (%)
Relaxed rates	8x clade shift	50	82	0	clade/background rate ratio 4.5 (true 8); branch rank corr. 0.69	56
Relaxed rates	constant BM	50	0	28	rate ratio 1.00 (true 1)	98
Rate through time	mid-history pulse	50	74	0	RMSE log rate 1.01; peak within 2 bins 86%	85 (per bin)
Rate through time	constant BM	50	0	24	RMSE log rate 0.15	100 (per bin)
Varying slopes	$\sigma_v = 0.6$	50	34	0	σ_v 0.56 (true 0.6); slope corr. 0.73	92 (σ_v); 87 (species slopes)
Varying slopes	$\sigma_v = 0$	50	0	12	σ_v 0.12 (true 0)	98 (species slopes)
Phylogeny + geography	$\sigma_g = 1$	30	87	0	geography share 0.72 (true 0.69); phylogeny share 0.19 (true 0.25)	93 (geo.); 97 (phylo.)
Phylogeny + geography	$\sigma_g = 0$	30	0	40	geography share 0.03 (true 0)	—

Computational scaling. Pruning time grew with exponent 1.08 (log-likelihood) and 0.97 (all- N LOO), dense time with exponent 2.95 and 2.75 (Fig. S4). At 2,000 tips one pruning likelihood took 3 ms against 750 ms dense, 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 with the reference BLAS; absolute values will differ, but the scaling should not. In the Stan implementation, one evaluation of the log density and its gradient (the operation that optimisation and HMC repeat) took 0.15 ms with the pruning likelihood at 100 tips and 2.9 ms at 5,000 tips (empirical exponent 1.36 between 1,000 and 5,000 tips; the algorithm is linear in N , and the measured times also include memory-access costs that grow with problem size), against 2.4

ms and 923 ms for the dense kernel at 100 and 1,000 tips (exponent 2.61). Complete fits add a start-up and input/output cost of about a second, which dominates for small trees. For full HMC sampling of the relaxed-rate kernel, which has $2N - 2$ latent log-rate increments, the cost per gradient evaluation grew from 0.07 ms at 100 tips to 0.89 ms at 1,000 tips (empirical exponent 1.10), and a complete run of four chains took 3, 11 and 40 s at 100, 300 and 1,000 tips, without divergent transitions. The number of gradient evaluations per run did not grow systematically with N in this range, but HMC cost in high dimensions generally grows faster than linearly, so these numbers support linear cost *per gradient*, not for whole analyses of very large trees. Timings were measured on a shared machine with other jobs running, so they are indicative only.

Fossil tips and identifiability. Without fossil tips, the maximised log-likelihoods of the OU and ACDC models 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 the difference was informative (Fig. S5): $\log L_{\text{OU}} - \log L_{\text{ACDC}}$ had median 3.4 (90% interval 0.0 to 9.8) when OU generated the data, with absolute difference above 2 in 74% of replicates, and median -16.5 when ACDC did (100% above 2 in absolute value). The squared-exponential kernel on a fossil tree with 60 length-scales examined failed to be numerically positive semi-definite at 73% of them. On the 5-tip counterexample of Appendix S1 the smallest eigenvalues were -0.038 (squared exponential), -0.006 (Matérn-3/2), -0.016 (Matérn-5/2) and 0.031 (exponential). A two-optimum OU model has the same kernel as the single-optimum model and a regime-dependent mean; its analytic mean and covariance matched 40,000 simulated tip vectors (maximum deviation 0.0056 for the mean, Monte Carlo s.e. up to 0.0030; 0.0043 for the covariance; Appendix S2).

Table 3: Additional simulations: weaker effects, a larger tree and misspecified data. Columns as in Table 2. Rank corr.: Spearman correlation between estimated and true branch rates (shift and OU scenarios only).

	Extension		Simpler	Rank corr.	Sampler
Scenario	Reps preferred (%)		preferred (%)	(GP)	time (s)
Relaxed rates, 3-fold shift	30	30	0	0.67	3

Scenario	Extension		Simpler	Rank corr.	Sampler
	Reps	preferred (%)	preferred (%)	(GP)	time (s)
Relaxed rates, 300 tips	15	100	0	0.52	15
Relaxed rates, OU data (misspecified)	30	0	23	0.16	3
Relaxed rates, jump data (misspecified)	30	13	0	—	4
Rate through time, weak pulse	30	10	0	—	—
Varying slopes, $\sigma_v = 0.3$	30	10	0	—	—
Phylogeny + geography, $\sigma_g = 0.5$	20	60	0	—	—

Additional simulations (Table 3). With weaker effects, power fell sharply: the relaxed-rate kernel detected a 3-fold shift in 30% of replicates (median clade/background ratio 2.0, interval coverage of the true ratio 57%), the rate-through-time kernel a pulse of amplitude 1 in 10% (per-bin coverage 0.91), the varying-slope kernel $\sigma_v = 0.3$ in 10% (median estimate 0.32, slope coverage 0.87) 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 (100%), with a median per-fit sampling time of 15 s, but it still shrank the ratio (median 4.4, coverage 20%), so the autocorrelated prior underestimated an abrupt shift at both tree sizes examined. Under misspecification the kernel was conservative. On OU data, whose implied rates rise smoothly towards the present rather than differing among lineages, it was never preferred over BM (0%; BM preferred in 23%) and its branch rates were only weakly correlated with the true ones (median Spearman 0.16): a lineage-autocorrelated model is not a detector of rate trends through time. With three large jumps it was preferred in 13% of replicates. The largest Pareto- $\hat{k}$ exceeded 0.7 in most relaxed-rate fits here too, so the detection rates are approximate.

