# Ubiquitous nonidentifiability of state-dependent diversification models (SSEs) is not problematic for phylogenetics

**Sergei Tarasov**[a,b,c,1] **and Josef Uyeda**[c]

[a]Finnish Museum of Natural History, Pohjoinen Raubtatiekatu 13, FI-00014 Helsinki, Finland; [b]National Institute for Mathematical and Biological Synthesis (NIMBioS), TN, USA; [c]Department of Biological Sciences, Virginia Tech 4076 Derring Hall 926 West Campus Drive Blacksburg, VA 24061, USA

This manuscript was compiled on March 5, 2024

**A recent study (Louca and Pennell, 2020) spotlighted the issue of model congruence, or asymptotic unidentifiability, in time-dependent birth-death models used for reconstructing species diversification histories on phylogenetic trees. The phenomenon of model congruence implies that any given timetree can be equally likely explained by various diversification scenarios. The present work investigates this issue in state-dependent speciation and extinction (SSE) models, which are widely employed to study trait-dependent diversification. We demonstrate that model unidentifiability is universal for SSEs, with every SSE belonging to an infinite class of congruent models. Importantly, we find that any trait-independent model is congruent with a trait-dependent model that possesses the same or fewer parameters, raising significant concerns for model selection and hypothesis testing. To address this challenge, we propose an analytical solution and discuss its implication for comparative phylogenetics and model development. Our results indicate that while SSE congruence must be carefully considered in diversification studies, it does not require a fundamental change in phylogenetic methodologies.**

Markov models | congruence class | model selection | discrete traits | speciation | extinction

A recent study by Louca and Pennell [LP] (1) raised a critical issue of model congruence occurring in the birth-death models used to reconstruct diversification histories. LP showed that any extant timetree can be explained by countless diversification scenarios with identical likelihood scores, indicating that any timetree is consistent with an infinite set of congruent models that remain unidentifiable regardless of the amount of observed data (2). This discovery has sparked considerable interest in the phylogenetic community, leading to the development of methods to address this issue (3–6).

LP's study was restricted to time-dependent diversification processes, characterized by constant speciation and extinction rates across lineages. They conjectured that similar issues may arise in state-dependent speciation and extinction (SSE) models, where diversification rates are allowed to vary among lineages [e.g., BiSSE, HiSSE, (7, 8)]. A recent study proved the identifiability of a particular subset of pure-birth SSE models (9), while another study proposed that SSE models are generally identifiable, albeit without any proof (10).

In this study, we demonstrate that model congruence is a universal phenomenon across all SSE models. To illustrate this, we employ a general model, HiClaSSE, which comprises commonly used SSEs (e.g., BiSSE, HiSSE, MuSSE, and ClaSSE) as its special cases. We find that congruence manifests in models that adhere to the lumpability criteria of Markov chains (11), presenting a scenario where a model with or without hidden states is congruent with an infinite set of models that include hidden states. This infinite set forms a congruence class where models may have drastically different diversification histories, and parameter counts, ranging from, at least two, to infinity. For example, each state-independent model is congruent with a model with the same or fewer parameters but with trait-dependent diversification. As noted by LP, such congruence poses a problem for model selection and hypothesis testing. We propose a solution that minimizes the number of hidden states in the optimal model, ensuring its uniqueness, and enabling efficient model selection among congruent classes. Importantly, our solution is available analytically and should not broadly affect traditional model selection practices in state-dependent diversification. We discuss the implications of this solution below.

Our findings suggest that the congruence identified in SSE models points towards straightforward procedures for model selection and evaluation in comparative phylogenetics, though it does raise important interpretational challenges as well as opportunities. Crucially, we conjecture that SSE models do not contain any problematic congruent classes as those identified by LP in birth-death models. In other words, the SSE congruence should be carefully considered when studying diversification, but does not necessitate a shift in phylogenetic practices. We support our conclusions with simulations using likelihood inferences with HiClaSSE model.

## Results

**HiClaSSE and hidden states.** The HiClaSSE model extends ClaSSE (12) and BiSSE-ness (13) by including hidden states, as in HiSSE (8). In addition to speciation and extinction events (7), HiClaSSE allows modeling two types of state-change events. The cladogenetic events which denote state transitions occurring with lineage bifurcation during speciation, and anagenetic events which denote state transitions without speciation. For simplicity, speciation events with and without state changes are termed coupled and decoupled speciations, respectively. HiClaSSE may include any number of states $k$.

Like other SSEs, HiClaSSE is a joint process that combines an observable trait and a diversification component. The trait represents a Continuous Time Markov Chain (CTMC) where discrete states change over time. The diversification component is a separate CTMC whose states are unobservable diversification regimes, each equipped with a specific specia-

The authors declare no competing interests.

[1]To whom correspondence should be addressed. E-mail: sergei.tarasov@helsinki.fi

tion and extinction rate to control lineage birth and death. Thus, in an SSE, every state is effectively hidden, belonging simultaneously to the trait state and the diversification regime. Observable states are modeled at the tips of tree using ambiguous character coding, they are the collections of the hidden states. For example, in a four-state HiSSE, each observable state may comprise two hidden states, while in a trait-free MiSSE (14), that models diversification shifts alone, all hidden states are mapped onto the single observable one. This framework allows for the construction of various SSE models. Note, our interpretation of hidden states, for biological clarity and mathematical precision, may differ from previous treatments (SI Appendix).

**HiClaSSE likelihood.** As for other SSE models (7, 15), the likelihood is calculated by solving a system of ODEs backward in time using two terms: $D_i(t)$, the probability of a lineage being in state $i$ at time $t$, and $E_i(t)$, the probability of a lineage in state $i$ not surviving until the present time. Representing them as column vectors, $\boldsymbol{D} = [D_1(t), \ldots, D_k(t)]$ and $\boldsymbol{E} = [E_1(t), \ldots, E_k(t)]$, yields the following ODEs expressed in array notation:

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{D}}{\mathrm{d}t} &= -\left(\mathbf{1}^T\boldsymbol{\Lambda}\mathbf{1} + \boldsymbol{\mu}\right) \circ \boldsymbol{D} + Q\boldsymbol{D} + \mathbf{1}^T\left(\tilde{\boldsymbol{\Lambda}} \circ \mathrm{I}_{2,1} \circ (\boldsymbol{D} \otimes \boldsymbol{E}^T)\right)\mathbf{1},\\ \frac{\mathrm{d}\boldsymbol{E}}{\mathrm{d}t} &= \boldsymbol{\mu} - \left(\mathbf{1}^T\boldsymbol{\Lambda}\mathbf{1} + \boldsymbol{\mu}\right) \circ \boldsymbol{E} + Q\boldsymbol{E} + \mathbf{1}^T\left(\tilde{\boldsymbol{\Lambda}} \circ \mathrm{I}_{1,\frac{1}{2}} \circ (\boldsymbol{E} \otimes \boldsymbol{E}^T)\right)\mathbf{1},\end{aligned} \quad [1]$$

where, "$\circ$" is an element-wise product of two arrays, $\boldsymbol{D} \otimes \boldsymbol{E}^T$ is outer product of two vectors, and $\mathrm{I}_{a,b}$ indicates square matrix where all diagonal elements are $a$'s while off-diagonals are $b$'s.

The speciation tensor $\boldsymbol{\Lambda} = \{\Lambda_1, \ldots, \Lambda_k\}$ contains $k \times k$ matrices $\Lambda_i$ with speciation rates ($\lambda_{ijz}$). For example, the notation $\lambda_{111}$ indicates a decoupled speciation event – when one lineage in state 1 specialties into two lineages with the same state. In contrast, the notation $\lambda_{122}$ indicates a coupled speciation – when one lineage in state 1 produces two lineages each with the state 2 (12, 15).

The extinction rates of a lineage are denoted by the vector $\boldsymbol{\mu} = [\mu_1(t), \ldots, \mu_k(t)]$. The instantaneous rate matrix $Q$ defines anagenetic transitions ($q_{ij}$) within a CTMC that governs the joint evolution of the trait states and diversification regimes (details in SI Appendix). HiClaSSE is a general model that includes various SSEs as submodels. For instance, HiClaSSE with two states and no cladogenetic events is BiSSE model (7).

**Correlated evolution in SSE models.** In SSEs, the parameters determine whether the diversification component and the trait evolve independently or in correlation. The parameterizations corresponding to trait-dependent diversification have been identified initially for simple models (e.g. BiSSE, ClaSSE). In these models, trait-dependence occurs when: (**C1**) there is variation in speciation and/or extinction rates between observable states (7); or (**C2**) a model includes cladogenetic events indicating that a speciation is coupled with a trait change (12, 13). However, this list is not exhaustive. It is often overlooked that SSEs with hidden states may introduce additional trait-dependent scenarios. While these conditions were initially formulated for binary traits (16), they can be readily extended to all SSEs (SI Appendix).

In general, these additional scenarios are applicable when condition (**C1**) is met, and they are observed in the following situations: (**C3**) when the rate of change in the diversification regime depends on the trait, (**C4**) when changes in both the trait and regime are interdependent, and (**C5**) when changes in the trait and regime occur simultaneously, typically when dual state transitions are allowed.

**Lumpability Conditions for SSE models.** If multiple states in an original CTMC can be aggregated into fewer states and the aggregated process remains Markovian, then the original CTMC is called "lumpable" for a given state partition (11). Lumping enables the creation of congruent models (see the next section), and applies generally to CTMCs on phylogenetic trees (17).

Lumpability occurs when specific symmetries in transition rates exist. Specifically, a CTMC is lumpable if, for any two aggregated states $\hat{s}_i$ and $\hat{s}_j$, the total transition rate $\hat{q}_{ij}$, from any original state $s_m$ in $\hat{s}_i$ to $\hat{s}_j$ remains constant. These $\hat{q}_{ij}$'s become the transition rates in the lumped CTMC (11).

SSE models, which have a greater complexity due to their diversification component, can also be lumped into another SSE with fewer states. By applying the lumpability principles for a CTMC (11) to the ODEs in Eq. 1, we have derived such criteria (SI Appendix).

An SSE model is lumpable if three conditions are met simultaneously: (**L1**) the rate matrix $Q$ must be lumpable, as previously specified for a CTMC; (**L2**) all $\mu$'s within each partition set must be equal, and the lumped extinction rate is that value; (**L3**) the speciation tensor presents a more complex situation. In this case, each lumped speciation rate $\hat{\lambda}_{abc}$ is composed of the original rates $\lambda_{ijz}$ which can be grouped based on the parent branch state (i.e., $i$). So, the speciation tensor is lumpable if the sum of the original $\lambda$'s in each group is the same. The lumped speciation rate is one sum that satisfies this condition.

**Lumpability and Model Congruence.** If within an observable state, several or all hidden states are lumped together, it reduces the the total number of states in an SSE, without altering the observable ones. Consequently, the original and lumped models are congruent (denoted by "$\cong$") and phylogenetically indistinguishable, as they share identical observable state spaces and likelihoods (SI Appendix).

