nature

https://doi.org/10.1038/s41586-020-2176-1

Supplementary information

Extant timetrees are consistent with a myriad of diversification histories

In the format provided by the authors and unedited

Stilianos Louca ✉ & Matthew W. Pennell ✉

Extant timetrees are consistent with a myriad of diversification histories
- Supplemental Information -

Stilianos Louca[1,2] & Matthew W. Pennell[3,4]

[1]*Department of Biology, University of Oregon, USA*
[2]*Institute of Ecology and Evolution, University of Oregon, USA*
[3]*Biodiversity Research Centre, University of British Columbia, Vancouver, Canada*
[4]*Department of Zoology, University of British Columbia, Vancouver, Canada*

Contents

S.1 Mathematical derivations

In the following, we provide mathematical derivations for various claims made in the main article. Some parts can be found in previous literature (1, 2, 3, 4, 5, 6, 7), but are included here for completeness.

S.1.1 General considerations

We begin with listing some basic mathematical properties of deterministic birth-death models that will be of use at various later stages. Here, by "deterministic model" we refer to a set of differential equations describing the expected number of extant species over time as well as the expected number of lineages in the timetree over time, based on the speciation and extinction rate of the original stochastic birth-death model (e.g., as done by Kubo *et al.* (4)). For large tree sizes, the LTTs generated by the stochastic model converge to the dLTT predicted by the deterministic model. Such a deterministic model is sometimes known as the "continuum limit" of the stochastic model (8).

Our starting point is some time-dependent speciation rate λ, some time-dependent extinction rate μ and some sampling fraction ρ (fraction of extant species included in the tree). Let τ denote time before present ("age"). The deterministic total diversity, i.e. the number of species predicted at any point in time according to the deterministic model, and conditional upon M_o extant species having been sampled at present-day, is obtained by solving the following differential equation backward in time:

$$\frac{dN}{d\tau} = N \cdot (\mu - \lambda), \tag{1}$$

with initial condition $N(0) = M_o/\rho$, i.e.:

$$N(\tau) := \frac{M_o}{\rho} \exp\left[\int_0^\tau [\mu(u) - \lambda(u)]\ du\right]. \tag{2}$$

The deterministic LTT (dLTT), i.e. the number of lineages represented in the final extant timetree at any time point according to the deterministic model, is given by:

$$M(\tau) = N(\tau) \cdot (1 - E(\tau)), \tag{3}$$

where $E(\tau)$ is the fraction of lineages extant at age τ that will be missing from the timetree (either due to extinction or not having been sampled). Note that for finite trees generated by the original stochastic model E is the probability that a lineage extant at age τ will be missing from the timetree. As explained by Morlon *et al.* (5), E satisfies the differential equation:

$$\frac{dE}{d\tau} =\mu - E \cdot (\lambda + \mu) + E^2\lambda, \quad E(0) = 1 - \rho. \tag{4}$$

We mention that the solution to Eq. (4) is provided by Morlon *et al.* (5, Eq. 2). Taking the derivative of both sides in Eq. (3), and then using Eq. (4) to replace $dE/d\tau$ as well as Eq. (1) to replace $dN/d\tau$ quickly leads to the differential equation:

$$\frac{dM}{d\tau} = M\lambda \cdot (E - 1), \tag{5}$$

with initial condition $M(0) = M_o$. The solution to this differential equation is:

$$M(\tau) = M_o \cdot \exp\left[\int_0^\tau \lambda(u) \cdot [E(u) - 1]\ du\right]. \tag{6}$$

Observe that E is a property purely of the model, and does not depend on the particular tree considered nor on M_o; together with Eq. (6), this shows that any two models either have equal dLTTs for every given M_o or they have non-equal dLTTs for every given M_o. Hence, model congruency is a property of two models, regardless of the data considered.

Pulled speciation rate: Defining the relative slope of the dLTT:

$$\lambda_{\mathrm{p}} := -\frac{1}{M}\frac{dM}{d\tau} \tag{7}$$

allows us to write Eq. (5) as follows:

$$\lambda_{\mathrm{p}} = \lambda \cdot (1 - E). \tag{8}$$

We note that $P(\tau) := 1 - E(\tau)$ is the probability that a lineage extant at age τ is represented in the extant timetree. P can thus be interpreted as a generalization of the present-day sampling fraction ρ to previous times. In fact, trimming a timetree at some age $\tau_o > 0$ (i.e., omitting anything younger than τ_o) would yield a new (shorter) timetree, whose tips are a random subset of the lineages that existed at age τ_1, each included at probability $P(\tau_o)$.

As becomes clear in Eq. (8), in the absence of extinction and if $\rho = 1$, the relative slope λ_{p} becomes equal to the speciation rate λ. In the presence of extinction, λ_{p} is artificially pulled downwards relative to λ towards the past. Reciprocally, under incomplete sampling λ is artificially pulled downwards near the present. We shall therefore henceforth call λ_{p} the "pulled speciation rate". Note that $\lambda_{\mathrm{p}}(\tau)$ is the expected density of branching events in the timetree at age τ, normalized by the expected number of lineages at that age. Since for a given M_o a model's dLTT is fully determined by λ_{p} and, reciprocally, λ_{p} is fully determined by the dLTT, two models are congruent if and only if they have the same pulled speciation rate.

Pulled diversification rate: Taking the derivative on both sides of Eq. (8) and using Eq. (4) to replace $dE/d\tau$ leads to:

$$\frac{d\lambda_{\mathrm{p}}}{d\tau} = \lambda_{\mathrm{p}} \cdot \left[\frac{1}{\lambda}\frac{d\lambda}{d\tau} - \mu + \lambda E\right] = \lambda_{\mathrm{p}} \cdot \left[\frac{1}{\lambda}\frac{d\lambda}{d\tau} + \lambda - \mu - \lambda \cdot (1 - E)\right] = \lambda_{\mathrm{p}} \cdot (r_{\mathrm{p}} - \lambda_{\mathrm{p}})\,, \tag{9}$$

where we defined the "pulled diversification rate":

$$r_{\mathrm{p}} := \lambda - \mu + \frac{1}{\lambda}\frac{d\lambda}{d\tau}. \tag{10}$$

Rearranging terms in Eq. (9) yields:

$$r_{\mathrm{p}} = \lambda_{\mathrm{p}} + \frac{1}{\lambda_{\mathrm{p}}}\frac{d\lambda_{\mathrm{p}}}{d\tau}, \tag{11}$$

which shows that r_{p} can be directly calculated from the dLTT. Reciprocally, λ_{p} is completely determined by r_{p} and some initial condition (i.e., λ_{p} specified at some fixed time), since one can just solve the differential equation for $\lambda_{\rm p}$ (see solution in Supplement S.1.6). We thus conclude that two birth-death models are congruent if and only if they have the same $r_{\rm p}$ and the same $\lambda_{\rm p}$ at some time point in the present or past (for example the same product $\rho\lambda_o$).

S.1.2 The likelihood in terms of the LTT and dLTT

In the following we show how the likelihood of an extant timetree under a birth-death model can be expressed purely in terms of the tree's LTT and the model's dLTT. We note that an alternative derivation was provided by Lambert *et al.* (6, §3.2). We begin with the case where the stem age is known and the likelihood is conditioned on the survival of the stem lineage; the alternative case where only the crown age is known is very similar and will be discussed at the end.

Our starting point is the likelihood formula described by Morlon *et al.* (5):

$$L = \frac{\rho^{n+1}\Psi(\tau_1,\tau_o)}{1-E(\tau_o)}\prod_{i=1}^{n}\lambda(\tau_i)\Psi(s_{i,1},\tau_i)\Psi(s_{i,2},\tau_i), \tag{12}$$

where n is the number of branching points (internal nodes), τ_o is the age of the stem, $\tau_1 > \tau_2 > .. > \tau_n$ are the ages (time before present) of the branching points, $s_{i,1}$, $s_{i,2}$ are the ages at which the daughter lineages originating at age τ_i themselves branch (or end at a tip), ρ is the tree's sampling fraction (fraction of present-day extant species included in the tree), $E(\tau)$ is the probability that a single lineage that existed at age τ would survive to the present and be represented in the tree (5, Eq. 2 therein), Ψ is defined as:

$$\Psi(s,\tau) := e^{R(\tau)-R(s)}\left[\frac{1+\rho\displaystyle\int_0^s \lambda(u)e^{R(u)}\,du}{1+\rho\displaystyle\int_0^\tau \lambda(u)e^{R(u)}\,du}\right]^2, \tag{13}$$

and $R(\tau)$ is defined as:

$$R(\tau) := \int_0^\tau [\lambda(u)-\mu(u)]\,du. \tag{14}$$

It is straightforward to confirm that Ψ satisfies the property $\Psi(s,\tau) = \Psi(0,\tau)/\Psi(0,s)$; using this property in Eq. (12) leads to:

$$L = \frac{\rho^{n+1}}{1-E(\tau_o)}\cdot\frac{\Psi(0,\tau_o)}{\Psi(0,\tau_1)}\prod_{i=1}^{n}\frac{\lambda(\tau_i)\Psi(0,\tau_i)^2}{\Psi(0,s_{i,1})\Psi(0,s_{i,2})}. \tag{15}$$

Since each internal node except for the root is the child of another internal node, the enumerator and denominator in Eq. (15) partly cancel out, eventually leading to:

$$L = \frac{\rho^{n+1}\Psi(0,\tau_o)}{1-E(\tau_o)}\prod_{i=1}^{n}\lambda(\tau_i)\Psi(0,\tau_i). \tag{16}$$

Since the set of branching times τ_i is completely determined by the LTT (branching events correspond to jumps in the LTT), we conclude that the likelihood of a tree is entirely determined by its LTT.