Empirical tree shapes. 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 8-fold shift in 77% of replicates and was preferred under the null in 1 of 30; it again shrank the shift (median recovered ratio 3.7, true 7.9) while locating it (median branch rank correlation 0.69). The varying-slope kernel detected $\sigma_v = 0.6$ in 20% of replicates (0 of 30 under the null), with σ_v estimated at 0.54 and 88% coverage of species slopes. For the model parameters, 115 of 120 replicates had $\hat{R} \leq 1.01$ in both fitted models (largest 1.014). Divergent transitions occurred in 65% of null replicates, where the variability parameter lies at its boundary of zero, and in 10% of effect replicates.

Comparison with multirateBM. multirateBM finished in 20 of 20 replicates. The two methods recovered branch rates equally well (median Spearman correlation with the true rates 0.69 for the relaxed-rate kernel and 0.69 for multirateBM) and shrank the shift similarly (median recovered ratio 2.7 and 2.8; true 7.8), as expected for two estimators of autocorrelated rates. The kernel, fitted by HMC with four chains, took a median of 8 s against 480 s for the penalised-likelihood fit of multirateBM, and in addition gives posterior uncertainty and leave-one-out densities. Timings come from one machine running other jobs and are indicative only.

Mammalian brain size

These results illustrate how the kernels and the checks around them are used; they are not a substantive analysis of brain evolution (see Methods). Of 1,552 species with brain and body masses, 1,349 matched the phylogeny (tree height 179 Myr). For all seven classical kernels, the GP and phylolm maximised log-likelihoods agreed to within $9.5\text{e-}07$ units (Table 4). Among the classical models, Pagel’s λ had the lowest AIC. The estimated allometric slope ranged from 0.49 to 0.54 depending on the kernel. Pagel’s δ reached the upper bound of its range ($\delta = 3.00$), which concentrates change towards the present, and improved on BM by 20.0 log-likelihood units, although it fit far worse than λ . The early-burst rate stopped at $r = 0.00$, where EB reduces to BM, because phylolm only allows $r \leq 0$ and so cannot represent an accelerating rate. That acceleration

appears instead in the OU fit. By the equivalence above (Appendix S1, Proposition 4), the fixed-root OU estimate ($\alpha = 2.88$) is the same model as an ACDC model with $r = 2\alpha = 5.76$, and the two give the same maximised log-likelihood on this ultrametric tree (843.856 and 843.856; BM 809.417). In these data, then, stabilising selection towards one optimum and rates of brain-size evolution increasing towards the present cannot be told apart, and both fit much worse than λ ($\lambda = 0.956$), whose small tip-level component (about 4% of the variance at the tips) improved the fit by far more than any change to the evolutionary process among the classical kernels.

Table 4: Brain–body allometry in mammals: classical kernels fitted by maximum likelihood with `phylolm` and with the GP implementation (tree rescaled to unit height).

Model	slope (phylolm)	slope (GP)	θ (phylolm)	θ (GP)	log L (phylolm)	log L (GP)	Δ AIC
BM	0.4901	0.4901	—	—	809.417	809.417	441.9
Pagel’s λ	0.5437	0.5436	0.9558	0.9558	1031.371	1031.371	0.0
Pagel’s κ	0.5324	0.5324	0.3532	0.3532	993.548	993.548	75.6
Pagel’s δ	0.5004	0.5004	3	3	829.424	829.424	403.9
Early burst	0.4901	0.4901	0	0	809.417	809.417	443.9
OU, fixed root	0.5217	0.5217	2.881	2.881	843.856	843.856	375.0
OU, random root	0.5217	0.5217	2.887	2.887	843.826	843.826	375.1

Comparing kernels with simpler alternatives. All 13 Bayesian fits converged (largest $\hat{R}$ for model parameters 1.004, no divergent transitions), and every species with Pareto- $\hat{k} > 0.7$ (0–19 per model) was refitted exactly. Fig. 6a ranks the models by leave-one-out predictive density. Three results stand out. First, the residual model mattered most. Student- t residuals improved on Gaussian residuals by 147.5 ± 55.6 elpd for the BM kernel; the species with the largest variance multipliers were *Lissodelphis borealis*, *Balaena mysticetus*, *Homo sapiens*, *Bos taurus*. Residual

variance growing with body size added much less (31.2 ± 36.4), and nothing once slopes varied (-0.8 ± 5.1). Second, the brain–body relationship was curved: the quadratic coefficient was -0.019 (90% CrI -0.023 to -0.014), concave on the log scale as reported by Venditti et al. (2024) (Fig. 6b), and the quadratic mean improved prediction (12.6 ± 9.8 with t residuals). Third, once heavy tails and curvature were modelled, phylogenetically varying slopes (23.1 ± 10.4) and relaxed rates (38.1 ± 19.6) still predicted left-out species better than the curved BM model with t residuals, by about two standard errors, but could not be distinguished from each other (14.9 ± 18.1). The remaining slope variation ($\sigma_v = 0.076$, 90% CrI 0.060 – 0.093) was smaller than without the controls (0.107). A comparison against BM with Gaussian residuals alone would have credited the varying-slope kernel with a gain of 79.7 ± 39.8 elpd, most of which came from heavy tails and curvature rather than from evolving slopes.

Leave-clade-out and phylogenetic uncertainty. Leaving out whole orders changed the picture for varying slopes: the curved varying-slope model predicted left-out orders worse than the curved BM model with t residuals (-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 relative to the curved t model), 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 trees from the posterior of Upham et al. (2019), the pooled leave-one-out ranking of the four models was the same as on the MCC tree (quadratic mean, relaxed rates 0; quadratic mean, varying slope + t -12.4 ; quadratic mean, BM + t residual -36.6 ; BM + residual -200.0), and the estimates varied little among trees (slope variability 0.069 – 0.087 ; quadratic coefficient -0.020 to -0.018). All but 1 of the 200 tree fits had $\hat{R} \leq 1.01$ (largest 1.012).