As an example, consider the trait-free MiSSE model ($M_1$ in Eq. 2) with three parameters: two speciation rates $\lambda_{AAA}$ and $\lambda_{BBB}$, a single transition rate $q$ between the diversification regimes, and a constant extinction rate $\mu$. This model is congruent with another model, $M_2$ (Eq 2), in which there are three diversification regimes, but the speciation rates for the regimes $B$ and $C$ are the same (i.e., $\lambda_{BBB} = \lambda_{CCC}$). Clearly, $M_2$ is lumpable given the state partition $\{\{A\}, \{B, C\}\}$ as it meets the lumpability conditions (**L1-3**). Since all states in $M_2$ are hidden, it collapses to $M_1$ under the lumping.

$$M_1 = \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{pmatrix} - & q \\ q & - \end{pmatrix} \end{array}, M_2 = \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \left(\begin{array}{c|cc} - & q & 0 \\ \hline q & - & q \\ q & 0 & - \end{array}\right) \end{array}, M_3 = \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \left(\begin{array}{c|cc} - & q & 0 \\ \hline 0 & - & q \\ q & 0 & - \end{array}\right) \end{array}. \quad [2]$$

**Table 1. The independent and dependent SSEs and their congruence to CID4 with four hidden states.**

| Model | np | pars | $\cong$ | Trait & Div. | Speciation |
|---|---|---|---|---|---|
| CID4 | 4 | $2\lambda_d\ 1q\ 1\mu$ | – | independent | decoupled ($\lambda_d$) |
| CID8 | 4 | $2\lambda_d\ 1q\ 1\mu$ | yes | independent | decoupled ($\lambda_d$) |
| COR8-C | 4 | $2\lambda_d\ 1q\ 1\mu$ | yes | correlated | decoupled ($\lambda_d$) |
| CLA8-C | 4 | $2\lambda_c\ 1q\ 1\mu$ | yes | correlated | coupled ($\lambda_c$) |
| EHE8-C | 2 | $1(\lambda_{c+d}{+}q)\ 1\mu$ | yes | correlated | both ($\lambda_{c+d}$) |
| COR8-NC | 4 | $2\lambda_d\ 1q\ 1\mu$ | no | correlated* | decoupled ($\lambda_d$) |
| CLA8-NC | 4 | $2\lambda_c\ 1q\ 1\mu$ | no | correlated | coupled* ($\lambda_c$) |

These SSEs differ in parameter count (np), number of parameters across model components (pars), the congruence relation to CID4 ($\cong$), trait-dependent diversification (Trait & Div), and the type of speciation rates. * indicates non-lumpable transition or speciation rates. See also, Fig. 3 and Table S1.

In contrast, a slightly different model $M_3$ (Eq. 2) is not congruent with either $M_1$ or $M_2$, as its rate matrix violates the lumpability condition (**L1**) for the partition $\{\{A\},\{B,C\}\}$ because the transitions $B \to A$ and $C \to A$ are not equal. Therefore, despite all three models in Eq. (2) have the same parameter count, $M_3$ stands as a distinct model in its own right. Note, the states of $M_3$ cannot be reduced to a two-state model. While sharing the same speciation rates as $M_1$, $M_3$ displays distinct dynamics in its diversification regimes. Consequently, the pattern of speciation and extinction events on a phylogeny will be different too.

**Hidden Expansion (HE).** In the example above, we illustrated how adding hidden states to model creates a congruent model, a process termed "hidden expansion" (HE). While our example above is a fairly trivial HE, this technique can be applied universally to any SSE to generate a congruent model, including those mapping to biologically distinct interpretations–which we will demonstrate in the next sections. For instance, $M_2$ can be further expanded, maintaining lumpability, to construct models with any number of hidden states. Given the limitless possibilities for maintaining lumpability, countless implementations of HE exist. Therefore, every SSE model is part of an infinite congruence class, a phenomenon we define as HE-congruence. Also, this congruence holds under any sampling fraction which specifies the estimated proportion of extant species (SI Appendix).

**Congruence between independent and dependent SSEs.** An SSE with trait-independent diversification is always congruent with a model implying dependence. Take, for example, the CID4 model, which represents a trait-independent diversification with two regimes (8). Assume the model has four parameters: two speciation rates, a single rate for state transitions, and a constant extinction (Table 1; Fig. 3; SI Appendix). Note, the numeral in the model denotes the count of hidden states.

Similar to the $M_1$ (Eq. 2), we can apply HE to CID4 by adding extra states to diversification regimes. Let us assume that this results in CID8 model, which is congruent to CID4 and has identical parameter count (Table 1; SI Appendix).

Now, we can begin modifying CID8 but keeping it lumpable with CID4. One way to achieve this is by enabling dual transitions between the trait and diversification regimes. This transformation results in a model with a trait-correlated diversification, which we call COR8-C. Yet COR8-C remains congruent with CID4 (Table 1; SI Appendix). To illustrate that this congruence is solely due to lumpability, we can invalidate it by prohibiting certain (but not all) dual transitions. This creates a non-congruent model COR8-NC that maintains the correlation but displays distinct diversification dynamics; it is state space is irreducible (Fig. 3; Table S1).

A different congruent model can be created by disabling all decoupled speciation events ($\lambda_{xxx}$'s) and permitting coupled events that adhere to lumpability principles. This transformation results in CLA8-C, which is similar to COR8-C, but all speciation events coincide with trait changes (Table 1; Fig. 3; SI Appendix). As before, this congruence occurs solely due to lumpability. If, instead, we allow "non-lumpable" cladogenetic events, we obtain the non-congruent model, such as CLA8-NC, which has distinct dynamics but similar biological interpretation with CLA8-C (Table 1; SI Appendix).

HE can generate congruent and correlated models from any CID model; its parameterization determines the state count of a congruent model. The resulting congruence class is also infinite. For instance, we can expand CID4 to CID16 instead of CID8, and make similar modifications as we did with CID8 to obtain COR8-C and CLA8-C.

**Equal Rate Hidden Expansion (EHE).** HE can be extended to a point when all permissible rates are equal, except extinctions. We refer to this transformation as Equal Rate Hidden Expansion (EHE).

Consider, for instance, a CID4 model with four parameter values: $\lambda_1 = 0.3$, $\lambda_2 = 0.1$, $q = 0.2$, and $\mu = 0.01$. Applying the EHE to this model results in the EHE8-C, where the diversification is trait-dependent, and coupled speciation events are allowed (Table 1). The EHE8-C has only two parameters compared to the four in the CID4; and it is akin to CLA8-C but all its $\lambda$'s and $q$'s have the same value of 0.1 (SI Appendix), while extinction rate remains unchanged ($\mu = 0.01$).

EHE transformation typically includes coupled speciation events along with trait-dependent $Q$ matrix. It reduces the number of parameters to the number of extinction categories plus one. The EHE can be applied to any SSE model, regardless of the state count or rate values. This universality arises from the properties of rational numbers, where any rational number can be expressed as the product of an integer $z$, and a rational multiplier $r$. If the rates are not rational, they can be approximated to the desired precision. As a result, the set of parameters ($\theta$'s) in an original SSE, excluding extinction, expands as $\{\theta_1 = z_1 r,\ \theta_2 = z_2 r,\ \ldots\ \theta_n = z_n r\}$. The common factor $r$ serves as the equal rate, while the integers $z_i$ represent additional rate terms in the EHE. Therefore, an SSE model that is sufficiently large to embed these additional rates represents the EHE transformation. We present a general algorithm that can be used to create EHE models (SI Appendix).

For a given SSE model, EHE generates an infinite congruence class because there is an infinite set of multipliers $r$ to use for the same original model. The choice of $r$ determines the state count in the EHE model.

**Semi-Congruent Behavior.** A universal parametric form of EHE for a given SSE does not exist because the number of hidden states in EHE depends on the original rate values

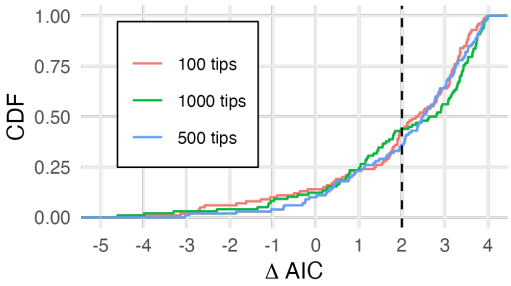


**Fig. 1.** Semi-congruent behavior shown using the cumulative distribution function (CDF) of ΔAIC. Positive values indicate support for semi-congruent EHE8-C model over the true model CID4. The simulations were performed for phylogenetic trees with 100, 500, and 1000 tips.

(SI Appendix). For example, varying parameter values in the CID4 from the previous section would result in different EHE models.

Consider fitting the CID4 model, from the previous example, to data that were originally generated by it. The ML estimates might be: $\lambda_1 = 0.336$, $\lambda_2 = 0.138$, $q = 0.229$, and $\mu = 0.0139$. When applying an EHE transformation to them, it would require thousands of hidden states due to the many decimals, which is present in all ML estimates. However, instead of using the true generating EHE model, let us fit EHE8-C (Table 1) that would be congruent if we round the estimates to one decimal. While EHE8-C is expected to yield a worse likelihood, it might be preferred by the Akaike Information Criterion (AIC) due to having fewer parameters. This exemplifies semi-congruent behavior, which may occur when there is no exact match between the SSE and EHE models.

Our simulations show that semi-congruence can lead to flawed model selection, regardless of the amount of data. For example, in 55-64% of the simulations, a semi-congruent EHE8-C received positive support ($\Delta\text{AIC} > 2$) over CID4, despite CID4 constantly having a better likelihood (Fig. 1).

For any SSE, a semi-congruent model can be constructed by reducing the respective EHE. The extent to which AIC favors a semi-congruent model over the original one depends on their parameter difference. The maximum possible AIC support ($AIC_m$) for the semi-congruent model is deterministic and calculates as: $AIC_m = 2(p_1 - p_2)$, where $p_1$ and $p_2$ represent the number of parameters in the compared models with $p_1 \geq p_2$.

**The Structure of a Congruence Class.** Each congruence class, generated by HE, consists of a unique non-lumpable model called the irreducible model ($M_0$) and an infinite set of lumpable models created by adding hidden states to $M_0$ (Fig. 2). Thus, all other models in the class have more hidden states than $M_0$. If $M_0$ already has hidden states, it represents a process that cannot be further reduced. For instance, CID4 and BiSSE are irreducible models with and without hidden states, respectively. In addition to the HE explained earlier, which lead to congruent models with the same or fewer parameters, it is straightforward to create models with more parameters.