Further, from Eq. (11) we know that the model's dLTT satisfies:

$$\lambda-\mu+\frac{d\ln\lambda}{d\tau} = \frac{d\ln\lambda_{\rm p}}{d\tau}-\frac{d\ln M}{d\tau}. \tag{17}$$

Integrating both sides of Eq. (17) yields:

$$R(\tau) + \ln\frac{\lambda(\tau)}{\lambda_o} = \int_0^\tau \left[\lambda - \mu + \frac{d\ln\lambda}{du}\right] du = \int_0^\tau \left[\frac{d\ln\lambda_\mathrm{p}}{du} - \frac{d\ln M}{du}\right] du = \ln\frac{\lambda_\mathrm{p}(\tau)}{\lambda_\mathrm{p}(0)} - \ln\frac{M(\tau)}{M_o}, \tag{18}$$

where M_o is the number of extant species included in the timetree. Hence:

$$e^{R(\tau)}\frac{\lambda(\tau)}{\lambda_o} = \frac{\lambda_\mathrm{p}(\tau)M_o}{\lambda_\mathrm{p}(0)M(\tau)}. \tag{19}$$

Using Eq. (19) in Eq. (13) yields:

$$\Psi(0,\tau) = \frac{\lambda_o}{\lambda(\tau)} \cdot \frac{\lambda_\mathrm{p}(\tau)M_o}{\lambda_\mathrm{p}(0)M(\tau)} \cdot \left[1 + \frac{\rho\lambda_o}{\lambda_\mathrm{p}(0)}M_o\int_0^\tau du\,\frac{\lambda_\mathrm{p}(u)}{M(u)}\right]^{-2} \tag{20}$$

Recall that $\rho\lambda_o = \lambda_\mathrm{p}(0)$ according to Eq. (8), so that Eq. (20) can be written as:

$$\Psi(0,\tau) = \frac{1}{\rho\lambda(\tau)} \cdot \frac{\lambda_\mathrm{p}(\tau)M_o}{M(\tau)} \cdot \left[1 + M_o\int_0^\tau du\,\frac{\lambda_\mathrm{p}(u)}{M(u)}\right]^{-2}. \tag{21}$$

Note that:

$$\frac{\lambda_\mathrm{p}}{M} = \frac{d}{d\tau}\frac{1}{M}. \tag{22}$$

Hence, Eq. (21) can be further simplified to:

$$\begin{aligned}\Psi(0,\tau) =& \frac{1}{\rho\lambda(\tau)} \cdot \frac{\lambda_\mathrm{p}(\tau)M_o}{M(\tau)} \cdot \left[1 + M_o\int_0^\tau du\,\frac{d}{du}\left(\frac{1}{M}\right)\right]^{-2} \\ =& \frac{1}{\rho\lambda(\tau)} \cdot \frac{\lambda_\mathrm{p}(\tau)M_o}{M(\tau)} \cdot \left[1 + M_o\left(\frac{1}{M(\tau)} - \frac{1}{M_o}\right)\right]^{-2} \\ =& \frac{\lambda_\mathrm{p}(\tau)M(\tau)}{\rho\lambda(\tau)M_o}.\end{aligned} \tag{23}$$

Inserting Eq. (23) into the likelihood formula (16) yields:

$$L = \frac{1}{[1-E(\tau_o)]\,\lambda(\tau_o)} \cdot \frac{\lambda_\mathrm{p}(\tau_o)M(\tau_o)}{M_o^{n+1}}\prod_{i=1}^n \lambda_\mathrm{p}(\tau_i)M(\tau_i). \tag{24}$$

Recall that $(1-E)\,\lambda = \lambda_\mathrm{p}$ according to Eq. (8), which when inserted into (24) yields:

$$L = \frac{M(\tau_o)}{M_o^{n+1}}\prod_{i=1}^n \lambda_\mathrm{p}(\tau_i)M(\tau_i). \tag{25}$$

Since $\lambda_\mathrm{p}M = -dM/d\tau$, Eq. (25) becomes:

$$\boxed{L = \frac{M(\tau_o)}{M_o^{n+1}}\prod_{i=1}^n \left[-\left.\frac{dM}{d\tau}\right|_{\tau_i}\right].} \tag{26}$$

A corollary of Eq. (26) is that for any given extant timetree, any two models with the same dLTT will also yield the same likelihood.

Note that the likelihood in Eq. (12) or equivalently Eq. (26) is conditioned upon the survival of the stem lineage, assuming that the stem age is known. If the stem age is unknown the likelihood should be conditioned upon the splitting at the root and the survival of the root's two daughter-lineages, as follows:

$$L_r = \frac{\rho^{n+1}}{\lambda(\tau_1)\cdot[1-E(\tau_1)]^2}\prod_{i=1}^{n}\lambda(\tau_i)\Psi(s_{i,1},\tau_i)\Psi(s_{i,2},\tau_i). \tag{27}$$

Note that Eq. (27) can be obtained from (12) by setting the stem age equal to the crown age ($\tau_o = \tau_1$) and adjusting the conditioning. Following a similar procedure as above, it is easy to show that L_r can be expressed in the following alternative forms:

$$L_r = \frac{\rho^{n+1}\Psi(0,\tau_1)}{\lambda(\tau_1)\cdot[1-E(\tau_1)]^2}\prod_{i=1}^{n}\lambda(\tau_i)\Psi(0,\tau_i), \tag{28}$$

and

$$\boxed{L_r = \frac{M^2(\tau_1)}{M_o^{n+1}}\prod_{i=2}^{n}\left[-\frac{dM}{d\tau}\bigg|_{\tau_i}\right].} \tag{29}$$

□

S.1.3 The likelihood in terms of $\lambda_{\rm p}$

In the following we show how the likelihood of an extant timetree under a birth-death model can be expressed purely in terms of the tree's LTT and the model's pulled speciation rate $\lambda_{\rm p}$.

We begin with the case where the stem age is known and the likelihood is conditioned on the survival of the stem lineage. Our starting point is the likelihood formula in Eq. (26):

$$L = \frac{M(\tau_o)}{M_o^{n+1}}\prod_{i=1}^{n}\left[-\frac{dM}{d\tau}\bigg|_{\tau_i}\right], \tag{30}$$

where M is the dLTT and $M_o := M(0)$. From Eq. (7) it is easy to obtain the following relationship between M and $\lambda_{\rm p}$:

$$M(\tau) = M_o e^{-\Lambda_{\rm p}(\tau)}, \tag{31}$$

where we defined:

$$\Lambda_{\rm p}(\tau) := \int_0^\tau \lambda_{\rm p}(s)\,ds. \tag{32}$$

Inserting Eq. (31) into Eq. (30) yields:

$$L = e^{-\Lambda_{\rm p}(\tau_o)}\prod_{i=1}^{n}\underbrace{\frac{-1}{M(\tau_i)}\frac{dM}{d\tau}\bigg|_{\tau_i}}_{\lambda_{\rm p}(\tau_i)}\cdot e^{-\Lambda_{\rm p}(\tau_i)}, \tag{33}$$

and hence:

$$\boxed{L = e^{-\Lambda_\mathrm{p}(\tau_o)} \prod_{i=1}^{n} \lambda_\mathrm{p}(\tau_i) \cdot e^{-\Lambda_\mathrm{p}(\tau_i)}.} \tag{34}$$

If only the crown age is known and the likelihood is conditioned on the splitting at the root and the survival of the root's two daughter-lineages (likelihood formula in Eq. (29)), we instead obtain the expression:

$$\boxed{L_r = \frac{e^{-\Lambda_\mathrm{p}(\tau_1)}}{\lambda_\mathrm{p}(\tau_1)} \prod_{i=1}^{n} \lambda_\mathrm{p}(\tau_i) \cdot e^{-\Lambda_\mathrm{p}(\tau_i)}.} \tag{35}$$

S.1.4 Calculating λ from r_p and μ

In the following we provide the general solution to the differential equation (2) in the main article:

$$\frac{d\lambda}{d\tau} = \lambda \cdot (r_\mathrm{p} + \mu^* - \lambda), \tag{36}$$

with initial condition:

$$\lambda(0) = \eta_o/\rho > 0. \tag{37}$$

We assume that r_p and μ^* are sufficiently "well-behaved", specifically that they are integrable over any finite interval. Observe that Eq. (36) is an example of a Bernoulli-type differential equation, as it can be written in the standard form:

$$\frac{d\lambda}{d\tau} = p(\tau)\lambda(\tau) + q(\tau)\lambda^\alpha(\tau), \tag{38}$$

where $\alpha = 2$, $p = r_\mathrm{p} + \mu^*$ and $q = -1$. Using the standard technique for solving Bernoulli differential equations (i.e., substituting $u = \lambda^{1-\alpha}$ to obtain a linear differential equation for u), it is straightforward to obtain the solution:

$$\boxed{\lambda(\tau) = \frac{\eta_o e^{\Lambda(\tau)}}{\rho + \eta_o \displaystyle\int_0^\tau e^{\Lambda(s)}\, ds},} \tag{39}$$

where we defined:

$$\Lambda(\tau) := \int_0^\tau [r_\mathrm{p}(s) + \mu^*(s)]\ ds. \tag{40}$$

Note that the solution in Eq. (39) is strictly positive and continuous, and hence λ is indeed a valid speciation rate.