Where slopes and rates differ. Fig. 7 maps the posterior slopes and relative rates from the two curved models. Among orders with at least 25 species that form a clade, mean species slopes ranged from 0.54 (Rodentia) to 0.64 (Cetacea), and relative rates from $0.27\times$ (Chiroptera) to $8.0\times$ (Cetacea) the median branch rate. These maps show what the kernels estimate; whether the patterns

reflect evolutionary processes, rather than data quality, taxonomic matching or remaining model inadequacy, would need further analyses.

Phylogeny and geography. Adding the geographic kernel to the phylogenetic and residual components barely changed the fit. The ML estimate attributed 0.3% of the variance in relative brain size to geography (length-scale $\rho = 2,790$ km), against 95% to phylogeny. The maximised log-likelihood rose by 2.13 units for two extra parameters ($\Delta\text{AIC} = 0.27$ in favour of the geographic kernel), and the leave-one-out gain at the ML estimates was 5.5 ± 2.9 , less than two standard errors. The allometric slope was unchanged (0.54 vs. 0.54). A range midpoint is not a meaningful location for marine species, so I repeated the fit without cetaceans, sirenians and pinnipeds (74 species). The geographic share then rose to 0.7% ($\rho = 1,360$ km), ΔAIC was 7.9 in favour of the geographic kernel, and the leave-one-out gain was 11.0 ± 4.5 , more than two standard errors. Once body size and phylogeny are accounted for, relative brain size therefore shows at most weak geographic structure at this scale: whether it is detectable depends on which species are included, and its share of the variance is small either way. Range midpoints are a crude summary of where species live, and a firmer answer would need better spatial data.

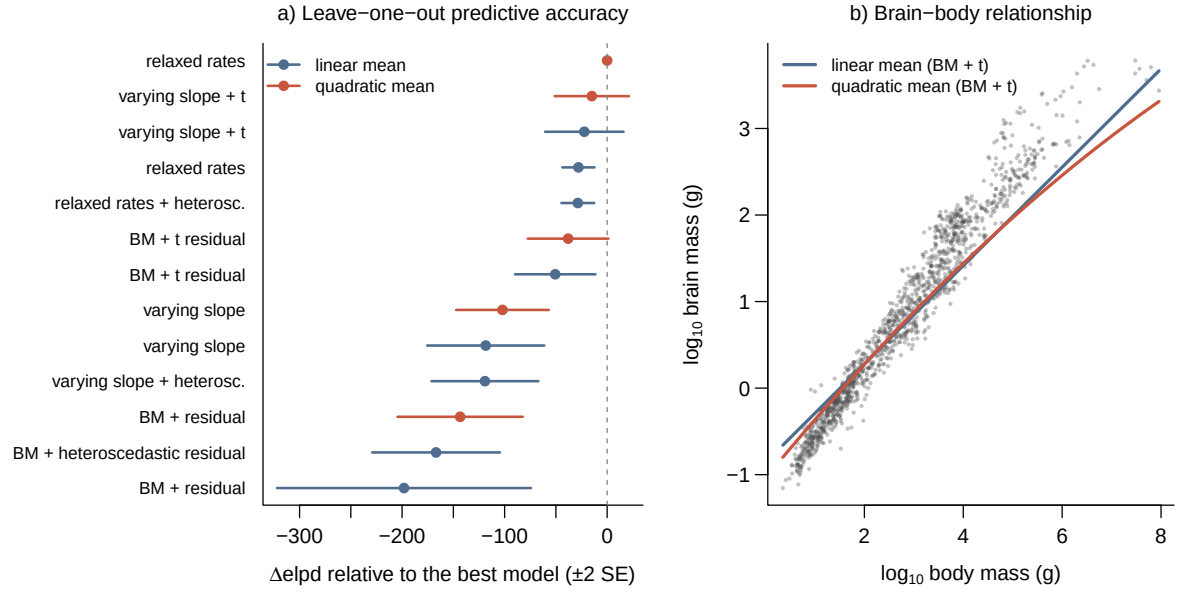


Figure 6. Mammalian brain size: kernels compared with simpler alternative explanations. (a) Expected log predictive density relative to the best model (exact leave-one-out: every species with $\text{Pareto-}\hat{k} > 0.7$ refitted; bars ± 2 SE) for models crossing the process kernel (BM, phylogenetically varying slope, relaxed rates), the residual model (Gaussian, heteroscedastic, Student- t) 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.