We identified three main types of HE: sub-expansion (e.g,. EHE), equivalent (e.g, CID8, COR8-C, CLA8-C), and super-expansion. These types create models with the fewer, same, or more parameters in the off-diagonal blocks of the $Q$ matrix,

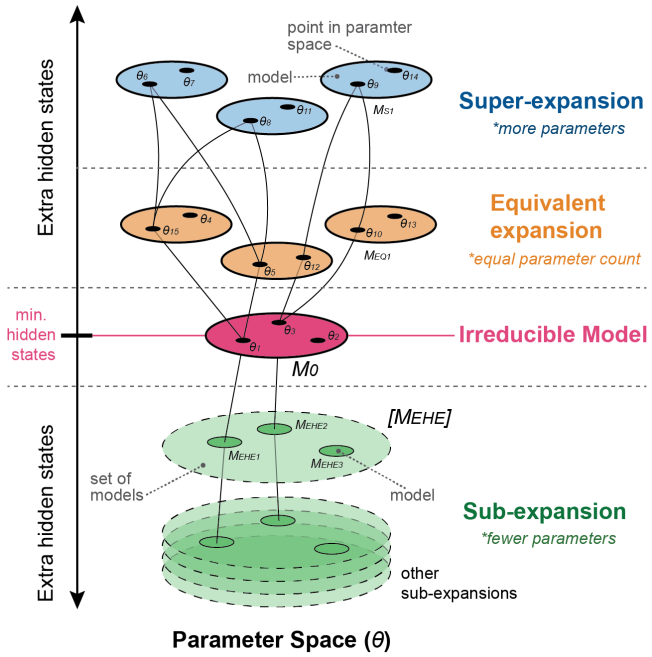


**Fig. 2.** Structure of congruence class generated by HE. The irreducible model $M_0$ is congruent with all equivalent and super-expansion models within the class (e.g, $M_0 \cong M_{EQ1} \cong M_{S1}$). Only specific parameter values of $M_0$ are congruent with sub-expansion models (e.g, $M_0(\theta_1) \cong M_{EHE1}$), while the entire $M_0$ is congruent with a set of sub-expansion models (e.g, $M_0 \cong [M_{EHE}]$). Note, congruence relationships are transitive, not all of them are shown for clarity; * denotes the parameter count for off-diagonal parameters in $Q$ (SI Appendix).

respectively. The off-diagonal parameters are essential for defining the parametric form (SI Appendix).

Sub-expansion models depend on specific parameter values in $M_0$. Each parameter setup corresponds to a distinct sub-expansion model, usually with fewer parameters than $M_0$. $M_0$ is not directly congruent with individual sub-expansion models but rather with sets of these models (e.g., $M_0 \cong [EHE]$).

## Discussion

By using the general HiClaSSE model, we demonstrated that any SSE is a member of a congruence class, indicating that there is an infinite set of models which always produce identical likelihoods irrespective of the amount of data. Thus, the congruence reported by LP in time-dependent models (1), is also ubiquitous in SSEs. This finding contradicts to the previous conclusion that SSEs have no congruence classes (10).

We have identified hidden expansion (HE) as the primary cause of SSE congruence. This congruence arises in models with hidden states that meet specific lumpability conditions and renders these models congruent with other models with or without hidden states. HE reveals an important source of uncertainty in the construction of SSEs – to what extent do the observable character states represent the "true" architecture of the underlying complexity of the evolutionary process? Hidden states can map on to a variety of unobserved phenomena, including diversification regimes, latent developmental, ecological, or genetic states or even observer biases in the construction of a character (see the "Two-Scientist Paradox", (17, 18)). Such challenges are familiar to biologists seeking to

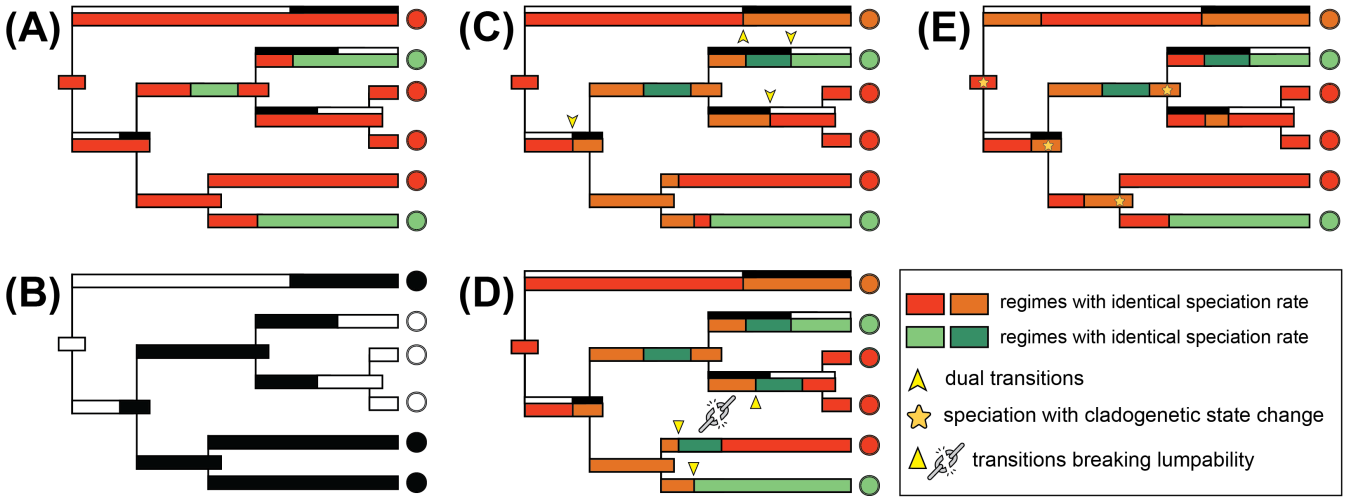


**Fig. 3.** Alternative evolutionary scenarios (Table 1). (A-B) The CID4 model illustrates the independent evolution of the diversification process with two distinct regimes (A), and the binary trait (B). The models in (C-D) are trait-dependent; they have four diversification regimes and two speciation rates. The COR8-C (C) and CLA8-C (E) models are congruent with CID4, while COR8-NC (D) is not congruent. In COR8-C, the trait-dependence arises due to dual transitions. The CLA8-C model is similar to COR8-C, but all speciation events are coupled with trait changes. The COR8-NC model is also similar to COR8-C but transitions between the orange and red diversification regimes are disallowed, resulting in a non-congruent model.

classify the complexity of a highly variable complex character into discrete categories for analysis by SSE models. With the increasing usage of hidden state models in macroevolutionary research, and the vast range of new models hidden expansion reveals, understanding the structure of congruence becomes essential for interpreting the results of SSE model selection.

HE congruence classes are akin to those found by LP, as both share a similar invariance in their rates. Namely, the pulled speciation and diversification rates in LP classes are similarly invariant as the distribution of individual rates with respect to observable states in HE classes. Note, this congruence also holds in models for ancestral state reconstruction alone, as they are submodels of SSEs with no diversification.

We showed that congruent models can vary in parameter count from, at least, two (sub-expansions) to infinity (super-expansions), and imply alternative evolutionary scenarios. In particular, any trait-independent CID model is congruent with an infinite set of trait-dependent models with equal or fewer parameters. Similarly, there are some trait-dependent models that will not be identifiable from a generating model that is independent since congruence is symmetric. This raises the fundamental question of whether we should rethink the suitability of SSE models for testing trait-dependent diversification hypothesis in comparative phylogenetics.

**Conventional Model Selection Fails.** Constructing SSE models requires translating biological hypotheses into SSE parameterizations, where the structure of parameters reflects the complexity of the diversification process under study, which is typically unknown beforehand. Therefore, standard phylogenetic practice recommends fitting multiple models and assessing their fit using, for example, ΔAIC or AIC weights (19).

Consider a scenario where data follows a trait-independent diversification process, as in the CID4 model (Table 1, Fig. 3A-B), and a researcher seeks to test it. Upon fitting the CID4, the researcher explores alternative models with with an extra diversification regimes and encounters congruent models suggesting different evolutionary scenarios. They may range from trait-dependent evolution as in COR8-C (Fig. 3C), to a scenario where, in addition to that dependence, all speciation events are coupled with state changes as in CLA8-C (Fig. 3E). As a result of the identical parameter count, these models have the same AIC scores as the CID4, and therefore do not reject the null hypothesis of independence. Instead, they indicate the need for further exploration of the model space.

The researcher examines variants of the discovered models with fewer parameters. This exploration may uncover semi-congruent models (Fig. 1), which, although not perfectly congruent, exhibit improved AIC scores compared to CID due to fewer parameters. Semi-congruent models imply trait-dependent diversification, with support ($AIC_m$) depending on data and parameter differences compared to the original process. When the original process has four or more parameters, it is possible to obtain a semi-congruent model with positive to substantial ΔAIC support. This phenomenon was common, in our simulations, the semi-congruent model was favored in 55-64% of trials. Consequently, better AIC scores lead the researcher to conclude that the diversification is trait-dependent.

Instead of surveying the model space, all congruent models can be derived analytically using the lumpability property as described above. An intriguing example is the EHE transformation, where any SSE can be represented by a congruent model with just two parameters, indicating trait-dependence. EHE models may include thousands of hidden states, which makes them computationally infeasible in practice. However, our analytical approaches allows their derivation without fitting them directly to data. In present example, it only requires parameter estimates from the CID model (SI Appendix). Consequently, regardless of the researcher's findings, there will always be an EHE model with a lower AIC. Thus, AIC fails to select the true scenario from a set of congruent models.

**Solution to model selection.** We propose applying the parsimony principle when selecting among congruent models, favor-

ing a model that explains rate variation with minimal number of hidden states. This leads to one unique solution for each congruence class, known as the irreducible model $M_0$ (Fig. 2). It can be found analytically by lumping the states in a reducible model as described above. In the previous example, this model was CID4, and all other (semi-)congruent models can be lumped into it.

The uniqueness of $M_0$ in congruence class renders model selection unambiguous. Thus, the proposed parsimony principle is merely a practical assumption for hypothesis testing with a minimum set of restrictions. It also implies that if the process were evolving under a different model from the congruence class (not $M_0$), we would never be able to uniquely identify it. The proposed penalty aligns with traditional model selection approach. Most SSEs in available software are irreducible and have been successfully used in numerous studies to explain real data. While, the identified congruence does not seem to alter standard practices in phylogenetics significantly, our argument does not imply that other models in the congruence class cannot more accurately reflect the true biological process of diversification. In certain cases, it may be necessary to consider factors other than parsimony in order to construct biologically plausible scenarios, making reducible models within a congruence class more credible than $M_0$.

Our main recommendation is to always use irreducible models for model comparison. When comparing models between different congruent classes, one needs to ensure that they are irreducible. It is trivial if a model has no hidden states because all such models are irreducible. Nevertheless, if using non-standard models, it is always recommended to check whether a smaller irreducible model exists using the lumpability conditions. Various congruent models can be constructed in RevBayes (15, 20), while those with decoupled speciation in HiSSE (8). We did not find any congruent models among the standard models offered by these software.

**The biological validity of SSE congruence.** The congruent models COR8-C, CLA8-C, and EHE8-C may appear odd at first. Unlike CID4, which pairs two diversification regimes with two speciation parameters (Fig. 3A), in these models some or all regimes share identical speciation parameters, potentially suggesting unnecessary redundancy in hidden states (Fig. 3C-D). This configuration is typical for congruent models, raising questions about their biological rationale. Are the congruent models merely theoretical constructs that may obscure comparative phylogenetics? We argue they are not.

Our MiSSE example shows that increasing number of hidden states, while retaining the number of speciation parameters produces different models with different evolutionary dynamics (e.g., $M_1$ vs. $M_3$ in Eq. 2). This finding indicates that the potential range of the MiSSE models is significantly larger than initially proposed (14). By simply increasing state count (i.e., "redundancy"), we can construct models with distinct diversification dynamics. Specifically, such modifications will probabilistically result in different sequences of speciation and extinction events across the phylogenetic tree and likewise change the probabilities of a given character pattern at tips.