For future reference, we mention that the above solution can be easily generalized to the case where the "initial condition" for λ is given at some arbitrary age τ_1, rather than at present-day. Specifically, the solution to the differential equation:

$$\frac{d\lambda}{d\tau} = \lambda \cdot (r_\mathrm{p} + \mu^* - \lambda), \tag{41}$$

with condition:

$$\lambda(\tau_1) = \lambda_1, \tag{42}$$

is given by:

$$\lambda(\tau) = \frac{\lambda_1 e^{\Lambda(\tau)}}{e^{\Lambda(\tau_1)} + \lambda_1 \int_0^{\tau} e^{\Lambda(s)}\, ds - \lambda_1 \int_0^{\tau_1} e^{\Lambda(s)}\, ds}. \tag{43}$$

Special cases:

- In the special case where r_{p} and μ^* are time-independent and $r_{\mathrm{p}} + \mu^* \neq 0$, the solution in Eq. (39) takes the form:

  $$\lambda(\tau) = \frac{P}{(P\rho/\eta_o - 1)e^{-P\tau} + 1}, \tag{44}$$

  where $P = r_{\mathrm{p}} + \mu^*$.
- If and only if $\mu^*(\tau) = \eta_o/\rho - r_{\mathrm{p}}(\tau)$, the solution in Eq. (39) is time-independent:

  $$\lambda(\tau) = \frac{\eta_o}{\rho}. \tag{45}$$

  Hence, for a fixed ρ, a congruence class can include at most one model with constant speciation rate; it includes exactly one model with constant speciation rate if and only if $\eta_o/\rho \geq \max_\tau r_{\mathrm{p}}(\tau)$.

S.1.5 Calculating λ from r_{p} and ε

In the following we show how the speciation rate λ can be calculated from the pulled diversification rate r_{p}, the present-day speciation rate λ_o and the ratio of extinction over speciation rate, $\varepsilon := \mu/\lambda$. Specifically, we provide the general solution to the following differential equation:

$$\frac{d\lambda}{d\tau} = \lambda \cdot \left[r_{\mathrm{p}} + (\varepsilon - 1)\lambda\right]. \tag{46}$$

We assume that r_{p} and ε are sufficiently "well-behaved", specifically that they are integrable over any finite interval. Observe that Eq. (46) is an example of a Bernoulli-type differential equation, as it can be written in the standard form:

$$\frac{d\lambda}{d\tau} = p(\tau)\lambda(\tau) + q(\tau)\lambda^{\alpha}(\tau), \tag{47}$$

where $\alpha = 2$, $p = r_{\mathrm{p}}$ and $q = \varepsilon - 1$. Using the standard technique for solving Bernoulli differential equations (i.e., substituting $u = \lambda^{1-\alpha}$ to obtain a linear differential equation for u), it is straightforward to obtain the solution:

$$\boxed{\lambda(\tau) = \frac{\lambda_o e^{R_{\mathrm{p}}(\tau)}}{1 + (1-\varepsilon)\cdot\lambda_o \int_0^{\tau} e^{R_{\mathrm{p}}(s)}\, ds},} \tag{48}$$

where we defined:

$$R_{\mathrm{p}}(\tau) := \int_0^{\tau} r_{\mathrm{p}}(s)\, ds. \tag{49}$$

In the special case where r_{p} is time-independent and non-zero, the solution in Eq. (48) simplifies to:

$$\lambda(\tau) = \frac{\lambda_o e^{r_{\mathrm{p}}\tau}}{1 + (1-\varepsilon)\cdot \dfrac{\lambda_o}{r_{\mathrm{p}}}\left(e^{r_{\mathrm{p}}s} - 1\right)}. \tag{50}$$

S.1.6 The likelihood in terms of the r_{p}

In the following we show how the likelihood of a tree under a birth-death model can be expressed solely in terms of the model's pulled diversification rate r_{p} and the product $\rho\lambda_o$. We first consider the case where the stem age is known and the likelihood is conditioned on the survival of the stem lineage (5); the alternative case where only the crown age is known and the likelihood is conditioned upon the survival of the root's two daughter lineages (Eq. 28) can be treated similarly and is briefly mentioned at the end.

Our starting point is the likelihood formula in Eq. (16), Supplement S.1.2. Define:

$$R_{\mathrm{p}}(\tau) := \int_0^{\tau} r_{\mathrm{p}}(u)\, du. \tag{51}$$

Then from the definition of r_{p} (Eq. 1 in the main article) we have:

$$R_{\mathrm{p}}(\tau) = \int_0^{\tau} [\lambda(u) - \mu(u)]\ du + \int_0^{\tau} \frac{d\ln\lambda}{du}\, du = R(\tau) + \ln\frac{\lambda(\tau)}{\lambda_o}. \tag{52}$$

Exponentiating (52) and rearranging yields:

$$e^{R(\tau)} = e^{R_{\mathrm{p}}(\tau)}\frac{\lambda_o}{\lambda(\tau)}. \tag{53}$$

Inserting Eq. (53) into the definition of Ψ in Eq. (13) yields:

$$\Psi(0,\tau) = e^{R_{\mathrm{p}}(\tau)}\frac{\lambda_o}{\lambda(\tau)}\left[1 + \rho\lambda_o\int_0^{\tau} e^{R_{\mathrm{p}}(u)}\, du\right]^{-2}. \tag{54}$$

Inserting Eq. (54) into the likelihood formula (16) yields:

$$L = \frac{(\rho\lambda_o)^{n+1}e^{R_{\mathrm{p}}(\tau_o)}}{[1-E(\tau_o)]\,\lambda(\tau_o)}\left[1 + \rho\lambda_o\int_0^{\tau_o} e^{R_{\mathrm{p}}(u)}\, du\right]^{-2}\prod_{i=1}^{n} e^{R_{\mathrm{p}}(\tau)}\left[1 + \rho\lambda_o\int_0^{\tau} e^{R_{\mathrm{p}}(u)}\, du\right]^{-2}. \tag{55}$$

Recall that $(1-E)\lambda = \lambda_{\mathrm{p}}$ according to Eq. (8), which when inserted into Eq. (55) yields:

$$L = \frac{(\rho\lambda_o)^{n+1}e^{R_{\mathrm{p}}(\tau_o)}}{\lambda_{\mathrm{p}}(\tau_o)}\left[1 + \rho\lambda_o\int_0^{\tau_o} e^{R_{\mathrm{p}}(u)}\, du\right]^{-2}\prod_{i=1}^{n} e^{R_{\mathrm{p}}(\tau)}\left[1 + \rho\lambda_o\int_0^{\tau} e^{R_{\mathrm{p}}(u)}\, du\right]^{-2}. \tag{56}$$

From Eqs. (8) and (11) we know that λ_{p} satisfies the initial value problem (Bernoulli differential equation):

$$\frac{d\lambda_{\mathrm{p}}}{d\tau} = \lambda_{\mathrm{p}} \cdot (r_{\mathrm{p}} - \lambda_{\mathrm{p}}), \quad \lambda_{\mathrm{p}}(0) = \rho\lambda_o. \tag{57}$$

It is straightforward to verify that the solution to Eq. (57) is given by:

$$\lambda_{\mathrm{p}}(\tau) = \frac{\rho\lambda_o e^{R_{\mathrm{p}}(\tau)}}{1 + \rho\lambda_o \int_0^\tau e^{R_{\mathrm{p}}(u)} du}. \tag{58}$$

Inserting the solution (58) into Eq. (56) yields the following expression for the likelihood:

$$\boxed{L = \left[1 + \rho\lambda_o \int_0^{\tau_o} e^{R_{\mathrm{p}}(u)}\, du\right]^{-1} (\rho\lambda_o)^n \prod_{i=1}^{n} e^{R_{\mathrm{p}}(\tau_i)} \left[1 + \rho\lambda_o \int_0^{\tau_i} e^{R_{\mathrm{p}}(u)}\, du\right]^{-2}.} \tag{59}$$