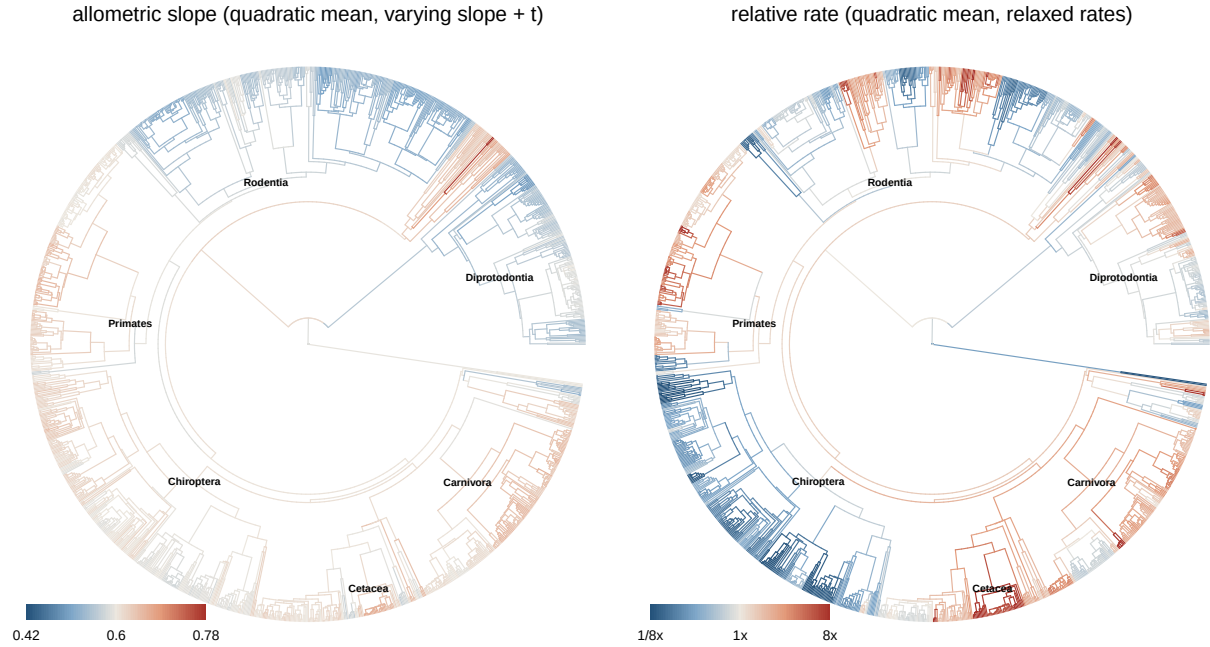


Figure 7. Posterior mean allometric slope (left; quadratic mean, varying slope + t residuals) and posterior median relative rate of brain-size evolution (right; quadratic mean, relaxed rates) along the mammal tree. Labels mark orders with at least 25 species that form a clade in this tree. The maps illustrate what the kernels estimate, not conclusions about mammalian brain evolution.

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. What a systematic treatment adds is a single notation, explicit conditions under which each representation holds, and 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 parameter 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, but it does not show that the models themselves are well behaved.

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, which is the phylogenetic mixed model in a different parameterisation (Housworth et al. 2004). Similarly, ancestral-state reconstruction under a process defined on the whole tree (e.g. BM, OU or early burst) is kriging at internal points. For OU, this kriging has to use the OU covariance between node and tip states, $\frac{\sigma^2}{2\alpha} e^{-\alpha d_{ab}} (1 - e^{-2\alpha t_{ab}})$ for a fixed root, evaluated at the heights of the internal nodes. The ultrametric edge variances of Table 1 cannot be used for this purpose, because they only reproduce the tip distribution and would give ACDC reconstructions instead. In the same way, on an ultrametric tree a single-optimum OU model is the same Gaussian model as an accelerating-rate model with $r = 2\alpha$, exactly for a fixed root and under REML for a stationary root. This equivalence has been noted before (Ho and Ané 2014b; Uyeda et al. 2015; Murrell 2018), but the edge representation turns it into a one-line consequence of Eq. (4) 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. Within-species variation does not help either, because it adds variance at the tips without changing the among-species covariance that both models share.

Second, the kernel framework is generative. 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. Here, I described four of these kernels: autocorrelated rates, a smooth rate-through-time function, phylogenetic plus spatial structure, and phylogenetically varying slopes. However, my simulations only characterise their behaviour within the scenarios I simulated, namely pure-birth trees of 80–300 tips, random 100-species 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 in the null scenarios, 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. For the relaxed-rate comparisons, and to a lesser extent the varying-slope comparisons, the largest Pareto- $\hat{k}$ exceeded 0.7 in a large share of replicates, so the detection and false-positive rates reported for these kernels (Table 2) should be regarded as approximate. Nonetheless, exact refitting of the relaxed-rate comparisons (Results) left the conclusions unchanged, and the same check (or moment matching) is cheap enough to be routine in other data sets. 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, 98% for species slopes), so the intervals are conservative when the process is absent. 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: an autocorrelated-rate prior smooths an abrupt change, so it locates shifts well (median branch rank correlation 0.69) but underestimates their size. The varying-slope kernel detected $\sigma_v = 0.6$ in only 34% of 100-tip data sets, although σ_v itself was estimated without obvious bias. Finally, the two-standard-error rule never selected an extended kernel under a null scenario, but it selected the simpler, true model in only 12–40% of null replicates and was otherwise undecided. It therefore behaves as a conservative test of the extended kernel rather than a symmetric model-selection rule. Standard errors of elpd differences are themselves unreliable for small samples and similar models (Sivula et al. 2020), which argues for reporting the full distribution of differences rather than a single threshold.

The varying-slope kernel provides a model-based treatment of the long-standing “taxon-level problem” in allometry (Lande 1979; Pagel and Harvey 1989). Instead of fitting