This insight extends to other SSEs. Due to trait-dependence, congruent models such as COR8-C and CLA8-C have essentially the same biological interpretation as their non-congruent counterparts such as COR8-NC and CLA8-NC. The non-congruent, COR8-NC and CLA8-NC stand as distinct models because they have no simpler congruent versions and represent the irreducible models in their congruence classes. Despite that, they have the same "redundancy" between the regimes and number of speciation parameters (Fig. 3D), such that their diversification pattern is unique up to the congruence class. Thus, "redundancy" enables the construction of distinct models that capture different aspects of the diversification process. Also, it does not conflict with evolutionary principles, thereby affirming the biological validity of the models with "redundancy". Consequently, the congruent models (COR8-C, CLA8-C, EHE8-C) with "redundancy" are also biologically valid. The unidentifiability of the congruent models with CID emerges merely as a mathematical coincidence due to the lumpability property. If it were possible to explicitly observe state changes throughout phylogenetic history, congruent models would be identifiable.

**Congruence and SSE model space.** The models discussed here may not have been used in any previous study, primarily due to the lack of methods for exploring the SSE model space. In previous example, when the researcher encountered the congruence, the extensive model exploration was merely an assumption. In practice, researchers typically use a few default models offered by available software, even though a thorough examination of the model space is essential.

A fundamental assumption of the ML estimator is that the true stochastic process, which generated the data, must lie within the family of models used to estimate the data (21). Using a model outside this family may fail to capture the true process, and may bias hypothesis testing. Independent and dependent SSEs do not form separate model families due to their complex interrelationships. For instance, there are non-congruent models (e.g., COR8-NC, CLA8-NC) with parametric forms similar to models within the CID4 congruence class (e.g., COR8-C, CLA8-C). This relationship indicates that certain CIDs are more closely related to dependent models than to other CIDs, and this complex interplay applies to all SSEs. Nevertheless, with sufficient data, these models can be distinguished using ML since they belong to different congruence classes.

SSE models are notably prone to high false positive rates in distinguishing between trait-independent and dependent evolution (22). This issue, as illustrated by congruence, can often stem from a violation of the ML fundamental assumption. Specifically, not including an adequate independent model in the analysis can lead to situations where a dependent model, even if wrong, appears to better fit the observed data than the selected independent model. Therefore, examining the full range of SSEs is essential in order to minimize the rate of false positives in model selection. However, this exploration necessitates the development of new methods [e.g., (23)].

HE-congruence emerges as a particularly valuable tool in this context. Firstly, it clarifies the asymptotic unidentifiability – why certain models cannot be distinguished based on available data alone. This insight could help create new SSE models while avoiding unidentifiability pitfalls. Secondly, it enables the differentiation between models that appear distinct yet are fundamentally identical (congruent) and those that are genuinely different (non-congruent). This property is crucial for constructing SSEs and hypothesis testing. Consequently, congruence significantly broadens the range of possible SSE models beyond the conventional scope. It also suggests that

the complex dynamics of diversification processes might not be fully captured by the narrow selection of standard models typically available in existing software.

**Are there other congruence classes?.** Other congruent classes may arise when distinct SSE models are assigned to different clades or tree slices (SI Appendix). Typically, such congruence leads to models with more parameters, and we do not anticipate how this might mislead model selection. Another potential type of congruent models could be caused by weak lumpability (11). It has been demonstrated that weak lumpability does not generally apply to phylogenetic trees (17). However, it is less tractable mathematically, and we cannot exclude the possibility that some specific type of it may be relevant.

Can we expect additional types of congruence for SSE models? A recent study (9) proved the identifiability of pure-birth SSEs with decoupled speciation events. Based on this and our results, we conjecture that problematic congruent classes in SSEs are improbable. Even if new congruent classes are found, they will be unlikely to present challenges for existing practices.

## Materials and Methods

**HiClaSSE and simulations.** The HiClaSSE model is implemented by modifying the functions from the diversitree R package (24).

To demonstrate the semi-congruent behaviour (Fig. 1), we simulated 100 datasets under the CID model ($\lambda_1 = 0.3$, $\lambda_2 = 0.1$, $q = 0.2$, and $\mu = 0.01$) with variable number of tips (100, 500, 1000) using HiSSE package (8). Next, we fitted the original CID model and semi-congruent model that was constructed from the original CID by EHE and has two parameters.

**Data, Materials, and Software Availability.** The software, data and code necessary for reproducing the results and the figures are available on github https://github.com/sergeitarasov/Congruent-SSE-CTMC.

**ACKNOWLEDGMENTS.** ST was supported by the Academy of Finland grant: 339576, and the three-year grant from University of Helsinki; ST partially conducted this work while a Postdoctoral Fellow at the National Institute for Mathematical and Biological Synthesis sponsored by the NSF Award DBI-1300426. JCU was funded by NSF-DEB-1942717 and NSF-DBI-1661516.

1. S Louca, MW Pennell, Extant timetrees are consistent with a myriad of diversification histories. *Nature* **580**, 502–505 (2020).
2. H Morlon, S Robin, F Hartig, Studying speciation and extinction dynamics from phylogenies: addressing identifiability issues. *Trends Ecol. & Evol.* **37** (2022).
3. BT Kopperud, AF Magee, S Höhna, Rapidly changing speciation and extinction rates can be inferred in spite of nonidentifiability. *Proc. Natl. Acad. Sci.* **120**, e2208851120 (2023).
4. S Höhna, BT Kopperud, AF Magee, Crabs: Congruent rate analyses in birth–death scenarios. *Methods Ecol. Evol.* **13**, 2709–2718 (2022).
5. J Andréoletti, H Morlon, Exploring congruent diversification histories with flexibility and parsimony. *Methods Ecol. Evol.* **14**, 2931–2941 (2023).
6. B Legried, J Terhorst, A class of identifiable phylogenetic birth–death models. *Proc. Natl. Acad. Sci.* **119**, e2119513119 (2022).
7. WP Maddison, PE Midford, SP Otto, Estimating a binary character's effect on speciation and extinction. *Syst. biology* **56**, 701–710 (2007).
8. JM Beaulieu, BC O'Meara, Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. *Syst. Biol.* **65**, 583–601 (2016).
9. D Dragomir, ES Allman, JA Rhodes, Parameter identifiability of a multitype pure-birth model of speciation. *J. Comput. Biol.* **30**, 277–292 (2023).
10. B O'Meara, J Beaulieu, Potential survival of some, but not all, diversification methods. *EcoEvoRxiv* (2021).
11. J Kemeny, J Snell, Finite markov chains. princeton (nj): van nostrand (1960).
12. EE Goldberg, B Igić, Tempo and mode in plant breeding system evolution. *Evolution* **66**, 3701–3709 (2012).
13. K Magnuson-Ford, SP Otto, Linking the investigations of character evolution and species diversification. *The Am. Nat.* **180**, 225–245 (2012).
14. T Vasconcelos, BC O'Meara, JM Beaulieu, A flexible method for estimating tip diversification rates across a range of speciation and extinction scenarios. *Evolution* **76**, 1420–1433 (2022).
15. M May, M Xavier, A unified phylogenetic model of state-dependent speciation, extinction and preservation with tree-wide events. *Prepr. https://bitbucket.org/mrmay/tensorphylo/src/master/* **1**, 1–14 (2020).
16. M Pagel, Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond. B* **255**, 37–45 (1994).
17. S Tarasov, Integration of anatomy ontologies and evo-devo using structured markov models suggests a new framework for modeling discrete phenotypic traits. *Syst. biology* **68**, 698–716 (2019).
18. S Tarasov, The invariant nature of a morphological character and character state: insights from gene regulatory networks. *Syst. biology* **69**, 392–400 (2020).
19. DS Caetano, BC O'Meara, JM Beaulieu, Hidden state models improve state-dependent diversification approaches, including biogeographical models. *Evolution* **72**, 2308–2324 (2018).
20. S Höhna, et al., Revbayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Syst. Biol.* **65**, 726–736 (2016).
21. L Held, D Sabanés Bové, Applied statistical inference. *Springer, Berlin Heidelberg, doi* **10**, 16 (2014).
22. DL Rabosky, EE Goldberg, Model inadequacy and mistaken inferences of trait-dependent speciation. *Syst. biology* **64**, 340–355 (2015).
23. M Pagel, A Meade, Bayesian analysis of correlated evolution of discrete characters by reversible-jump markov chain monte carlo. *The Am. Nat.* **167**, 808–825 (2006).
24. RG FitzJohn, Diversitree: comparative phylogenetic analyses of diversification in r. *Methods Ecol. Evol.* **3**, 1084–1092 (2012).

## Supporting Information Text

## HiClaSSE and hidden states

In order to avoid a common misunderstanding about what hidden states are in an SSE model, we provide a theoretical background below to help understand this concept.

**The individual components of an SSE model.** An SSE is a joint model that considers both observable trait evolution and diversification regimes simultaneously. To understand SSE models, including HiClaSSE, it is practical to start with the CID model, where trait and diversification are independent. The CID model is equivalent to two separate models — one for diversification and one for trait component. In other words, CID is a separable model, whose likelihood is the product of the likelihoods of the individual models. Let us consider the simplest such model, CID4, with four states, describing the evolution of a binary trait and binary diversification regime.

The trait component is a traditional CTMC, represented by a rate matrix $Q_t$, with each state being observable (Eq. 1). This model is traditionally used for ancestral state reconstruction of binary traits.

The diversification component is another CTMC characterized by a rate matrix $Q_r$ (Eq. 1), where the regimes are discrete states (A and B) like in the trait's CTMC. Each state is equipped with corresponding speciation $(\lambda_A, \lambda_B)$ and extinction $(\mu_A, \mu_B)$ rates, governing the birth and death of lineages. It is convenient to think of this CTMC from a data generation perspective. When this process generates a phylogenetic tree, it transitions between regimes, and depending on the regime, a lineage can split into two lineages or go extinct according to the speciation and extinction rates. The diversification component alone is known as a trait-free MiSSE model.

$$Q_r = \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{pmatrix} - & r_1 \\ r_2 & - \end{pmatrix} \end{array}, \quad Q_t = \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} - & q_1 \\ q_2 & - \end{pmatrix} \end{array}, \quad Q_j = \begin{array}{c} \begin{array}{cccc} A0 & B0 & A1 & B1 \end{array} \\ \begin{pmatrix} - & r_1 & q_1 & 0 \\ r_2 & - & 0 & q_1 \\ q_2 & 0 & - & r_1 \\ 0 & q_2 & r_2 & - \end{pmatrix} \end{array}. \quad [1]$$

**Joint CID process.** By combining the diversification and trait components, we obtain a four-state transition rate matrix $Q_j$ for the joint process (Eq. 1). Each state in this matrix simultaneously belongs to a diversification regime and a trait state. Each state is associated with speciation rates $(\lambda_A, \lambda_B, \lambda_A, \lambda_B)$ and extinction rates $(\mu_A, \mu_B, \mu_A, \mu_B)$, following the order of states in $Q_j$. As speciation and extinction rates are independent, it implies that pairs of joint states 1A and 2A, as well as 1B and 2B, have the same speciation and extinction rates as well as the $Q$ matrix as in Eq. 1.