In the alternative case where only the crown age is known, and the likelihood is conditioned on the splitting at the root and the survival of the root's two daughter lineages, we obtain the following expression for the likelihood:

$$\boxed{L_r = e^{-R_{\mathrm{p}}(\tau_1)} (\rho\lambda_o)^{n-1} \prod_{i=1}^{n} e^{R_{\mathrm{p}}(\tau_i)} \left[1 + \rho\lambda_o \int_0^{\tau_i} e^{R_{\mathrm{p}}(u)}\, du\right]^{-2}.} \tag{60}$$

Corollary: A corollary of the above results is that two models have the same likelihood function if and only if they have the same r_{p} and product $\rho\lambda_o$. We note that this corollary can also be derived using previous results by Lambert *et al.* (6), as follows. Lambert *et al.* (6) showed that the distribution of a model's generated LTTs is entirely determined by a single function F, which is given by the inverse of the tail distribution function of coalescence ages:

$$F(\tau) = 1 + \rho \int_0^\tau \lambda(s) e^{R(s)}\, ds, \tag{61}$$

where R is defined as in Eq. (14). Since $F(0)$ is always 1, two models $(\lambda_1, \mu_1, \rho_1)$ and $(\lambda_2, \mu_2, \rho_2)$ have identical F ($F_1 = F_2$) if and only if $dF_1/d\tau = dF_2/d\tau$ at all ages τ, that is:

$$\rho_1\lambda_1(\tau)e^{R_1(\tau)} = \rho_2\lambda_2(\tau)e^{R_2(\tau)}. \tag{62}$$

Equation (62) holds if and only if $\rho_1\lambda_1(0) = \rho_2\lambda_2(0)$ and:

$$\begin{aligned} &\rho_1\lambda_1(\tau)e^{R_1(\tau)} \left(\frac{1}{\lambda_1(\tau)}\frac{d\lambda_1(\tau)}{d\tau} + \lambda_1(\tau) - \mu_1(\tau)\right) \\ =&\rho_2\lambda_2(\tau)e^{R_2(\tau)} \left(\frac{1}{\lambda_2(\tau)}\frac{d\lambda_2(\tau)}{d\tau} + \lambda_2(\tau) - \mu_2(\tau)\right), \end{aligned} \tag{63}$$

where condition (63) was obtained by differentiating both sides in Eq. (62). Combining Eq. (62) and Eq. (63) yields:

$$\lambda_1 - \mu_1 + \frac{1}{\lambda_1}\frac{d\lambda_1}{d\tau} = \lambda_2 - \mu_2 + \frac{1}{\lambda_2}\frac{d\lambda_2}{d\tau}. \tag{64}$$

Since the two sides of Eq. (64) correspond to the pulled diversification rates of the models, it follows that two models are congruent if and only if they have the same product $\rho\lambda_o$ and the same $r_{\rm p}$.

S.1.7 Congruent models have the same probability distribution of generated tree sizes

In the following, we show that the distribution of extant timetree sizes generated by a birth-death model, either conditional upon the age and survival of the stem, or conditional upon the age of the root and the survival of its two daughter lineages, is the same for all models in a congruence class.

Consider a birth-death process with parameters (λ, μ, ρ), starting from a single lineage at some time before present τ_o and ultimately resulting in a timetree at age 0, comprising only extant species that are included at some probability ρ. The probability that the timetree will comprise n tips can be expressed using formulas first derived by Kendall *et al.* (9):

$$\begin{aligned} &P(n) = (1 - E(\tau_o)) \cdot (1 - H) \cdot H^{n-1}, \quad n \geq 1 \\ &P(0) = E(\tau_o), \end{aligned} \tag{65}$$

where $E(\tau_o)$ is the probability that a lineage existing at age τ_o will be missing from the timetree (as defined previously), H is defined as:

$$H := \frac{\rho \displaystyle\int_0^{\tau_o} e^{R(s)} \lambda(s)\, ds}{1 + \rho \displaystyle\int_0^{\tau_o} e^{R(s)} \lambda(s)\, ds}, \tag{66}$$

and R was previously defined in Eq. (14). Note that the formula in Eq. (65) can be readily obtained using equations 8, 10b and 11 in (9), after setting the time variable therein equal to τ_o (i.e. $t = \tau_o$), switching from time to age ($\tau = \tau_o - t$), and adding the term $-\delta(\tau) \ln \rho$ to the extinction rate (where δ is the Dirac distribution, peaking at age 0) to account for incomplete species sampling. As shown previously in Eq. (53), we have

$$e^{R(\tau)} = e^{R_{\rm p}(\tau)} \frac{\lambda_o}{\lambda(\tau)}, \tag{67}$$

where $R_{\rm p}$ is defined as:

$$R_{\rm p}(\tau) := \int_0^{\tau} r_{\rm p}(u)\, du, \tag{68}$$

and $r_{\rm p}$ is the pulled diversification rate. Inserting Eq. (67) into Eq. (66) allows us to write H as follows:

$$H = \frac{\rho\lambda_o \displaystyle\int_0^{\tau_o} e^{R_{\rm p}(s)}\, ds}{1 + \rho\lambda_o \displaystyle\int_0^{\tau_o} e^{R_{\rm p}(s)}\, ds}. \tag{69}$$

Since $\rho\lambda_o$, $r_{\rm p}$ and $R_{\rm p}$ are the same for all models in a congruence class, H is also constant across the congruence class.

The probability of obtaining a tree of size $n \geq 1$ conditional upon the age of the stem lineage (τ_o) and its survival to the present, denoted $P_{\text{stem}}(n)$, is given by the ratio $P(n)/(1-E(\tau_o))$, i.e.:

$$\boxed{P_{\text{stem}}(n) = (1-H)\cdot H^{n-1}.} \tag{70}$$

Since H is constant across a congruence class, the same also holds for $P_{\text{stem}}(n)$ for any n. The probability of obtaining a tree of size $n \geq 1$ conditional upon the splitting of the root at age τ_o and the survival of its two daughter lineages, denoted $P_{\text{root}}(n)$, can be derived in a similar way, as follows. The probability that the two daughter lineages survive, conditional upon the split at age τ_o, is given by the product:

$$P(\text{daughter lineages survive} \mid \text{split at } \tau_o) = (1-E(\tau_o))^2. \tag{71}$$

The probability that the two daughter lineages survive and the timetree has size $n \geq 1$, conditional upon the split at age τ_o, is given by the following sum of probabilities:

$$\begin{aligned}
&P(\text{daughter lineages survive and tree has size } n \mid \text{split at } \tau_o)\\
&= \sum_{k=1}^{n-1} P(k)P(n-k)\\
&= (1-E(\tau_o))^2 \sum_{k=1}^{n-1}(1-H)\cdot H^{k-1}\cdot(1-H)\cdot H^{n-k-1}\\
&= (1-E(\tau_o))^2(1-H)^2\sum_{k=1}^{n-1} H^{n-2}\\
&= (n-1)\cdot(1-E(\tau_o))^2(1-H)^2H^{n-2}.
\end{aligned} \tag{72}$$

Dividing Eq. (72) by Eq. (71) yields the desired probability:

$$\boxed{P_{\text{root}}(n) = (n-1)\cdot(1-H)^2H^{n-2}.} \tag{73}$$

Since H is constant across the congruence class, the same also holds for $P_{\text{root}}(n)$.

□

S.1.8 On the nature of congruence classes

In the following, we provide a formal definition of model congruence classes, and point out an analogy to the concept of object congruency in geometry. A basic background in abstract algebra is assumed.

In geometry, two objects are called congruent if they exhibit similar geometric properties, such as identical angles between corresponding lines and identical distances between corresponding points. More precisely, two geometric objects (sets of points in Euclidean space $\mathbb{R}^n$) are called congruent if one set can be transformed into the other set by means of an isometry, i.e. a mapping that preserves distances between pairs of points (via translations, rotations, and/or reflections). Object congruency is a type of equivalence relation, and hence the set of models congruent to some focal object is an equivalence class. The set of all isometries is itself a group (known as "Euclidean group") that acts on the set of geometric objects, and congruence classes of objects correspond to "orbits" under the action of isometries (10). By analogy, two birth-death models are called "congruent" if they exhibit similar statistical properties in terms of their generated extant timetrees and LTTs (see main text and Supplement S.1). As we show below, congruence classes can be interpreted as the orbits of a group of mappings acting on model space that preserve dLTTs (just as isometries preserve distances in Euclidean space).

While the "congruence" relationship is well-defined (two models defined on the same age interval are congruent if their dLTTs exist as unique solutions to the differential equation (5) and are identical at all ages), the precise nature of a "congruence class" depends on the particular model space considered, i.e. the types of time-dependent curves one is willing to consider for λ and μ and the values one is willing to consider for ρ. For example, one might choose to only consider continuous, or only continuously differentiable, or only twice continuously differentiable functions λ and μ and so on. One might also want to restrict ages within a specific interval $[0, \tau_o]$, where τ_o is chosen to cover all relevant stem or crown ages possibly encountered in real data, and one might want to fix ρ to a specific value. Once a model space $\mathcal{B}$ (i.e., the function spaces for λ and μ, and the allowed values for ρ) is chosen, the congruence class of a model $x \in \mathcal{B}$ is the set of all models $y \in \mathcal{B}$ congruent to that model; note that any two models in a congruence class are themselves congruent to each other. Congruence specifies an equivalence relation on $\mathcal{B}$, and congruence classes are the corresponding equivalence classes within $\mathcal{B}$.