separate regressions within arbitrarily chosen taxonomic ranks, it allows the slope itself to evolve along the tree and estimates how much it varies. Phylogenetic random-slope models with the same marginal covariance can also be fitted in mixed-model software (Hadfield and 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 and variable-rates approaches (Eastman et al. 2011; Venditti et al. 2011), to penalised-likelihood branch rates (Revell 2021) and to relaxed random-walk models fitted in a Bayesian phylogenetic framework (Gill et al. 2017). Its closest precedent is the relaxed random walk of Fisher et al. (2021), which also samples branch-specific rate multipliers by HMC with linear-time gradients. The main differences are the autocorrelated (Brownian) prior on log rates, and the fact that the relaxed-rate kernel is just one component that can be combined with others in the same model. Its cost per gradient evaluation grows roughly linearly with the number of species and, on subtrees of the mammal tree, it recovered branch rates as well as multirateBM while running much faster (Results). Nevertheless, a comparison with reversible-jump methods and with Bayesian relaxed-rate models in BEAST (Suchard et al. 2018; Fisher et al. 2021) and RevBayes (Höhna et al. 2016) is still needed. Other extensions follow the same recipe: Lévy-type jumps (Landis et al. 2013) as heavy-tailed priors on branch increments; OU shift models (Butler and King 2004; Uyeda and Harmon 2014; Bastide et al. 2018) as regime-specific mean functions (for models with shared α and σ^2 , such as OUM, the OU kernel is unchanged and only the mean differs, as I verified by simulation in Appendix S2, whereas models with regime-specific σ^2 or α , such as OUMV and OUMA, change the kernel as well); multivariate and factor models (Hassler et al. 2022) as structured coregionalisation; and phylogenetic uncertainty (Villemereuil et al. 2012) by marginalising over a posterior sample of trees. Many other kernels could be envisaged for different macroevolutionary scenarios, for example rates that depend on a discrete character or on an environmental time series (e.g. temperature through time), kernels that combine phylogenetic with ecological or trophic similarity, or covariance structures for reticulate evolution along phylogenetic networks.

Third, the GP framework comes with general tools for inference and model checking.

774 Leave-one-out cross-validation (Sundararajan and Keerthi 2001; Bürkner et al. 2021) compares
775 models with different numbers and kinds of parameters based on their predictive performance,
776 without relying on the asymptotic justification of AIC, which is unreliable near parameter
777 boundaries. For Gaussian models, the conditional LOO densities have a closed form and, for tree
778 kernels, they can be computed in linear time by two-pass message passing. When combined with
779 importance sampling over posterior draws, this approximates full LOO-CV without refitting. This
780 approximation can fail when some observations are highly influential, so the Pareto- $\hat{k}$ diagnostics
781 should always be checked. Posterior predictive kriging also gives ancestral states and imputations
782 that carry parameter uncertainty with them. The phylogp package provides these tools for any
783 kernel, together with exact refits for influential species, leave-clade-out cross-validation,
784 linear-time predictions at tips and internal nodes, fits across a sample of trees, and a numerical
785 check of whether two kernels imply the same distribution for the tips.

786 The mammal example illustrates why these checks matter. When compared with BM and
787 Gaussian residuals, both varying slopes and relaxed rates appeared to improve prediction
788 considerably. However, most of that improvement came from heavy-tailed residuals and a curved
789 brain–body relationship, and leaving out whole orders reversed the advantage of varying slopes.
790 Therefore, kernel extensions should be compared with simpler explanations that leave the same
791 statistical signature, and assessed with more than one form of cross-validation, before their
792 predictive gains are interpreted as evidence for a particular process. Whilst this example does not
793 provide definitive answers about brain evolution, it does show how the workflow can be used to
794 question an apparently strong result. These results are specific to this data set and set of models,
795 and further analyses are required to establish the biological meaning of the signal that remains.

796 The GP view is complementary to existing general frameworks for Gaussian models on
797 trees. The GLInv family of Mitov et al. (2020), the models of Bastide et al. (2021) and the unifying
798 framework of Manceau et al. (2017) are organised around the generating process (i.e. how a trait
799 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. Other kernels, such as the additive spatial kernel or the varying-slope product kernel, are easy to state in the GP language but awkward to express as branch-wise processes, although both can be fitted as mixed models in general software (Freckleton and Jetz 2009; Hadfield and Nakagawa 2010).

What, then, does the GP view add to existing software? Specialised packages such as `phylolm` and `geiger` estimate the parameters of the classical kernels (α , κ , δ , the EB rate) internally, and mixed-model software such as `brms` and `MCMCglmm` fits phylogenetic random intercepts and random slopes, and with them the varying-slope kernel, from a fixed covariance matrix. In addition, several of the pieces described here also exist elsewhere: `phylolm` uses a linear-time algorithm (Ho and Ané 2014a), `MCMCglmm` exploits the sparse inverse of the relatedness matrix including ancestral nodes (Hadfield and Nakagawa 2010), `RPANDA` fits rates that are arbitrary functions of time or of an environmental variable (Clavel and Morlon 2017), `BEAST` fits relaxed random walks with linear-time gradients (Suchard et al. 2018; Fisher et al. 2021), and `RevBayes` lets users specify custom models of continuous-trait evolution, including relaxed Brownian motion, as probabilistic graphical models (Höhna et al. 2016). What has been missing until now is a framework organised around the covariance function, in which (i) kernels can be combined freely (i.e. sums, products and hierarchical priors) and fitted with the same code; (ii) a kernel can be checked for validity on the tree at hand (which, as shown here, can fail for distance-based kernels on trees with fossils), 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 `phyllogp` package provides these, and treats kernel parameters like any other parameter in the model. That said, the specialised packages are still the better choice for models whose likelihood is not a Gaussian process 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 are special cases of a common framework,

not that the existing software should be replaced.

Kernel validity deserves particular attention. 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 of distance remain valid. This is particularly relevant for total-evidence analyses that combine extinct and extant species, which is precisely where GPs on trees are most attractive. In these analyses, kernels should be built from Eq. (4) 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/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 (4) the $O(N)$ pruning recursion avoids this (Felsenstein 1973; Ho and Ané 2014a), and latent tree processes can be sampled at nodes with linear cost per gradient. Kernels that break the Markov structure of the tree, such as spatial kernels, require dense algebra or approximations (sparse inducing points, nearest-neighbour GPs, or SPDE/GMRF methods; Lindgren et al. (2011); Rue et al. (2009); Matsuba et al. (2024)).