**What are the hidden states?.** All states in the joint model $Q_j$ are essentially hidden because we are uncertain about which states our data at the tree tips correspond to. For instance, if we know that a tip is in state 1, we do not know whether it belongs to $1A$ or $1B$. Therefore, the observable states in our model consist of sets of hidden states. For example, an observable state 1 consists of the hidden states $1A$ and $1B$. It is helpful to conceptualize this model as having a two-layered structure: a hidden layer that represents the actual states of the model and an observable layer to which the hidden states are mapped. The evolution takes place exclusively at the hidden layer, and its results are subsequently re-mapped to the observable layer. This two-layered approach serves as a general framework for constructing various SSE models, accommodating a wide range of scenarios. In model formalism, these two layers are modeled using ambiguous coding at the tips of tree.

For likelihood computation, the data at tips should be encoded as probability vectors, instead of integers like in character matrix. For instance, for a binary trait, data at the tips are encoded as vectors: $(1, 0)$ or $(0, 1)$, representing the observed states 0 and 1, respectively. In cases where the tip's state is uncertain, we use an ambiguous coding, such as $(1, 1)$, which implies that both states are possible.

The same approach is adopted by SSEs. For example, the CID model (Eq. 1) requires the data to be encoded as (1,1,0,0) and (0,0,1,1) corresponding to the observable states 0 and 1 respectively. For the two-state MiSSE model (e.g., $Q_r$ in Eq. 1), all tips are encoded as $(1, 1)$, indicating that all states are hidden, which emphasizes our lack of knowledge about the diversification regimes at the tips, which the model aims to infer.

**Correlated SSE models.** The SSE model displaying correlated evolution between trait and diversification can be created from CID by altering rate symmetries in $Q_j$ or by rewiring the association between hidden states and speciation/extinction rates. See the section "Correlated Evolution in SSE Models".

**BiSSE model and hidden states.** Any SSE model, including those without hidden states like the BiSSE model, can be viewed as an instance of the general "hidden state" framework. By modifying the mapping of speciation and extinction rates in the CID model (Eq. 1), we can create a correlated process. For example, we can set the rates as follows: speciation $(\lambda_A, \lambda_A, \lambda_B, \lambda_B)$ and extinction $(\mu_A, \mu_A, \mu_A, \mu_B)$ while keeping $Q_j$ the same. This modified model, which we will call COR4, is congruent to the BiSSE model because the rate matrix of COR4 (i.e., $Q_j$ in Eq. 1) and the diversification rates are lumpable with respect to the state partition $\{\{A0, B0\}, \{A1, B1\}\}$. So, the corresponding BiSSE model is defined by $Q_t$ (Eq. 1) and the diversification rates are: $(\lambda_A, \lambda_B)$, $(\mu_A, \mu_B)$. Although the BiSSE model does not formally have hidden states, it can be viewed as a simplified version of a more complex hidden state model.

## HiClaSSE likelihood

The HiClaSSE description follows the general framework for SSE models. We will use lowercase symbols for scalars (e.g., $a$), italic-bold symbols for column vectors (e.g., $\boldsymbol{a}$), capital letters for matrices (e.g., $A$), and bold capital letters for sets of matrices, called tensors (e.g., $\mathbf{A}$). The element-wise product, denoted as $\circ$, involves multiplying each element of two arrays with the same dimensionality. The notation $\boldsymbol{u} \otimes \boldsymbol{v}^T$ signifies the outer product of two vectors, where $X\boldsymbol{u}$ represents the multiplication of matrix $X$ by vector $\boldsymbol{u}$, and $XY$ indicates the multiplication of matrix $X$ by matrix $Y$.

We consider that HiClaSSE represents a lineage-varying state-dependent birth-death process with $k$ discrete states and the following types of events: (1) speciation events, where a lineage splits into two; (2) extinction events, resulting in the death of a lineage; and (3) state-change events, where a lineage change from one state to another.

**Speciation events.** We denote by $\boldsymbol{\Lambda}$ the speciation array that represents a set of $k$ matrices $\boldsymbol{\Lambda} = \{\Lambda_1, \Lambda_2, \ldots, \Lambda_k\}$ characterizing the rates for speciation events. Each matrix $\Lambda_i$ has dimension $k \times k$, with upper triangular entities populated with $\lambda$ rates (others are zero). Matrix $\Lambda_i$ has entities $\lambda_{ijz}$, which refer to the speciation rates that occur when a lineage in state $i$ speciates and produces two daughter lineages one in state $j$ and another in state $z$:

$$\boldsymbol{\Lambda} = \left\{ \Lambda_1 = \begin{pmatrix} \lambda_{111} & \lambda_{112} & \ldots & \lambda_{11k} \\ 0 & \lambda_{122} & \ldots & \lambda_{12k} \\ & & & \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & 0 & \lambda_{1kk} \end{pmatrix} \Lambda_2 = \begin{pmatrix} \lambda_{211} & \lambda_{212} & \ldots & \lambda_{21k} \\ 0 & \lambda_{222} & \ldots & \lambda_{22k} \\ & & & \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & 0 & \lambda_{2kk} \end{pmatrix} \quad \ldots \quad \Lambda_k = \begin{pmatrix} \lambda_{k11} & \lambda_{k12} & \ldots & \lambda_{k1k} \\ 0 & \lambda_{k22} & \ldots & \lambda_{k2k} \\ & & & \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & 0 & \lambda_{kkk} \end{pmatrix} \right\}. \quad [2]$$

All these events fall in two categories – those simultaneously occurring with state changes and those which do not. We refer to them as coupled (e.g., $\lambda_{111}$) and decoupled (e.g., $\lambda_{122}$) speciation events respectively.

Note that $\Lambda_i$ is intentionally an upper triangular matrix to visualize unique rate entities and avoid rate duplication (as $\lambda_{112} = \lambda_{121}$). However, in some calculations below, we will require a symmetric speciation matrix ($\tilde{\Lambda}_i$) where the zeros in $\Lambda_i$ are filled with rates from the upper triangle. Such matrices can be obtained through the following algebraic operation:

$$\tilde{\Lambda}_i = \Lambda_i + \Lambda_i^T - \mathrm{Diag}(\mathrm{diag}(\Lambda_i)), \quad [3]$$

where, superscript $T$ denotes the transpose of matrix, $\mathrm{diag}(\Lambda_i)$ extracts the diagonal of $\Lambda_i$, and $\mathrm{Diag}(\mathrm{diag}(\Lambda_i))$ forms a diagonal matrix with the diagonal elements of $\Lambda_i$. These symmetric matrices also form an array $\tilde{\boldsymbol{\Lambda}} = \{\tilde{\Lambda}_1, \tilde{\Lambda}_2, \ldots, \tilde{\Lambda}_k\}$.

We also need to define a matrix-wise operation on the matrix arrays ($\boldsymbol{\Lambda}$ and $\tilde{\boldsymbol{\Lambda}}$). For example, a notation $\boldsymbol{\Lambda} \circ X$ indicates that each $\Lambda_i$ in $\boldsymbol{\Lambda}$ should be multiplied via $\circ$ by matrix $X$:

$$\boldsymbol{\Lambda} \circ X = \{\Lambda_1 \circ X, \Lambda_2 \circ X, \ldots, \Lambda_k \circ X\}. \quad [4]$$

Similar matrix-wise operations apply to $\boldsymbol{\Lambda}$ and $\tilde{\boldsymbol{\Lambda}}$ for vector multiplication (e.g., $\mathbf{1}^T\boldsymbol{\Lambda}\mathbf{1}$). We will use the matrices $I_{2,1}$ and $I_{1,\frac{1}{2}}$ in matrix-wise operations on arrays. $I_{2,1}$ is a matrix where all diagonal elements are 2s, while off-diagonals are 1s. It multiplies a diagonal element in each $\tilde{\Lambda}_i$ by two. $I_{1,\frac{1}{2}}$ is a matrix where all diagonal elements are 1s, while off-diagonals are 1/2s. It divides the off-diagonal elements in each $\tilde{\Lambda}_i$ by two. These matrices are necessary to control for the number of speciation and extinction events in calculations.

**Extinction events.** The extinction rates for each state are given by the column vector $\boldsymbol{\mu}$:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix} \quad [5]$$

**State-change events.** The instantaneous rate matrix matrix $Q$ defines anagenetic state changes and represents a traditional CTMC. Its entities are the transition rates $q_{ij}$.

**Probability Terms.** The probability of observing a particular branch at time $t$, given that the branch is in state $i$, is denoted by the scalar $D_i(t)$. The probability of lineage extinction at time $(t)$, given the lineage is in state $i$, is the scalar $E_i(t)$. These two entities are equivalent to $D_{N,i}(t)$ and $E_i(t)$ from BiSSE. Note, that we have dropped the index $N$ from $D_{N,i}(t)$ for brevity. The entities $D_i(t)$ and $E_i(t)$ can be presented in vector form as:

$$\boldsymbol{D} = \begin{bmatrix} D_1(t) \\ D_2(t) \\ \vdots \\ D_k(t) \end{bmatrix}, \quad \boldsymbol{E} = \begin{bmatrix} E_1(t) \\ E_2(t) \\ \vdots \\ E_k(t) \end{bmatrix}. \quad [6]$$

**Differential Equations and Likelihood.** The probability terms evolve over time and branch according to the following system of differential equations, which is written in array notation as:

$$\frac{\mathrm{d}\boldsymbol{D}}{\mathrm{d}t} = -\left(\mathbf{1}^T\boldsymbol{\Lambda}\mathbf{1} + \boldsymbol{\mu}\right) \circ \boldsymbol{D} + Q\boldsymbol{D} + \mathbf{1}^T\left(\tilde{\boldsymbol{\Lambda}} \circ I_{2,1} \circ (\boldsymbol{D} \otimes \boldsymbol{E}^T)\right)\mathbf{1}$$

$$\frac{\mathrm{d}\boldsymbol{E}}{\mathrm{d}t} = \boldsymbol{\mu} - \left(\mathbf{1}^T\boldsymbol{\Lambda}\mathbf{1} + \boldsymbol{\mu}\right) \circ \boldsymbol{E} + Q\boldsymbol{E} + \mathbf{1}^T\left(\tilde{\boldsymbol{\Lambda}} \circ I_{1,\frac{1}{2}} \circ (\boldsymbol{E} \otimes \boldsymbol{E}^T)\right)\mathbf{1}, \quad [7]$$

this system is equivalent to BiSSE-ness model extended for $k$ states. Note, that each of the vectors, $\frac{\mathrm{d}\boldsymbol{D}}{\mathrm{d}t}$ and $\frac{\mathrm{d}\boldsymbol{E}}{\mathrm{d}t}$, characterizes $k$ differential equations.

At each ancestral node ($A$) the diversification probabilities ($\boldsymbol{D_A}$) calculate as:

$$\boldsymbol{D_A} = \mathbf{1}^T\left(\tilde{\boldsymbol{\Lambda}} \circ I_{1,\frac{1}{2}} \circ (\boldsymbol{D_M} \otimes \boldsymbol{D_N}^T)\right)\mathbf{1}, \quad [8]$$

given that this node has two daughter lineages ($M$ and $N$) with the the diversification probabilities $\boldsymbol{D_M}$ and $\boldsymbol{D_N}$ respectively. This formula is equivalent to how BiSSE-ness combines the probabilities at nodes.