For technical reasons, in this section we shall only consider the space of birth-death models (denoted $\mathcal{B}$) with strictly positive λ, μ and ρ and continuously differentiable λ and μ defined over some age interval $[0, \tau_o] \subseteq \mathbb{R}$. Let $\mathcal{C}^1_+[0, \tau_o]$ denote the set of all continuously differentiable real-valued strictly positive functions defined on the interval $[0, \tau_o]$. For any $S_o \in (0, \infty)$ and any $f \in \mathcal{C}^1_+[0, \tau_o]$, define $S[S_o, f] \in \mathcal{C}^1_+[0, \tau_o]$ as the solution to the following initial value problem:

$$\frac{dS[S_o, f]}{d\tau} = S[S_o, f](\tau) \cdot [f(\tau) - S[S_o, f](\tau)], \quad S[S_o, f](0) = S_o. \tag{74}$$

It is straightforward to verify that the solution to the above problem is given by:

$$S[S_o, f](\tau) = \frac{S_o e^{F(\tau)}}{1 + S_o \int_0^\tau e^{F(s)}\, ds}, \tag{75}$$

where we denoted:

$$F(\tau) := \int_0^\tau f(s)\, ds. \tag{76}$$

For any arbitrary $\alpha \in (0, \infty)$ and $\beta \in \mathcal{C}^1_+[0, \tau_o]$, let $g_{\alpha,\beta} : \mathcal{B} \to \mathcal{B}$ be a transformation of birth-death models defined as follows:

$$g_{\alpha,\beta}(\lambda, \mu, \rho) := \left(S\left[\lambda/\alpha, \lambda - \mu + \frac{1}{\lambda}\frac{d\lambda}{d\tau} + \beta\mu\right], \beta\mu, \alpha\rho \right). \tag{77}$$

Note that $g_{\alpha,\beta}$ is dLTT-preserving, that is, it maps models to models within the same congruence class. Indeed, the variable

$$\lambda^* := S\left[\lambda/\alpha, \lambda - \mu + \frac{1}{\lambda}\frac{d\lambda}{d\tau} + \beta\mu\right] \tag{78}$$

is exactly the speciation rate of a model with extinction rate $\mu^* := \beta\mu \in \mathcal{C}^1_+[0, \tau_o]$ and sampling fraction $\rho^* := \alpha\rho \in (0, \infty)$, congruent to the original model (λ, μ, ρ). The set of all such transformations,

$$G := \left\{ g_{\alpha,\beta} : \alpha \in (0, \infty), \beta \in \mathcal{C}^1_+[0, \tau_o] \right\}, \tag{79}$$

constitutes a group with group operation:

$$g_{\alpha,\beta} \circ g_{\tilde{\alpha},\tilde{\beta}} := g_{\alpha\tilde{\alpha},\beta\tilde{\beta}} \tag{80}$$

and identity element $g_{1,1}$. The group G acts on the set of birth-death models, while preserving dLTTs. Abstractly, each mapping $g \in G$ corresponds to an "isometric" transformation in model space that preserves the statistics of generated extant timetrees and dLTTs, in analogy to how rotations, translations or reflections preserve distances in Euclidean space.

Note that not all dLTT-preserving mappings defined on $\mathcal{B}$ are members of G. It turns out, however, that G is large enough to completely generate congruence classes in $\mathcal{B}$. In other words, for any model $(\lambda, \mu, \rho) \in \mathcal{B}$, the orbit:

$$G(\lambda, \mu, \rho) := \{g(\lambda, \mu, \rho) : g \in G\} \tag{81}$$

is exactly the congruence class of the model; indeed, for any congruent model $(\lambda^*, \mu^*, \rho^*) \in \mathcal{B}$ one can find a transformation $g_{\alpha,\beta} \in G$ such that $(\lambda^*, \mu^*, \rho^*) = g_{\alpha,\beta}(\lambda, \mu, \rho)$, by choosing $\alpha := \rho^*/\rho$ and $\beta := \mu^*/\mu$.

S.2 Why standard model selection methods cannot resolve model congruencies

Here we explain why model selection methods based on parsimony or "Occam's razor", such as the Akaike Information Criterion (AIC; 11) and the Bayesian Information Criterion (BIC; 12) that penalize excessive parameters, generally cannot resolve the identifiability issues discussed in the main article. First, there is generally little reason to believe that the simplest scenario in a congruence class will be the one closest to the truth. Indeed, even if the true model is included in a congruence class, it will almost always be the case that there are both simpler and more complex scenarios within the same congruence class (e.g., Extended data figure 2) and, crucially, all of these alternative models remain equally likely even with infinitely large datasets.

Second, if one were to apply AIC or BIC, it is unclear how to quantify the complexity of a diversification scenario in comparison with alternative scenarios, which may be described using different functional forms. It is tempting to think that one could simply count the number of parameters. However, any given curve can be written using various alternative functional forms parameterized in distinct ways (recall that ultimately we wish to approximately estimate the curves λ and μ, not the parameters of some functional form); the number of parameters is a property of parameterized sets of curves, not of a single curve. Even if that were not the case, the number of parameters conventionally associated with a given functional form need not necessarily reflect our intuition about complexity. In addition, different members of a congruence class may be described with different functional forms involving the same number of parameters. For example, the diversification scenario with linear extinction rate ($\mu = \alpha + \beta \cdot \tau$) and constant speciation rate ($\lambda = \gamma$) (3 parameters, assuming complete species sampling) is congruent to an alternative and markedly different scenario with zero extinction rate ($\mu^* = 0$) and λ^* defined as the solution to the differential equation $d\lambda^*/d\tau = \lambda^* \cdot (\gamma - \alpha - \beta\tau - \lambda)$ with initial condition $\lambda^*(0) = \gamma$ (again 3 parameters); there is no reason to prefer one congruent scenario over the other based on the number of parameters or biological realism.

Third, even when fitting models of the same functional form, as explained in the main article the maximum-likelihood model will a priori tend to be the one closest to the congruence class of the true process, rather than the true process itself, and neither AIC nor BIC would resolve this (since all other allowed models would have the same number of parameters but lower likelihood). Extended data figure 6 shows examples where maximum-likelihood fitted models, chosen among a wide range of model complexities based on AIC, grossly fail to estimate the true rates even when fitting to a massive tree with 1,000,000 tips, despite the fact that the allowed model sets (i.e., functional forms) could in principle have accurately reproduced the true rates (i.e., model inadequacy is not the issue).

Note that model selection methods based on significance tests (e.g., likelihood ratio tests) cannot be used either for selecting between congruent scenarios (in fact the likelihood ratio would always be 1). Significance tests for model selection are designed for situations where simpler models should be preferred when statistical support for more complex models is lacking due to limited data size, and where eventually, i.e., with increasing data, statistical support can accumulate for more complex models that capture the additional complexity of the studied process without the risk of overfitting.

S.3 Why previous studies failed to detect model congruencies

In practice, reconstructions of λ and μ over time are typically performed by selecting among a limited set of allowed models, i.e., considering specific functional forms described by a finite number of parameters (13, 14, 15, 16, 17, 5). In these situations it is generally unlikely that the allowed model set intersects a given congruence class more than once (see Supplement S.7 for mathematical justification). For example, when considering only constant-rate birth-death models and assuming that ρ is fixed (as is usually the case; 18), each congruence class reduces to a single combination of λ and μ. Likelihood functions defined over a limited allowed model set thus generally don't exhibit ridges associated with congruence classes, and may even exhibit a unique global maximum in the space of considered parameters, leaving the impression that λ and μ have been estimated close to their true values. Our findings suggest that this impression is almost certainly false. Instead, obtained estimates for λ and μ are almost always going to be a random outcome that depends on the particular choice of allowed models, such as the functional forms considered for λ and μ, and will be as close as possible to the congruence class of the truth rather than close to the truth itself. Unless one has reasons to prefer specific functional forms for λ and μ (e.g., based on a mechanistic macroevolutionary model; 19), fitted λ and μ are unlikely to resemble the true rates even if in principle the functional forms considered are flexible enough to resemble the true λ and μ (see Supplement S.10 for examples using simulations and real data).

Previous studies have failed to recognize the breadth of model congruencies because they typically only consider a limited set of candidate models at a time, both when analyzing real datasets as well as when assessing parameter identifiability via simulations. For example, if a tree was generated by an exponentially decaying λ and μ (e.g., via simulations), then fitting an exponential functional form will of course yield accurate estimates of the exponents; however if the generating process was only approximately exponential and better described by another gradually decaying function, then fitting an exponential curve could even lead to opposite trends (examples in Fig. 2, Extended data figure 2 and Supplement S.10).