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. One interesting next step would be to use the same approach to build kernels for some of the processes mentioned above, such as rates that depend on a discrete character or trait evolution along phylogenetic networks.

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 (the reanalysis of the mammal data and the additional simulations), 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. All code, data and results are publicly archived (see Data availability) so that every result can be independently reproduced.

Supplementary Material

Code and data: see Data availability. Appendix S1 (proofs), Appendix S2 (multi-optimum Ornstein–Uhlenbeck models), Appendix S3 (comparison with R packages, Table S1; rate-through-time sampler diagnostics and prior sensitivity; species removed in the mammal sensitivity analyses) and Figures S1–S5 are provided as online supplementary documents.

Data availability

The methods are implemented in the R package `phylogp` (<https://github.com/ThomasPueschel/phylogp>; version 0.2.0 used here archived at Zenodo, doi:[10.5281/zenodo.2347207](https://doi.org/10.5281/zenodo.2347207)), which includes tests against `phylolm`, `phytools` and dense Gaussian computations, and tutorials. All analysis scripts, simulated data and results underlying every figure and table are archived at Zenodo (doi:[10.5281/zenodo.2347207](https://doi.org/10.5281/zenodo.2347207)). The empirical data come from published sources: brain and body masses from Burger et al. (2019a) (Dryad, doi:[10.5061/dryad.2r62k7s](https://doi.org/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).

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.
- Blomberg S.P., Garland T., Ives A.R. 2003. Testing for phylogenetic signal in comparative data:

892 Behavioral traits are more labile. *Evolution*. 57:717–745.

893 Blomberg S.P., Lefevre J.G., Wells J.A., Waterhouse M. 2012. [Independent contrasts and PGLS](#)
894 [regression estimators are equivalent](#). *Systematic Biology*. 61:382–391.

895 Borovitskiy V., Azangulov I., Terenin A., Mostowsky P., Deisenroth M., Durrande N. 2021.
896 Matérn Gaussian processes on graphs. *Proceedings of the 24th international conference on*
897 *artificial intelligence and statistics*. 130:2593–2601.

898 Brahmantio B., Bartoszek K., Yapar E. 2026. [Bayesian inference of mixed Gaussian phylogenetic](#)
899 [models](#). *BMC Bioinformatics*. 27:77.

900 Burger J.R., George M.A., Leadbetter C., Shaikh F. 2019a. The allometry of brain size in mammals.
901 *Journal of Mammalogy*. 100:276–283.

902 Burger J.R., George M.A., Leadbetter C., Shaikh F. 2019b. [Data from: The allometry of brain size](#)
903 [in mammals](#).

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

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

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

- 910 Carpenter B., Gelman A., Hoffman M.D., Lee D., Goodrich B., Betancourt M., Brubaker M., Guo
911 J., Li P., Riddell A. 2017. Stan: A probabilistic programming language. *Journal of Statistical*
912 *Software*. 76:1–32.
- 913 Clavel J., Escarguel G., Merceron G. 2015. mvMORPH: An R package for fitting multivariate
914 evolutionary models to morphometric data. *Methods in Ecology and Evolution*. 6:1311–1319.
- 915 Clavel J., Morlon H. 2017. [Accelerated body size evolution during cold climatic periods in the](#)
916 [Cenozoic](#). *Proceedings of the National Academy of Sciences of the USA*. 114:4183–4188.
- 917 Cooper N., Thomas G.H., Venditti C., Meade A., Freckleton R.P. 2016. A cautionary note on the
918 use of Ornstein Uhlenbeck models in macroevolutionary studies. *Biological Journal of the*
919 *Linnean Society*. 118:64–77.
- 920 Cressie N.A.C. 1993. *Statistics for spatial data*. New York: Wiley.
- 921 Dellacherie C., Martínez S., San Martín J. 2014. [Inverse M-matrices and ultrametric matrices](#).
922 Cham: Springer.
- 923 Deza M.M., Laurent M. 1997. *Geometry of cuts and metrics*. Berlin: Springer.
- 924 Diggle P.J., Ribeiro P.J. Jr. 2007. *Model-based geostatistics*. New York: Springer.
- 925 Duvenaud D., Lloyd J., Grosse R., Tenenbaum J., Ghahramani Z. 2013. Structure discovery in
926 nonparametric regression through compositional kernel search. *Proceedings of the 30th*
927 *international conference on machine learning*. 28:1166–1174.

- 928 Eastman J.M., Alfaro M.E., Joyce P., Hipp A.L., Harmon L.J. 2011. A novel comparative method
929 for identifying shifts in the rate of character evolution on trees. *Evolution*. 65:3578–3589.
- 930 Felsenstein J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous
931 characters. *American Journal of Human Genetics*. 25:471–492.
- 932 Felsenstein J. 1985. Phylogenies and the comparative method. *The American Naturalist*. 125:1–15.
- 933 Felsenstein J. 2005. Using the quantitative genetic threshold model for inferences between and
934 within species. *Philosophical Transactions of the Royal Society B*. 360:1427–1434.
- 935 Feragen A., Lauze F., Hauberg S. 2015. Geodesic exponential kernels: When curvature and
936 linearity conflict. *Proceedings of the IEEE conference on computer vision and pattern
937 recognition (CVPR)*.:3032–3042.
- 938 Fisher A.A., Ji X., Zhang Z., Lemey P., Suchard M.A. 2021. [Relaxed random walks at scale](#).
939 *Systematic Biology*. 70:258–267.
- 940 Freckleton R.P. 2012. Fast likelihood calculations for comparative analyses. *Methods in Ecology
941 and Evolution*. 3:940–947.
- 942 Freckleton R.P., Jetz W. 2009. Space versus phylogeny: Disentangling phylogenetic and spatial
943 signals in comparative data. *Proceedings of the Royal Society B*. 276:21–30.
- 944 Gabry J., Češnovar R., Johnson A., Bröder S. 2024. Cmdstanr: R interface to CmdStan.
- 945 Garland T. Jr., Ives A.R. 2000. [Using the past to predict the present: Confidence intervals for](#)

946 [regression equations in phylogenetic comparative methods](#). *The American Naturalist*.
 947 155:346–364.