The likelihood of the tree ($\mathcal{L}$) calculates at the root as:

$$\mathcal{L} = \mathbf{1}^T\frac{\boldsymbol{\pi}\boldsymbol{D_R}}{w}, \quad [9]$$

where, $\boldsymbol{D_R}$ are the diversification probabilities at the root, and $\boldsymbol{\pi}$ is the initial vector of probabilities for each state. The term $w$ (Eq. 10) conditions the overall likelihood on clade survival (1), indicating that neither of the two daughter lineages at the root is allowed to become extinct before the present. Note, both the root and likelihood equation are equivalent to BiSSE-ness model too.

$$w = \mathbf{1}^T\boldsymbol{\pi}\mathbf{1}^T\left(\tilde{\boldsymbol{\Lambda}} \circ I_{1,\frac{1}{2}} \circ \left((1-\boldsymbol{E_R}) \otimes (1-\boldsymbol{E_R})^T\right)\right)\mathbf{1}. \quad [10]$$

**Example: HiClaSSE2.** Following Eq. (7), the HiClaSSE model with two states is defined by these ODEs:

$$\begin{aligned}
D_0'(t) &= 2D_0(t)E_0(t)\lambda_{000} + D_0(t)E_1(t)\lambda_{001} + D_1(t)E_0(t)\lambda_{001} + 2D_1(t)E_1(t)\lambda_{011} - D_0(t)(\lambda_{000} + \lambda_{001} + \lambda_{011} + \mu_0 + q_{01}) + D_1(t)q_{01} \\
D_1'(t) &= 2D_0(t)E_0(t)\lambda_{100} + D_0(t)E_1(t)\lambda_{101} + D_1(t)E_0(t)\lambda_{101} + 2D_1(t)E_1(t)\lambda_{111} - D_1(t)(\lambda_{100} + \lambda_{101} + \lambda_{111} + \mu_1 + q_{10}) + D_0(t)q_{10} \\
E_0'(t) &= E_0(t)^2\lambda_{000} + E_0(t)E_1(t)\lambda_{001} + E_1(t)^2\lambda_{011} - E_0(t)(\lambda_{000} + \lambda_{001} + \lambda_{011} + \mu_0 + q_{01}) + E_1(t)q_{01} + \mu_0 \\
E_1'(t) &= E_0(t)^2\lambda_{100} + E_0(t)E_1(t)\lambda_{101} + E_1(t)^2\lambda_{111} - E_1(t)(\lambda_{100} + \lambda_{101} + \lambda_{111} + \mu_1 + q_{10}) + E_0(t)q_{10} + \mu_1,
\end{aligned} \quad [11]$$

where, $\boldsymbol{f}'$ indicates $\boldsymbol{f}' = \frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}t}$.

These equations (Eq. 7) are equivalent to BiSSE-ness model, and thereby ClaSSE, but have a different parametrization. The BiSSE-ness model has 10 parameters: $(\lambda_0, \lambda_1, \mu_0, \mu_1, q_{01}, q_{10}, p_{0c}, p_{0a}, p_{1c}, p_{1a})$. These parameters include two speciation rates ($\lambda_0$, $\lambda_1$) that are weighted by probabilities representing anagenetic or cladogenetic speciation events. The parameter $p_{0c}$ is the probability of cladogenetic speciation when the ancestral lineage is in state 0, which can be asymmetric ($p_{0a}$) or symmetric ($1 - p_{0a}$). Similar logic applies to $p_{1c}$ and $p_{1a}$. The mapping between BiSSE-ness parameters (RHS) and HiClaSSE2 (LHS) is as follows:

$$\begin{aligned}
\lambda_{000} &= \lambda_0(1 - p_{0c}) \\
\lambda_{001} &= \lambda_0 p_{0c} p_{0a} \\
\lambda_{011} &= \lambda_0 p_{0c}(1 - p_{0a}) \\
\lambda_{100} &= \lambda_1 p_{1c}(1 - p_{1a}) \\
\lambda_{101} &= \lambda_1 p_{1c} p_{1a} \\
\lambda_{111} &= \lambda_1(1 - p_{1c})
\end{aligned} \quad [12]$$

## Correlated Evolution in SSE Models

Mark Pagel proposed conditions to determine whether two binary traits should be considered correlated or independent. We can extend these conditions to SSE models. Let us take a SSE model with two speciation and extinction rates: $(\lambda_1, \lambda_2, \lambda_1, \lambda_2)$, and $(\mu_1, \mu_2, \mu_1, \mu_2)$, and the rate matrix:

$$\begin{array}{cccc} A0 & B0 & A1 & B1 \end{array}$$
$$\begin{pmatrix} - & q_{12} & q_{13} & 0 \\ q_{21} & - & 0 & q_{24} \\ q_{31} & 0 & - & q_{34} \\ 0 & q_{42} & q_{43} & - \end{pmatrix} \quad [13]$$

For this model to be independent (i.e., CID), the following conditions should hold: $q_{13} = q_{24}, q_{31} = q_{42}, q_{12} = q_{34}, q_{21} = q_{43}$. If these conditions are not met, the model implies the following dependencies:

$$\begin{aligned} q_{13} \neq q_{24}: &\quad \text{change in trait depends upon diversification regime} \\ q_{31} \neq q_{42}: &\quad \text{change in trait depends upon diversification regime} \\ q_{12} \neq q_{34}: &\quad \text{change in diversification regime depends upon trait} \\ q_{21} \neq q_{43}: &\quad \text{change in diversification regime depends upon trait.} \end{aligned} \tag{14}$$

Consequently, if two different types of conditions from Eq. (14) are met simultaneously (e.g., $q_{13} \neq q_{24}$ and $q_{12} \neq q_{34}$), then the changes in both the trait and diversification regime are interdependent.

It is worth noting that there are additional conditions suggesting correlation. If any of the left diagonal rates, indicating dual transitions (e.g., $A0 \rightarrow B1$ is allowed), has a non-zero value, then a change in trait implies an immediate change in diversification regimes, and both are mutually dependent. Note that all these correlation conditions are only valid when we have at least two speciation or extinction rates (as in the CID model of this section). In the trivial case of equal speciation and extinction rates across all states, the diversification is naturally not affected by the trait since its invariable.

Non-independence can be proven by the fact that an independent SSE is separable, thereby it can be constructed from two individual models, one for trait and another for diversification. If this is not possible, then the SSE is correlated. This construction can be achieved by manually deriving the joint probabilities. Alternatively, for constructing the joint rate matrix $Q_j$, there is also a convenient algebraic operation using the Kronecker sum ($\oplus$) that combines individual matrices.

$$Q_j = Q_r \oplus Q_t = Q_r \otimes I_{Q_t} + I_{Q_t} \otimes Q_r, \tag{15}$$

where, $Q_r$ and $Q_t$ are the individual rate matrices for diversification regimes and traits, $I_{Q_r}$ and $I_{Q_t}$ are the identity matrices for the respective $Q$'s, and $\otimes$ denotes the Kronecker product. Consequently, all joint matrices that cannot be constructed using Eq. (15) imply correlation.

As an example, consider the models: COR8-C and COR8-NC from Eq. (28) (Table S1). Their rate matrices cannot be produced by combining individual trait and diversification models using Eq. (15). Therefore, they are not separable and describe dependent evolution. COR8-NC is a distinct model, while COR8-C is congruent with CID4 (Table S1) and shares the same likelihood.

## Lumpability Conditions for SSE models

**Lumpability of a basic CTMC.** Let us begin by discussing lumpability in the context of a basic CTMC and then extend it to SSE models. Suppose, we have an initial CTMC that describes the evolution of a four-state trait $\{s_1, s_2, s_3, s_4\}$ and is defined by the rate matrix $Q$:

$$Q = \begin{array}{c} \begin{array}{cccc} s_1 & s_2 & s_3 & s_4 \end{array} \\ \left(\begin{array}{cc|cc} - & q_{12} & q_{13} & q_{14} \\ q_{21} & - & q_{23} & q_{24} \\ \hline q_{31} & q_{32} & - & q_{34} \\ q_{41} & q_{42} & q_{43} & - \end{array}\right) \end{array}. \tag{16}$$

The aggregation of states at the level of the rate matrix means constructing a smaller matrix with fewer states and transitions. Suppose that we are willing to aggregate $Q$ into a two-state process $\hat{Q}$, with states $\{\hat{s}_1, \hat{s}_2\}$, using the following partition of the initial states: $\hat{s}_1 = \{s_1, s_2\}$, $\hat{s}_2 = \{s_3, s_4\}$. This partition can be visualized on $Q$ by splitting the rates into four partition blocks shown with a horizontal and a vertical line within the matrix in the Eq. (16). Each block includes four initial rates whose function should yield new rates in the aggregated matrix. So, the aggregated matrix should be:

$$\hat{Q} = \begin{array}{c} \begin{array}{cc} \hat{s}_1 & \hat{s}_2 \end{array} \\ \begin{pmatrix} - & \hat{q}_{12} \\ \hat{q}_{21} & - \end{pmatrix} \end{array}. \tag{17}$$

The lumpability property ensures that the transition rates of the aggregated CTMC can be calculated from the original CTMC, allowing us to model them using the aggregated process. The lumpability property described in the main text can be restated in the context of partition blocks within the rate matrix. We refer to this restatement as the "row-wise sum rule" (RWR). Both interpretations are equivalent.

The RWR implies the following: *the original CTMC is lumpable with respect to a given partition of states when the row-wise sum of rates within one partition block in the original $Q$ is the same for all rows within the given partition block, and this property holds for all blocks in $Q$. The rates in the aggregated matrix represent simply the row-wise sums of the original rates.*

For $Q$ in the equation Eq. (16), the RWR is maintained if the following two equalities hold:

$$\begin{aligned}\hat{q}_{12} &= q_{13} + q_{14} = q_{23} + q_{24} \\ \hat{q}_{21} &= q_{31} + q_{32} = q_{41} + q_{42}\end{aligned} \quad [18]$$

These equalities imply lumpability of $Q$ for the given state partition; $\hat{q}_{12}$ and $\hat{q}_{12}$ represent the new rates in the lumped process $\hat{Q}$, shown in the equation Eq. (17). Note, to prove lumpability, it is enough to show that the RWR holds for the off-diagonal blocks because it implies that the main diagonal blocks maintain the RWR too. This property will be helpful later, in classifying hidden expansions. In the lumped process, the initial probability vector at the tree root ($\boldsymbol{\pi}$) should be aggregated, too: by adding up state probabilities from the original process belonging to the same partition subset.

The lumpability property is general and applies to all types of CTMCs. Lumpability does not depend on the time over which a CTMC evolves nor the initial probability vector at the tree root; if a CTMC is lumpable, it is lumpable under any value of the initial vector. If the lumpability property does not hold, then the simple form for the rate matrix for the aggregated process does not exist since it is not Markovian.

**Lumpability and backward ODEs.** The previous example demonstrated lumpability for forward-in-time equations. However, in phylogenetics, likelihood is estimated using backward-in-time equations starting at $t = 0$. Let us first consider these equations for a basic CTMC from Eq. (16) to demonstrate our approach.