S.4 Cetacean diversification as an example

Here we discuss a real-world example where the existence of congruent scenarios has major macroevolutionary implications. Steeman *et al.* (20) reconstructed past speciation rates of Cetaceans (whales, dolphins, and porpoises) based on an extant timetree and using maximum-likelihood (assuming $\mu = 0$). Steeman *et al.* (20) found a temporary increase of λ during the late Miocene-early Pliocene (Extended data figure 3b), suggesting a potential link between Cetacean radiations and concurrent paleoceanographic changes. However, alternatively to assuming $\mu = 0$, one could assume that μ was close to λ, consistent with common observations from the fossil record (21). For example, by setting $\mu = 0.9 \cdot \lambda$ one obtains a congruent scenario in which λ no longer peaks during the late Miocene-early Pliocene but instead exhibits a gradual slowdown throughout most of Cetacean evolution (Extended data figure 3b). Both scenarios are similarly complex and both could have generated the timetree at equal probabilities.

S.5 Implications for inferring environmental correlations

Studies that test whether diversification dynamics are influenced by some environmental or geological variable X (e.g., temperature), either by testing for correlations between X and the estimated λ or μ (22, 23) or by fitting models in which λ or μ are explicit functions of X (24, 25, 26), will generally lead to unreliable conclusions. In the first scenario, since λ and μ themselves cannot be reliably estimated, any observed correlations with X will likely be inaccurate. In the second scenario, specifying λ or μ as functions of X (e.g., assuming $\mu = \alpha X + \beta$ and fitting the coefficients α and β) is essentially equivalent to choosing particular functional forms for λ or μ. Unless these functional forms are strongly justified on mechanistic grounds (which they usually aren't in practice), the model coefficients fitted to the data will correspond to a model closest to the true diversification history's congruence class, but not the necessarily the true diversification history itself. Hence, any resulting conclusions about the effects of X on λ and μ will often be wrong.

S.6 Potential implications for trait-dependent diversification models

Trait data combined with phylogenetic data, modeled using trait-dependent diversification models with time-variable rates (e.g., time-dependent Binary State Speciation and Extinction models; 25), might perhaps resolve the issues discussed here, although in our opinion the chances for that are rather slim. On the one hand, the additional data (tip character states) could provide sufficient information to resolve ambiguities; on the other hand the number of degrees of freedom is also larger, since each character state can exhibit distinct λ and μ over time (not to mention that character transition rates might themselves be unknown functions of time). The likelihood functions of trait-dependent diversification models tend to be substantially more complex than for the birth-death model, and hence a resolution of their identifiability is far beyond the scope of this article.

S.7 Typical model sets do not exhibit congruence ridges

In the following we explain why it is unlikely in practice that a limited set of allowed models (e.g., considered for maximum-likelihood estimation) will intersect any given congruence class more than once, and that it is especially unlikely that multiple intersections of a congruence class form a sub-manifold in parameter space (i.e., a "congruence ridge"). Consider a set of allowed models, parameterized through n independent parameters $q_1, .., q_n \in \mathbb{R}$, i.e. such that the speciation and extinction rates of a model are given as functions of age (τ) and the chosen parameters ($\mathbf{q} \in \mathbb{R}^n$):

$$\lambda = \lambda(\tau, \mathbf{q}), \quad \mu = \mu(\tau, \mathbf{q}). \tag{82}$$

For simplicity, assume that the sampling fraction ρ is given (identifiability issues associated with uncertainties in ρ are already well known; 27, 28, 29, 30).

Now consider some particular choice of parameters, $\mathbf{q}$, with corresponding PDR:

$$r_{\mathrm{p}}(\tau,\mathbf{q}) = \lambda(\tau,\mathbf{q}) - \mu(\tau,\mathbf{q}) + \frac{1}{\lambda(\tau,\mathbf{q})}\frac{\partial\lambda(\tau,\mathbf{q})}{\partial\tau}, \tag{83}$$

and present-day speciation rate $\lambda(0,\mathbf{q})$. For any other choice of parameters $\mathbf{h}\in\mathbb{R}^n$, the corresponding model would be in the same congruence class as the first model if and only if $\lambda(0,\mathbf{h}) = \lambda(0,\mathbf{q})$ and $r_{\mathrm{p}}(\tau,\mathbf{h}) = r_{\mathrm{p}}(\tau,\mathbf{q})$ for all ages $\tau\geq 0$, in other words $\lambda(\cdot,\mathbf{h})$ must be a solution to the initial value problem:

$$\frac{\partial\lambda(\tau,\mathbf{h})}{\partial\tau} = \lambda(\tau,\mathbf{h})\cdot\left[r_{\mathrm{p}}(\tau,\mathbf{q}) - \lambda(\tau,\mathbf{h}) + \mu(\tau,\mathbf{h})\right], \quad \lambda(0,\mathbf{h}) = \lambda(0,\mathbf{q}). \tag{84}$$

Unless the functional forms of λ and μ have been specifically designed for this purpose, it is generally unlikely that Eq. (84) will be satisfied for some $\mathbf{h}\neq\mathbf{q}$.

A stronger argument for the low probability of congruence ridges can be made as follows. Suppose that $\mathbf{q}$ was part of a congruence ridge, i.e. a sub-manifold in parameter space belonging to the same congruence class. Then there must exist a curve in parameter space, i.e. a one-parameter function $\mathbf{h}:[-\varepsilon,\varepsilon]\to\mathbb{R}^n$, passing through $\mathbf{q}$ (e.g., say $\mathbf{h}(0)=\mathbf{q}$), such that:

$$r_{\mathrm{p}}(\tau,\mathbf{h}(s)) = r_{\mathrm{p}}(\tau,\mathbf{q}), \tag{85}$$

and such that:

$$\lambda(0,\mathbf{h}(s)) = \lambda(0,\mathbf{q}), \tag{86}$$

for all $s\in[-\varepsilon,\varepsilon]$ and all $\tau\geq 0$. Taking the derivative of Eq. (85) with respect to s at 0 yields:

$$\sum_{i=1}^{n}\left.\frac{\partial r_{\mathrm{p}}}{\partial q_i}\right|_{(\tau,\mathbf{q})}\cdot\left.\frac{dh_i}{ds}\right|_{s=0} = 0. \tag{87}$$

Denote $\mathbf{H} := \frac{d\mathbf{h}}{ds}\big|_{s=0}$ and $\mathbf{R}(\tau) := \frac{\partial r_{\mathrm{p}}}{\partial\mathbf{q}}\big|_{\tau,\mathbf{q}}$. Then the condition in Eq. (87) can be written in vector notation:

$$\mathbf{R}(\tau)^{\mathrm{T}}\cdot\mathbf{H} = 0. \tag{88}$$

Note that $\mathbf{H}$ can be interpreted as the "velocity vector" along the ridge curve $\mathbf{h}$ at the point $\mathbf{q}$, and hence condition (88) means that the ridge must move perpendicular to the direction of steepest descent of r_{p}. Observe that condition (88) must be satisfied for all ages $\tau\geq 0$. Hence, for any arbitrary choice $\tau_1,\tau_2,..,\tau_m\geq 0$, we obtain the following m linear equations that must be satisfied by $\mathbf{H}$:

$$\begin{gathered}\mathbf{R}(\tau_1)^{\mathrm{T}}\cdot\mathbf{H} = 0.\\ \vdots \\ \mathbf{R}(\tau_m)^{\mathrm{T}}\cdot\mathbf{H} = 0.\end{gathered} \tag{89}$$

Unless the functional forms of λ and μ are specifically designed for this purpose, the system in Eq. (89) will almost certainly be over-determined if m is chosen sufficiently high ($m\gg n$). Hence, in practice, for a chosen set of allowed models and a given point $\mathbf{q}$ in parameter space, a congruence ridge will almost never exist at that point.

□

S.8 Interpreting the PDR and other congruence-invariant variables

Here we discuss various interpretations and uses of the pulled diversification rate (PDR) and other related variables. Given that the PDR is a composite quantity that depends on both λ and μ (Eq. 1), properly interpreting the estimated PDR in terms of actual speciation/extinction rates remains the responsibility of the investigator. Previous work has shown that the PDR can indeed yield valuable insight into diversification dynamics and can be useful for testing alternative hypotheses (7). For example, sudden rate transitions (e.g., due to mass extinction events) almost always lead to fluctuations in the PDR; thus, a relatively constant PDR over time would be indicative of constant — or only slowly changing — speciation and extinction rates.