948 Gill M.S., Ho L.S.T., Baele G., Lemey P., Suchard M.A. 2017. [A relaxed directional random walk](#)
 949 [model for phylogenetic trait evolution](#). *Systematic Biology*. 66:299–319.

950 Gneiting T. 2013. [Strictly and non-strictly positive definite functions on spheres](#). *Bernoulli*.
 951 19:1327–1349.

952 Grafen A. 1989. The phylogenetic regression. *Philosophical Transactions of the Royal Society of*
 953 *London B*. 326:119–157.

954 Guzmán Naranjo M., Becker L. 2022. [Statistical bias control in typology](#). *Linguistic Typology*.
 955 26:605–670.

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

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

961 Hansen T.F. 1997. Stabilizing selection and the comparative analysis of adaptation. *Evolution*.
 962 51:1341–1351.

963 Hansen T.F., Martins E.P. 1996. Translating between microevolutionary process and
 964 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., others. 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.

Hassler G.W., Gallone B., Aristide L., Allen W.L., Tolkoﬀ M.R., Holbrook A.J., others. 2022.

Principled, practical, flexible, fast: A new approach to phylogenetic factor analysis. *Methods in*

Ecology and Evolution. 13:2181–2197.

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

models. *Systematic Biology*. 63:397–408.

Ho L. si T., Ané C. 2014b. Intrinsic inference diﬃculties for trait evolution with

Ornstein-Uhlenbeck models. *Methods in Ecology and Evolution*. 5:1133–1146.

Hoffmann T. 2023. [gptoolsStan: Gaussian processes on graphs and lattices in 'stan'](#).

Höhna S., Landis M.J., Heath T.A., Boussau B., Lartillot N., Moore B.R., Huelsenbeck J.P.,

Ronquist F. 2016. [RevBayes: Bayesian phylogenetic inference using graphical models and an](#)

[interactive model-specification language](#). *Systematic Biology*. 65:726–736.

Housworth E.A., Martins E.P., Lynch M. 2004. The phylogenetic mixed model. *The American*

Naturalist. 163:84–96.

- 984 Ives A.R., Helmus M.R. 2011. Generalized linear mixed models for phylogenetic analyses of
985 community structure. *Ecological Monographs*. 81:511–525.
- 986 Ives A.R., Midford P.E., Garland T. 2007. Within-species variation and measurement error in
987 phylogenetic comparative methods. *Systematic Biology*. 56:252–270.
- 988 Jhwueng D.-C. 2025. [A literature review of stochastic modeling for phylogenetic comparative](#)
989 [analysis in trait evolution](#). *Mathematics*. 13:361.
- 990 Jones K.E., Bielby J., Cardillo M., Fritz S.A., O'Dell J., Orme C.D.L., Safi K., Sechrest W., Boakes
991 E.H., Carbone C., Connolly C., Cutts M.J., Foster J.K., Grenyer R., Habib M., Plaster C.A.,
992 Price S.A., Rigby E.A., Rist J., Teacher A., Bininda-Emonds O.R.P., Gittleman J.L., Mace
993 G.M., Purvis A. 2009. [PanTHERIA: A species-level database of life history, ecology, and](#)
994 [geography of extant and recently extinct mammals](#). *Ecology*. 90:2648.
- 995 Jones N.S., Moriarty J. 2013. Evolutionary inference for function-valued traits: Gaussian process
996 regression on phylogenies. *Journal of the Royal Society Interface*. 10:20120616.
- 997 Lande R. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain:body size
998 allometry. *Evolution*. 33:402–416.
- 999 Landis M.J., Schraiber J.G., Liang M. 2013. Phylogenetic analysis using Lévy processes: Finding
1000 jumps in the evolution of continuous traits. *Systematic Biology*. 62:193–204.
- 1001 Lindgren F., Rue H., Lindström J. 2011. An explicit link between Gaussian fields and Gaussian
1002 Markov random fields: The stochastic partial differential equation approach. *Journal of the*
1003 *Royal Statistical Society B*. 73:423–498.

- 1004 Lynch M. 1991. Methods for the analysis of comparative data in evolutionary biology. *Evolution*.
1005 45:1065–1080.
- 1006 Manceau M., Lambert A., Morlon H. 2017. [A unifying comparative phylogenetic framework](#)
1007 [including traits coevolving across interacting lineages](#). *Systematic Biology*. 66:551–568.
- 1008 Martins E.P., Hansen T.F. 1997. Phylogenies and the comparative method: A general approach to
1009 incorporating phylogenetic information into the analysis of interspecific data. *The American*
1010 *Naturalist*. 149:646–667.
- 1011 Matsuba M., Fukasawa K., Aoki S., Akasaka M., Ishihama F. 2024. [Scalable phylogenetic](#)
1012 [Gaussian process models improve the detectability of environmental signals on local extinctions](#)
1013 [for many Red List species](#). *Methods in Ecology and Evolution*. 15:756–768.
- 1014 McElreath R. 2020. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca
1015 Raton, FL: CRC Press.
- 1016 Meagher J.P., Damoulas T., Jones K.E., Girolami M. 2018. [Phylogenetic Gaussian processes for](#)
1017 [bat echolocation](#). *Statistical data science*. London: World Scientific. p. 111–124.
- 1018 Mitov V., Bartoszek K., Asimomitis G., Stadler T. 2020. Fast likelihood calculation for multivariate
1019 Gaussian phylogenetic models with shifts. *Theoretical Population Biology*. 131:66–78.
- 1020 Murrell D.J. 2018. [A global envelope test to detect non-random bursts of trait evolution](#). *Methods*
1021 *in Ecology and Evolution*. 9:1739–1748.
- 1022 O’Meara B.C., Ané C., Sanderson M.J., Wainwright P.C. 2006. Testing for different rates of

1023 continuous trait evolution using likelihood. *Evolution*. 60:922–933.