The system of backward-in-time equations for calculating likelihood over a tree branch can be written in matrix form as $\boldsymbol{D}' = Q\boldsymbol{D}$, where $\boldsymbol{D}$ is:

$$\boldsymbol{D} = \begin{bmatrix} D_1(t) \\ D_2(t) \\ D_3(t) \\ D_4(t) \end{bmatrix}. \quad [19]$$

This system expands as:

$$\begin{aligned}
D_1'(t) &= -(q_{12} + q_{13} + q_{14})D_1(t) &&+ q_{12}D_2(t) &&+ q_{13}D_3(t) &&+ q_{14}D_4(t), \\
D_2'(t) &= q_{21}D_1(t) &&- (q_{21} + q_{23} + q_{24})D_2(t) &&+ q_{23}D_3(t) &&+ q_{24}D_4(t), \\
D_3'(t) &= q_{31}D_1(t) &&+ q_{32}D_2(t) &&- (q_{31} + q_{32} + q_{34})D_3(t) &&+ q_{34}D_4(t), \\
D_4'(t) &= q_{41}D_1(t) &&+ q_{42}D_2(t) &&+ q_{43}D_3(t) &&- (q_{41} + q_{42} + q_{43})D_4(t).
\end{aligned} \quad [20]$$

Now, we want to construct a smaller CTMC given the state partition $\hat{s}_1 = \{s_1, s_2\}$, $\hat{s}_2 = \{s_3, s_4\}$ as above. Our approach involves first deriving a necessary condition for the smaller CTMC to be Markovian, then applying it to the system of the original ODEs, and finally deriving the relationships between the rates that are necessary and sufficient to maintain lumpability.

For the smaller CTMC to be Markovian in the backward equation, it is necessary that we do not distinguish between the original states that belong to the same lumped state. Thus, the $D_i(t)$'s belonging to the same lumped state must be equiprobable for any time $t$. In our example, this means that $D_1(t) = D_2(t)$ and $D_3(t) = D_4(t)$. Putting this conditions in equation Eq. (20) and after some algebra gives:

$$\begin{aligned}
D_1'(t) &= (q_{13} + q_{14})(D_3(t) - D_1(t)), \\
D_2'(t) &= (q_{23} + q_{24})(D_3(t) - D_1(t)), \\
D_3'(t) &= (q_{31} + q_{32})(D_1(t) - D_3(t)), \\
D_4'(t) &= (q_{41} + q_{42})(D_1(t) - D_3(t)).
\end{aligned} \quad [21]$$

Clearly, for this system to be lumpable the same equalities as in Eq. (18) should hold. This allows us to rewrite these four equations using just two equations that characterize the lumped process from Eq. (17):

$$\begin{aligned}
D_{12}'(t) &= \quad -\hat{q}_{12}D_{12}(t) + \hat{q}_{12}D_{34}(t) = \hat{q}_{12}(D_{34}(t) - D_{12}(t)) \\
D_{34}'(t) &= \quad \hat{q}_{21}D_{12}(t) - \hat{q}_{21}D_{34}(t) = \hat{q}_{21}(D_{12}(t) - D_{34}(t)),
\end{aligned} \quad [22]$$

where $D_{12}(t) = D_1(t) = D_2(t)$ and $D_{34}(t) = D_3(t) = D_4(t)$; the dependencies between the rates are the same as in Eq. (18). We will employ this approach to derive lumpability conditions for SSE models.

**Lumpability Conditions for SSE.** The ODEs for SSE models can become quite extensive and involve numerous parameters depending on the number of states. For simplicity, let us illustrate the method discussed in the previous section by applying it to a HiClaSSE3 with states denoted as $\{0, 1, 2\}$. Our goal is to lump these states into two states given the partition $\{\{0\}, \{1, 2\}\}$. We assume that the original SSE model has the maximum number of parameters, which include the extinction rates $\boldsymbol{\mu} = [\mu_0, \mu_1, \mu_2]$, and the transition rate matrix $Q_3$:

$$Q_3 = \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \left(\begin{array}{c|cc} - & q_{01} & q_{02} \\ \hline q_{10} & - & q_{12} \\ q_{20} & q_{21} & - \end{array}\right) \end{array}. \quad [23]$$

The coupled and decoupled speciation rates are specified by $\mathbf{\Lambda}$ tensor:

$$\mathbf{\Lambda} = \left\{ \Lambda_0 = \left( \begin{array}{c|cc} \lambda_{000} & \lambda_{001} & \lambda_{002} \\ \hline 0 & \lambda_{011} & \lambda_{012} \\ 0 & 0 & \lambda_{022} \end{array} \right) \middle| \Lambda_1 = \left( \begin{array}{c|cc} \lambda_{100} & \lambda_{101} & \lambda_{102} \\ \hline 0 & \lambda_{111} & \lambda_{112} \\ 0 & 0 & \lambda_{122} \end{array} \right), \Lambda_2 = \left( \begin{array}{c|cc} \lambda_{200} & \lambda_{201} & \lambda_{202} \\ \hline 0 & \lambda_{211} & \lambda_{212} \\ 0 & 0 & \lambda_{222} \end{array} \right) \right\}. \quad [24]$$

To lump the model, we set $D_1(t) = D_2(t)$ and use the ODEs fromEq. (7), which yields:

$$\boxed{\begin{aligned}
D_0'(t) &= 2D_0E_0\lambda_{000} + D_0E_1(\lambda_{001}+\lambda_{002}) + D_1E_0(\lambda_{001}+\lambda_{002}) + 2D_1E_1(\lambda_{011}+\lambda_{012}+\lambda_{022}) - D_0(\lambda_{000}+\lambda_{001}+\lambda_{002}+\lambda_{011}+\lambda_{012}+\lambda_{022}+\mu_0+q_{01}+q_{02}) + \\
&\quad + D_1(q_{01}+q_{02}), \\
D_1'(t) &= 2D_0E_0\lambda_{100} + D_0E_1(\lambda_{101}+\lambda_{102}) + D_1E_0(\lambda_{101}+\lambda_{102}) + 2D_1E_1(\lambda_{111}+\lambda_{112}+\lambda_{122}) - D_1(\lambda_{100}+\lambda_{101}+\lambda_{102}+\lambda_{111}+\lambda_{112}+\lambda_{122}+\mu_1+q_{10}) + D_0q_{10}, \\
D_2'(t) &= 2D_0E_0\lambda_{200} + D_0E_1(\lambda_{201}+\lambda_{202}) + D_1E_0(\lambda_{201}+\lambda_{202}) + 2D_1E_1(\lambda_{211}+\lambda_{212}+\lambda_{222}) - D_1(\lambda_{200}+\lambda_{201}+\lambda_{202}+\lambda_{211}+\lambda_{212}+\lambda_{222}+\mu_2+q_{20}) + D_0q_{20}, \\
\\
E_0'(t) &= E_0^2\lambda_{000} + E_0E_1(\lambda_{001}+\lambda_{002}) + E_1^2(\lambda_{011}+\lambda_{012}+\lambda_{022}) - E_0(\lambda_{000}+\lambda_{001}+\lambda_{002}+\lambda_{011}+\lambda_{012}+\lambda_{022}+\mu_0+q_{01}+q_{02}) + E_1(q_{01}+q_{02}) + \mu_0, \\
E_1'(t) &= E_0^2\lambda_{100} + E_0E_1(\lambda_{101}+\lambda_{102}) + E_1^2(\lambda_{111}+\lambda_{112}+\lambda_{122}) - E_1(\lambda_{100}+\lambda_{101}+\lambda_{102}+\lambda_{111}+\lambda_{112}+\lambda_{122}+\mu_1+q_{10}) + E_0q_{10} + \mu_1, \\
E_2'(t) &= E_0^2\lambda_{200} + E_0E_1(\lambda_{201}+\lambda_{202}) + E_1^2(\lambda_{211}+\lambda_{212}+\lambda_{222}) - E_1(\lambda_{200}+\lambda_{201}+\lambda_{202}+\lambda_{211}+\lambda_{212}+\lambda_{222}+\mu_2+q_{20}) + E_0q_{20} + \mu_2,
\end{aligned}} \quad [25]$$

where $D$'s and $E$'s are, indeed, $D(t)$'s and $E(t)$'s, with $t$ omitted for brevity. You may notice that the rates across these equations group together based on the same $D$ and $E$ terms. For example, for $D_1'(t)$, there is an entity $D_0E_1(\lambda_{101}+\lambda_{102})$ and for $D_2'(t)$, there is $D_0E_1(\lambda_{201}+\lambda_{202})$. If $\lambda_{101}+\lambda_{102} = \lambda_{201}+\lambda_{202}$, then the the two entities are equal. Thus, we can make the HiClaSSE3 model lumpable by ensuring that all such entities are equal. According to our state partition, HiClaSSE3 is lumpable if:

$$\begin{aligned}
\lambda_{100} &= \lambda_{200} \\
\lambda_{101}+\lambda_{102} &= \lambda_{201}+\lambda_{202} \\
\lambda_{111}+\lambda_{112}+\lambda_{122} &= \lambda_{211}+\lambda_{212}+\lambda_{222} \\
q_{10} &= q_{20} \\
\mu_1 &= \mu_2.
\end{aligned} \quad [26]$$

If all of these conditions are satisfied, it allows us to rewrite the HiClaSSE3 model from Eq. (25) as a HiClaSSE2 model with two states as shown in Eq. (11). The following relationships exist between the HiClaSSE3 (LHS) and HiClaSSE2 (RHS):

$$\begin{aligned}
\lambda_{000} &= \lambda_{000} \\
\lambda_{001}+\lambda_{002} &= \lambda_{001} \\
\lambda_{011}+\lambda_{012}+\lambda_{022} &= \lambda_{011} \\
\lambda_{100} &= \lambda_{100} \\
\lambda_{101}+\lambda_{102} &= \lambda_{101} \\
\lambda_{111}+\lambda_{112}+\lambda_{122} &= \lambda_{111} \\
\lambda_{200} &= \lambda_{100} \\
\lambda_{201}+\lambda_{202} &= \lambda_{101} \\
\lambda_{211}+\lambda_{212}+\lambda_{222} &= \lambda_{111} \\
q_{01}+q_{02} &= q_{01} \\
q_{10} &= q_{10} \\
q_{20} &= q_{10} \\
\mu_0 &= \mu_0 \\
\mu_1 &= \mu_1 \\
\mu_2 &= \mu_1,
\end{aligned} \quad [27]$$

**The lumpability algorithm.** For an SSE to be lumpable, each of its components: $\mathbf{\Lambda}$, $\boldsymbol{\mu}$, and $Q$, should satisfy lumpability conditions. While the lumpability conditions for $\boldsymbol{\mu}$ and $Q$ are straightforward and have been discussed earlier in detail, the conditions for the $\mathbf{\Lambda}$ tensor may be cumbersome. If we think of $\mathbf{\Lambda}$ not as a set of matrices but as a 3D array, then its lumpability is analogous to the RWR, but in three dimensions.

To check if $\mathbf{\Lambda}$ is lumpable given a certain state partition $\mathcal{B}$, you can follow this algorithm: (1) split each $\Lambda_i$ matrix in $\mathbf{\Lambda}$ into partition blocks, as done with $Q$, according to $\mathcal{B}$ (e.g., Eq. 24); (2) Then partition $\mathbf{\Lambda}$ itself by grouping individual matrices into subsets according to $\mathcal{B}$ (e.g., as shown by the vertical bar after $\Lambda_0$ in Eq. 24); (3) $\mathbf{\Lambda}$ is lumpable if, for each subset of matrices, the sum of rates across identical partition blocks is the same, and this condition holds for all subsets and blocks.