The PDR can be used to obtain other useful variables. For example, it is straightforward to confirm that the PDR and the total diversity N satisfy the following relationship:

$$\frac{N(\tau)}{N(0)} \cdot \frac{\lambda_o}{\lambda(\tau)} = \exp\left[-\int_0^\tau r_{\rm p}(u)\,du\right]. \tag{90}$$

Observe that the left hand side of this equation, henceforth called deterministic "pulled normalized diversity" (dPND), corresponds to the ratio of deterministic total diversity at some age τ over the assumed present-day total diversity $N(0)$, modulated by the factor $\lambda_o/\lambda(\tau)$. Like the PDR, the dPND is the same for all models in a congruence class, and can be readily estimated from extant timetrees (Fig. 8c). As becomes apparent from Eq. (90), while the dPND can yield information on variations of past diversity, the amount of information depends on how well λ can be constrained a priori.

Another useful derived variable is the "pulled extinction rate", or PER (7), defined as:

$$\mu_{\rm p} := \lambda_o - r_{\rm p}. \tag{91}$$

The PER is equal to the extinction rate μ if λ is time-independent, but differs from μ in most other cases. Note that calculating the PER requires knowing the present-day speciation rate λ_o, which can be estimated from the timetree if the sampling fraction ρ is known (simply divide the estimated $\rho\lambda_o$ by ρ). The present-day PER is related to the present-day extinction rate as follows:

$$\mu_{\rm p}(0) = \mu(0) - \frac{1}{\lambda_o}\frac{d\lambda}{d\tau}\bigg|_{\tau=0}. \tag{92}$$

Observe that if the present-day speciation rate changes only slowly, the present-day PER will resemble the present-day extinction rate. Further, since $\mu(0)$ is non-negative, we can obtain the following lower bound for the exponential rate at which λ changes:

$$\frac{1}{\lambda_o}\frac{d\lambda}{d\tau}\bigg|_{\tau=0} \geq -\mu_{\rm p}(0). \tag{93}$$

In particular, if the estimated $\mu_{\rm p}(0)$ is negative, this is evidence that λ is currently decreasing over time.

We emphasize that other useful parameterizations (invariants) of congruence classes may also exist, each being appropriate for different purposes. For example, the distribution of branching ages (an invariant of congruence classes) can be described using a single random variable H, whose tail distribution function is well-described (6, Proposition 5 therein), and which is particularly useful for efficiently simulating timetrees.

S.9 Fitting congruence classes instead of models

The discussion in the main article revealed that speciation and extinction rates constitute partly interchangeable (and thus partly redundant) parameters that cannot be completely resolved from extant timetrees alone, no matter how large the dataset. Extant timetrees do, however, contain the proper information to estimate the pulled diversification rate $r_{\rm p}$ and η_o (recall that $\eta_o = \rho\lambda_o$), and may thus be used to at least identify the congruence class from which a tree was likely generated. Indeed, for sufficiently large data sizes, $\lambda_{\rm p}$, $r_{\rm p}$ and η_o can be directly calculated from the slope and curvature of the tree's LTT (7), which shows that it is possible to design asymptotically consistent statistical estimators for these variables (simulation examples in Extended data figure 8). In fact, it is straightforward to design maximum-likelihood estimators for $\lambda_{\rm p}$, $r_{\rm p}$ and η_o, as illustrated below.

Since each congruence class corresponds to a unique $r_{\rm p}$ and η_o, the $r_{\rm p}$ and η_o can be used to parameterize the space of congruence classes; on this space the likelihood function no longer exhibits the highly problematic ridges seen in the original model space. We thus suggest describing birth-death models in terms of $r_{\rm p}$ and η_o, rather than λ and μ, when fitting models to timetrees. Since the likelihood function can be expressed directly in terms of $r_{\rm p}$ and η_o (Supplement S.1.6), such a parameterization is suitable for maximum-likelihood or Bayesian estimation methods. Reciprocally, since every given $r_{\rm p}$ and η_o correspond to a unique and non-empty congruence class (as shown in the main article), any $r_{\rm p}$ and η_o estimated from an extant timetree will represent at least one biologically meaningful scenario. It is thus possible to directly fit congruence classes, rather than individual models, via maximum-likelihood. A similar reasoning can also be applied to the pulled speciation rate $\lambda_{\rm p}$, which provides an alternative representation of congruence classes.

To demonstrate this approach, we created software for fitting $r_{\rm p}$ and η_o to extant timetrees via maximum likelihood. The code is integrated into the R package `castor` (31) as function `fit_hbd_pdr_on_grid`. The function accepts as input an extant timetree, and an arbitrary number of discrete ages at which to estimate $r_{\rm p}$, assuming $r_{\rm p}$ varies linearly or polynomially between those ages. In other words, the functional form for $r_{\rm p}$ fitted is that of a piecewise linear or piecewise polynomial (spline) curve with pre-specified knot ages. The function also accepts optional lower and upper bounds for the fitted $r_{\rm p}$ and/or η_o. The code then maximizes the likelihood of the tree, given by Eq. (56) in the Supplement, by iteratively refining the $r_{\rm p}$ values on the age grid and/or η_o. Optionally, one can limit the evaluation of the likelihood function to a smaller "truncated" age interval than covered by the tree, i.e. some age interval $[0, \tau^*]$, where τ^* may be smaller than the root age. This may be useful for avoid estimation errors towards older ages due to a small number of lineages in the tree. The likelihood formula for the "truncated" case can be easily obtained by assuming that the tree is split into multiple sub-trees, each originating at the truncation age, and considering each sub-tree an independent realization of the same birth-death process and subject to the same sampling fraction ρ. To avoid non-global local optima, the fitting can be repeated multiple times, each time starting at random start values for the fitted parameters, and the best fit among all repeats is kept. We also developed similar computer code for fitting the pulled speciation rate $\lambda_{\rm p}$ to extant timetrees, implemented as function `fit_hbd_psr_on_grid` in the R package `castor`.

Extended data figure 8 shows an example where either the $r_{\rm p}$ or $\lambda_{\rm p}$ were accurately fitted to an extant timetree, simulated under a birth-death scenario subject to an early rapid radiation event and followed by a mass extinction event. In this example, we limited fitting to ages where the LTT was over 500 lineages (i.e., $M(\tau^*) = 500$), and repeated the fitting 100 times to avoid non-global local optima.

S.10 Fitting birth-death models to trees yields unreliable results

To illustrate the identifiability issues discussed in the main article and the fact that these cannot be resolved using common parsimony methods, we simulated and analyzed two massive extant timetrees ($\sim$114,000 and $\sim$785,000 tips) via a birth-death process, subject to a mass extinction event (both trees) and a rapid radiation event (second tree). Instead of fitting models of the exact same functional form as used in the simulations, we fitted generic piecewise-linear curves for λ and μ that could in principle take various alternative shapes (including approximately the shapes used for the simulations), and visually compared the estimated profiles to their true profiles (Extended data figures 5a–f). Specifically, we fitted λ and μ at multiple discrete time points, treating the rates at each time point as free parameters, while assuming a known ρ. Despite the enormous tree sizes, and despite the fact that the fitted models reproduced the trees' LTTs and the true $r_{\rm p}$ extremely well (Extended data figures 5a,c,d,f), the estimated λ and μ were far from their true values and even exhibited spurious trends (Extended data figures 5b,e). This is consistent with our expectation that the particular combination of fitted λ and μ is essentially a random pick from the periphery of the true process's congruence class. In contrast, when we fixed μ to its true profile, λ was accurately estimated (Extended data figure 7), consistent with the expectation that any given μ and ρ fully determine the corresponding λ in the congruence class.

We also examined a large extant timetree of 79,874 seed plant species published by Smith *et al.* (32) (tree "GBMB"), and estimated λ and μ over the last 100 million years using two alternative approaches (methods details in Supplement S.11). In the first approach, we fitted generic piecewise-linear curves for λ and μ, similarly to the previous example. In the second approach, we fitted parameterized time curves for λ and μ that included an exponential as well as a polynomial term (5). Even though both approaches yielded similar estimates for $r_{\rm p}$, and both accurately reproduced the tree's LTT, they yielded markedly different λ and μ (Extended data figures 5d–f). This observation is consistent with our argument that, depending on the precise set of models considered, the estimated λ and μ will generally be a random pick from the underlying (true or close-to-true) congruence class.

To illustrate our point that common model selection approaches such as minimizing the Akaike Information Criterion (AIC) (11) cannot resolve the identifiability issues discussed, we also fitted a series of models of variable complexity to a massive timetree of 1,000,000 tips. The tree was simulated based on origination and extinction rates of marine invertebrate genera, previously estimated from marine invertebrate fossil data (33) (Fig. 2D in the main article). We fitted two types of models: piecewise constant models and piecewise linear models. In piecewise constant models (sometimes also referred to as "birth-death-shift" models; 17) the rates λ and μ have constant values within discrete time intervals, with every time interval exhibiting distinct λ and μ. In piecewise linear models λ and μ vary linearly between discrete time points. For both model types we considered various temporal grid sizes, ranging from 5 up to 15 grid points, thus including sufficient model complexity for approximating the true rates. In all cases the time grid points where located at equidistant intervals between the present and the tree's root. For each model type (piecewise constant or piecewise linear) and for each grid size we estimated the free parameters (either the rates within each interval, or the rates at each grid point, respectively) via maximum likelihood using the function `fit_hbd_model_on_grid` in the R package `castor`. Only the most recent 100 million years were considered for fitting, in order to focus estimations on times with greater lineage density in the simulated tree (towards the root estimated rates will be inaccurate regardless of the arguments presented in this paper). Fitting was repeated 20 times with random start parameters to avoid local non-global optima. Among each model type, we then kept the maximum likelihood model with smallest AIC value, shown in Extended data figure 6. As expected, estimated rates were highly inaccurate and missed important features, despite the fact that we were using a massive tree of 1,000,000 tips, and the fact that the tree's LTT was almost perfectly matched by the models' dLTTs.