1024 Orme D., Freckleton R., Thomas G., Petzoldt T., Fritz S., Isaac N., Pearse W. 2023. Caper:
1025 Comparative analyses of phylogenetics and evolution in R.

1026 Pagel M. 1999. Inferring the historical patterns of biological evolution. *Nature*. 401:877–884.

1027 Pagel M.D., Harvey P.H. 1989. Taxonomic differences in the scaling of brain on body weight
1028 among mammals. *Science*. 244:1589–1593.

1029 Paradis E., Schliep K. 2019. Ape 5.0: An environment for modern phylogenetics and evolutionary
1030 analyses in R. *Bioinformatics*. 35:526–528.

1031 Pennell M.W., Eastman J.M., Slater G.J., Brown J.W., Uyeda J.C., FitzJohn R.G., Alfaro M.E.,
1032 Harmon L.J. 2014. Geiger v2.0: An expanded suite of methods for fitting macroevolutionary
1033 models to phylogenetic trees. *Bioinformatics*. 30:2216–2218.

1034 Pinheiro J., Bates D., R Core Team. 2023. Nlme: Linear and nonlinear mixed effects models.

1035 Rasmussen C.E., Williams C.K.I. 2006. Gaussian processes for machine learning. Cambridge, MA:
1036 MIT Press.

1037 Revell L.J. 2012. Phytools: An R package for phylogenetic comparative biology (and other things).
1038 *Methods in Ecology and Evolution*. 3:217–223.

1039 Revell L.J. 2021. A variable-rate quantitative trait evolution model using penalized-likelihood.
1040 *PeerJ*. 9:e11997.

- 1041 Rue H., Martino S., Chopin N. 2009. Approximate Bayesian inference for latent Gaussian models
1042 by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society B*.
1043 71:319–392.
- 1044 Schluter D., Price T., Mooers A.Ø., Ludwig D. 1997. [Likelihood of ancestor states in adaptive](#)
1045 [radiation](#). *Evolution*. 51:1699–1711.
- 1046 Schoenberg I.J. 1938. Metric spaces and positive definite functions. *Transactions of the American*
1047 *Mathematical Society*. 44:522–536.
- 1048 Sivula T., Magnusson M., Matamoros A.A., Vehtari A. 2020. Uncertainty in Bayesian
1049 leave-one-out cross-validation based model comparison.
- 1050 Smaers J.B., Rothman R.S., Hudson D.R., Balanoff A.M., Beatty B., Dechmann D.K.N., others.
1051 2021. The evolution of mammalian brain size. *Science Advances*. 7:eabe2101.
- 1052 Suchard M.A., Lemey P., Baele G., Ayres D.L., Drummond A.J., Rambaut A. 2018. [Bayesian](#)
1053 [phylogenetic and phylodynamic data integration using BEAST 1.10](#). *Virus Evolution*.
1054 4:vey016.
- 1055 Sundararajan S., Keerthi S.S. 2001. Predictive approaches for choosing hyperparameters in
1056 Gaussian processes. *Neural Computation*. 13:1103–1118.
- 1057 Thorne J.L., Kishino H. 2002. Divergence time and evolutionary rate estimation with multilocus
1058 data. *Systematic Biology*. 51:689–702.
- 1059 Upham N.S., Esselstyn J.A., Jetz W. 2019. Inferring the mammal tree: Species-level sets of

- 1060 phylogenies for questions in ecology, evolution, and conservation. PLoS Biology.
1061 17:e3000494.
- 1062 Uyeda J.C., Caetano D.S., Pennell M.W. 2015. [Comparative analysis of principal components can](#)
1063 [be misleading](#). Systematic Biology. 64:677–689.
- 1064 Uyeda J.C., Harmon L.J. 2014. A novel Bayesian method for inferring and interpreting the
1065 dynamics of adaptive landscapes from phylogenetic comparative data. Systematic Biology.
1066 63:902–918.
- 1067 Uyeda J.C., Zenil-Ferguson R., Pennell M.W. 2018. Rethinking phylogenetic comparative methods.
1068 Systematic Biology. 67:1091–1109.
- 1069 Vehtari A., Gelman A., Gabry J. 2017. Practical Bayesian model evaluation using leave-one-out
1070 cross-validation and WAIC. Statistics and Computing. 27:1413–1432.
- 1071 Venditti C., Baker J., Barton R.A. 2024. [Co-evolutionary dynamics of mammalian brain and body](#)
1072 [size](#). Nature Ecology & Evolution. 8:1534–1542.
- 1073 Venditti C., Meade A., Pagel M. 2011. Multiple routes to mammalian diversity. Nature.
1074 479:393–396.
- 1075 Villemereuil P. de, Wells J.A., Edwards R.D., Blomberg S.P. 2012. Bayesian models for
1076 comparative analysis integrating phylogenetic uncertainty. BMC Evolutionary Biology. 12:102.