S.11 Fitting birth-death models to seed plants

An extant timetree of 79,874 seed plant species, constructed using GenBank sequence data with a backbone provided by Magallón *et al.* (34), was obtained from the Supplemental Material published by Smith *et al.* (32, tree "CBMB"). The sampling fraction was calculated based on the tree size and the number of extant seed plant species estimated at 422,127 by Govaerts (35). As mentioned in Supplement S.10, two approaches were used to fit λ and μ over time. In the first approach, λ and μ were allowed to vary independently at 8 discrete and equidistant time points (assuming piecewise linearity between grid points) and were estimated via maximum-likelihood using the function `fit_hb_model_on_grid` in the R package `castor` (31) (options "`condition='stem', relative_dt=0.001`"). Fitting was repeated 100 times using random start parameters to avoid local non-global optima in the likelihood function. The PDR was then estimated from the fitted λ and μ using the formula in Eq. (1) and using central finite differences to calculate derivatives on the time grid. In the second approach, λ and μ were assumed to be of the following general functional forms:

$$\lambda(\tau) = \max\left(0, p_1 \cdot e^{-p_2 \cdot \tau/\tau_\mathrm{r}} + p_3 + p_4 \cdot \frac{\tau}{\tau_\mathrm{r}} + p_5 \cdot \left(\frac{\tau}{\tau_\mathrm{r}}\right)^2 + p_6 \cdot \left(\frac{\tau}{\tau_\mathrm{r}}\right)^3 + p_7 \cdot \left(\frac{\tau}{\tau_\mathrm{r}}\right)^4\right) \tag{94}$$

$$\mu(\tau) = \max\left(0, q_1 \cdot e^{-q_2 \cdot \tau/\tau_\mathrm{r}} + q_3 + q_4 \cdot \frac{\tau}{\tau_\mathrm{r}} + q_5 \cdot \left(\frac{\tau}{\tau_\mathrm{r}}\right)^2 + q_6 \cdot \left(\frac{\tau}{\tau_\mathrm{r}}\right)^3 + q_7 \cdot \left(\frac{\tau}{\tau_\mathrm{r}}\right)^4\right), \tag{95}$$

where τ_r is the age of the root and $p_1, .., p_7, q_1, .., q_7$ are fitted parameters. Parameters were fitted using the `castor` function `fit_hbd_model_parametric` (options "`condition='stem', relative_dt=0.001, param_min=-100, param_max=100`"). As in the first approach, fitting was repeated 100 times to avoid local non-global optima. In both approaches, the likelihood only incorporated branching events at ages between 0 and 130 Myr, since the LTT and any parameter estimates become much less reliable at older ages.

References

[1] D. G. Kendall, On some modes of population growth leading to RA Fisher's logarithmic series distribution. *Biometrika* **35**, 6–15 (1948).

[2] P. H. Harvey, R. M. May, S. Nee, Phylogenies without fossils. *Evolution* **48**, 523–529 (1994).

[3] S. Nee, R. M. May, P. H. Harvey, The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **344**, 305–311 (1994).

[4] T. Kubo, Y. Iwasa, Inferring the rates of branching and extinction from molecular phylogenies. *Evolution* **49**, 694–704 (1995).

[5] H. Morlon, T. L. Parsons, J. B. Plotkin, Reconciling molecular phylogenies with the fossil record. *Proceedings of the National Academy of Sciences* **108**, 16327–16332 (2011).

[6] A. Lambert, T. Stadler, Birth–death models and coalescent point processes: The shape and probability of reconstructed phylogenies. *Theoretical Population Biology* **90**, 113–128 (2013).

[7] S. Louca, *et al.*, Bacterial diversification through geological time. *Nature Ecology & Evolution* **2**, 1458–1467 (2018).

[8] P. Kotelenez, *Stochastic Ordinary and Stochastic Partial Differential Equations: Transition from Microscopic to Macroscopic Equations*, Stochastic Modelling and Applied Probability (Springer New York, 2007).

[9] D. G. Kendall, *et al.*, On the generalized "birth-and-death" process. *The Annals of Mathematical Statistics* **19**, 1–15 (1948).

[10] V. Climenhaga, A. Katok, *From Groups to Geometry and Back*, Student Mathematical Library (American Mathematical Society, Providence, Rhode Island, USA, 2017).

[11] H. Akaike, Likelihood of a model and information criteria. *Journal of Econometrics* **16**, 3–14 (1981).

[12] G. Schwarz, Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464 (1978).

[13] D. L. Rabosky, Likelihood methods for detecting temporal shifts in diversification rates. *Evolution* **60**, 1152–1164 (2006).

[14] D. L. Rabosky, Laser: a maximum likelihood toolkit for detecting temporal shifts in diversification rates from molecular phylogenies. *Evolutionary Bioinformatics* **2** (2006).

[15] D. L. Rabosky, I. J. Lovette, Explosive evolutionary radiations: decreasing speciation or increasing extinction through time? *Evolution: International Journal of Organic Evolution* **62**, 1866–1875 (2008).

[16] D. Silvestro, J. Schnitzler, G. Zizka, A bayesian framework to estimate diversification rates and their variation through time and space. *BMC Evolutionary Biology* **11**, 311 (2011).

[17] T. Stadler, Mammalian phylogeny reveals recent diversification rate shifts. *Proceedings of the National Academy of Sciences* **108**, 6187–6192 (2011).

[18] H. Morlon, Phylogenetic approaches for studying diversification. *Ecology Letters* **17**, 508–525 (2014).

[19] M. A. McPeek, The ecological dynamics of clade diversification and community assembly. *The American Naturalist* **172**, E270–E284 (2008).

[20] M. E. Steeman, *et al.*, Radiation of extant cetaceans driven by restructuring of the oceans. *Systematic Biology* **58**, 573–585 (2009).

[21] C. R. Marshall, Five palaeobiological laws needed to understand the evolution of the living biota. *Nature Ecology & Evolution* **1**, 165 (2017).

[22] P. J. Mayhew, G. B. Jenkins, T. G. Benton, A long-term association between global temperature and biodiversity, origination and extinction in the fossil record. *Proceedings of the Royal Society B: Biological Sciences* **275**, 47–53 (2008).

[23] J. A. Esselstyn, R. M. Timm, R. M. Brown, Do geological or climatic processes drive speciation in dynamic archipelagos? the tempo and mode of diversification in southeast asian shrews. *Evolution* **63**, 2595–2610 (2009).

[24] A. N. Egan, K. A. Crandall, Divergence and diversification in North American Psoraleeae (Fabaceae) due to climate change. *BMC Biology* **6**, 55 (2008).

[25] J. L. Cantalapiedra, *et al.*, Dietary innovations spurred the diversification of ruminants during the Caenozoic. *Proceedings of the Royal Society B: Biological Sciences* **281**, 20132746 (2014).

[26] F. L. Condamine, J. Rolland, H. Morlon, Macroevolutionary perspectives to environmental change. *Ecology Letters* **16**, 72–85 (2013).

[27] T. Stadler, On incomplete sampling under birth–death models and connections to the sampling-based coalescent. *Journal of Theoretical Biology* **261**, 58–66 (2009).

[28] H. Morlon, M. D. Potts, J. B. Plotkin, Inferring the dynamics of diversification: A coalescent approach. *PLOS Biology* **8**, e1000493 (2010).

[29] T. Stadler, How can we improve accuracy of macroevolutionary rate estimates? *Systematic Biology* **62**, 321–329 (2013).

[30] T. Stadler, M. Steel, Swapping birth and death: Symmetries and transformations in phylodynamic models. *Systematic Biology* **68**, 852–858 (2019).

[31] S. Louca, M. Doebeli, Efficient comparative phylogenetics on large trees. *Bioinformatics* **34**, 1053–1055 (2018).

[32] S. A. Smith, J. W. Brown, Constructing a broadly inclusive seed plant phylogeny. *American Journal of Botany* **105**, 302–314 (2018).

[33] J. Alroy, Dynamics of origination and extinction in the marine fossil record. *Proceedings of the National Academy of Sciences* **105**, 11536–11542 (2008).

[34] S. Magallón, S. Gómez-Acevedo, L. L. Sánchez-Reyes, T. Hernández-Hernández, A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytologist* **207**, 437–453 (2015).

[35] R. Govaerts, How many species of seed plants are there? *Taxon* **50**, 1085–1090 (2001).