## Lumpability and Model Congruence

If the original model is lumpable with respect to the hidden states, then the lumped model is congruent with the original one. This can be understood by examining $D(t)$'s and $E(t)$'s. When calculating likelihood over a tree branch using Eq. (7), $D(t)$'s and $E(t)$'s belonging to the same observable state are identical due to the lumpability assumption. This implies that they are also the same at nodes and the root (Eq. 8). Therefore, lumping by aggregating hidden states essentially means that $D(t)$'s and $E(t)$'s at the root between the two models only differ in the number of states, while the likelihood values at the root (Eq. 9) are identical. Accordingly, the lumpability holds regardless of the sampling fraction, since it simply alters the relative value of $D(t)$'s and $E(t)$'s.

## Congruence between independent and dependent SSEs

The characteristics of all the discussed models are provided in Table S1 (the top part) and Eq. (28). These models were derived from CID4 using HE. The expanded models maintain the same parameter count of four parameters and consist of eight states. In these models, the speciation rates are anagenetic, except for CLA8-C and CLA8-NC, which have mostly cladogenetic rates. The congruent models (CID8, COR8-C, CLA8-C) can be reduced to CID4 through the following states partitioning: $\{\{a0, b0\}, \{c0, d0\}, \{a1, b1\}, \{c1, d1\}\}$, resulting in the states $\{A0, B0, A1, B1\}$. The rate matrix of COR8-C is lumpable, but COR8-NC is not lumpable as it violates the RWR, making it non-congruent. Similarly, the speciation tensor $\mathbf{\Lambda}$ of CLA8-C is lumpable, while that of CLA8-NC is not lumpable, resulting in non-congruence for the latter model as well.

$$Q_{CID4} = \begin{array}{c} \begin{array}{cccc} A0 & B0 & A1 & B1 \end{array} \\ \begin{pmatrix} - & q & q & 0 \\ q & - & 0 & q \\ q & 0 & - & q \\ 0 & q & q & - \end{pmatrix} \end{array} \tag{28a}$$

$$Q_{CID8} = \begin{array}{c} \begin{array}{cccccccc} a0 & b0 & c0 & d0 & a1 & b1 & c1 & d1 \end{array} \\ \left(\begin{array}{cc|cc|cc|cc} - & q & q & 0 & q & 0 & 0 & 0 \\ q & - & 0 & q & 0 & q & 0 & 0 \\ \hline q & 0 & - & q & 0 & 0 & q & 0 \\ 0 & q & q & - & 0 & 0 & 0 & q \\ \hline q & 0 & 0 & 0 & - & q & q & 0 \\ 0 & q & 0 & 0 & q & - & 0 & q \\ \hline 0 & 0 & q & 0 & q & 0 & - & q \\ 0 & 0 & 0 & q & 0 & q & q & - \end{array}\right) \end{array} \tag{28b}$$

$$Q_{COR8-C} = \begin{array}{c} \begin{array}{cccccccc} a0 & b0 & c0 & d0 & a1 & b1 & c1 & d1 \end{array} \\ \left(\begin{array}{cc|cc|cc|cc} - & q & q & 0 & 0 & q & 0 & 0 \\ q & - & 0 & q & q & 0 & 0 & 0 \\ \hline q & 0 & - & q & 0 & 0 & 0 & q \\ 0 & q & q & - & 0 & 0 & q & 0 \\ \hline 0 & q & 0 & 0 & - & q & q & 0 \\ q & 0 & 0 & 0 & q & - & 0 & q \\ \hline 0 & 0 & 0 & q & q & 0 & - & q \\ 0 & 0 & q & 0 & 0 & q & q & - \end{array}\right) \end{array} \tag{28c}$$

$$Q_{COR8-NC} = \begin{array}{c} \begin{array}{cccccccc} a0 & b0 & c0 & d0 & a1 & b1 & c1 & d1 \end{array} \\ \left(\begin{array}{cc|cc|cc|cc} - & q & q & 0 & 0 & q & 0 & 0 \\ q & - & 0 & q & 0 & 0 & 0 & 0 \\ \hline q & 0 & - & q & 0 & 0 & 0 & q \\ 0 & q & q & - & 0 & 0 & 0 & 0 \\ \hline 0 & q & 0 & 0 & - & q & q & 0 \\ 0 & 0 & 0 & 0 & q & - & 0 & q \\ \hline 0 & 0 & 0 & q & q & 0 & - & q \\ 0 & 0 & 0 & 0 & 0 & q & q & - \end{array}\right) \end{array} \tag{28d}$$

## Equal Rate Hidden Expansion (EHE)

The CID model used in the EHE transformation example consists of four parameters with specific parameter values. Its expansion results in EHE8-C, which has only two parameters and includes cladogenetic speciation events (Table S1, the bottom part; and Eq. 29).

**Table S1. The independent and dependent SSEs and their congruence to CID4 with four parameters**

| Model | $\cong$ CID4 | $Q$ Matrix | Speciation or other parameters |
|---|---|---|---|
| CID4 | | $Q_{CID4}$ | $\boldsymbol{\lambda_{CID4}} = (\lambda_1, \lambda_2, \lambda_1, \lambda_2)$ |
| CID8 | yes | $Q_{CID8}$ | $\boldsymbol{\lambda_{CID8}} = (\lambda_1, \lambda_1, \lambda_2, \lambda_2, \lambda_1, \lambda_1, \lambda_2, \lambda_2)$ |
| COR8-C | yes | $Q_{COR8-C}$ | $\boldsymbol{\lambda_{CID8}}$ |
| CLA8-C | yes | $Q_{COR8-C}$ | $\lambda_{022} = \lambda_{133} = \lambda_{202} = \lambda_{311} = \lambda_1$ & $\lambda_{466} = \lambda_{577} = \lambda_{646} = \lambda_{757} = \lambda_2$ |
| COR8-NC | no | $Q_{COR8-NC}$ | $\boldsymbol{\lambda_{CID8}}$ |
| CLA8-NC | no | $Q_{COR8-C}$ | $\lambda_{004} = \lambda_{116} = \lambda_{202} = \lambda_{311} = \lambda_1$ & $\lambda_{422} = \lambda_{577} = \lambda_{646} = \lambda_{757} = \lambda_2$ |
| CID4 | | $Q_{CID4}$ | $\boldsymbol{\lambda_{CID4}} = (0.3, 0.1, 0.3, 0.1), q = 0.2, \mu = 0.01$ |
| EHE8-C | yes | $Q_{EHE-C}$ | $\lambda_{000} = \lambda_{200} = \lambda_{111} = \lambda_{311} = \lambda_{444} = \lambda_{555} = \lambda_{666} = \lambda_{377} = 0.1, q = 0.1, \mu = 0.01$ |

Top of the table shows models which are congruent to CID4; their rate matrices are given in Eq. (28); all models have single extinction rate $\mu$; for CLA8-C and CLA8-NC $\lambda_1$ and $\lambda_2$ refer to the respective $\lambda$'s from $\boldsymbol{\lambda_{CID8}}$. Bottom of the table shows congruence between CID and EHE8-C with known parameter values; their rate matrices are given in Eq. (29). Note, all models have four parameters, except EHE8-C that has two.

Note that in the EHE example, the rates are exact multipliers of 0.1 for demonstration purposes. This is done to simplify the visualization. If other values with many decimals were used, EHE transformations could result in rate matrices with thousands of states, making it impractical to visualize them.

$$Q_{CID4} = \begin{array}{c} \begin{array}{cccc} A0 & B0 & A1 & B1 \end{array} \\ \begin{pmatrix} - & 0.2 & 0.2 & 0 \\ 0.2 & - & 0 & 0.2 \\ 0.2 & 0 & - & 0.2 \\ 0 & 0.2 & 0.2 & - \end{pmatrix} \end{array} \cong Q_{EHE8-C} = \begin{array}{c} \begin{array}{cccccccc} a0 & b0 & c0 & d0 & a1 & b1 & c1 & d1 \end{array} \\ \left(\begin{array}{cc|cc|cc|cc} - & 0 & 0.1 & 0.1 & 0.1 & 0.1 & 0 & 0 \\ 0 & - & 0.1 & 0.1 & 0.1 & 0.1 & 0 & 0 \\ \hline 0.1 & 0.1 & - & 0 & 0 & 0 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0 & - & 0 & 0 & 0.1 & 0.1 \\ \hline 0.1 & 0.1 & 0 & 0 & - & 0 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0 & 0 & 0 & - & 0.1 & 0.1 \\ \hline 0 & 0 & 0.1 & 0.1 & 0.1 & 0.1 & - & 0 \\ 0 & 0 & 0.1 & 0.1 & 0.1 & 0.1 & 0 & - \end{array}\right) \end{array}. \quad [29]$$

**The EHE Algorithm.** The EHE algorithm expands any CID model with decoupled separation events to an EHE model. It can also be adopted for any SSE. This algorithm expands each initial state with an equal number of additional hidden states.

To calculate the minimal EHE model, follow these steps:

1. Pool together the rates from $Q$ (only non-negative rates) and $\mathbf{\Lambda}$ into a single vector. Represent them as irreducible fractions of two integers and find their least common denominator ($lcd$).

2. Define the equal rate as $r = 1/lcd$.

3. Find the maximum rates in $\mathbf{\Lambda}$ and $Q$. Denote them as $\lambda_{max}$ and and $q_{max}$ respectively.

4. Calculate the necessary number of hidden states per initial state, represented as $N_\Lambda$ and $N_Q$, corresponding to $\mathbf{\Lambda}$ and $Q$ respectively. For the rate matrix $Q$, this number is:

$$N_Q = \frac{q_{max}}{r}$$

   For the speciation tensor $\mathbf{\Lambda}$ is:

$$N_\Lambda = \lceil u \rceil, \quad u = \frac{-1 + \sqrt{1 + 8\frac{\lambda_{max}}{r}}}{2}, \quad [30]$$

   where $\lceil \cdot \rceil$ represents the ceiling function. The ceiling of a number $x$ is the smallest integer that is greater than or equal to $x$. It is required since the number of hidden states is an integer. The formula from Eq. (30) derives from the number of different lambda rates $N_{\lambda_{ijz}}$ per initial CID state in the expanded model, which is:

$$N_{\lambda_{ijz}} = \frac{N_\Lambda^2 + N_\Lambda}{2}, \quad [31]$$

   setting $N_{\lambda_{ijz}} = \lambda_{max}/r$, and solving Eq. (31) yields Eq. (30).

5. Select the maximum value to determine the final number of hidden states $N_h$ needed for the observable one: $N_h = \max(N_Q, N_\Lambda)$.

6. Create the appropriate EHE model, corresponding to $N_h$, and distribute the rates to maintain lumpability.

For the CID4 from Eq. (29) and Table S1, the algorithm operates as follows. The three rates $\{0.1, 0.3, 0.2\}$ can be represented as fractions: 1/10, 3/10, and 2/10; their $lcd = 10$, hence $r = 1/lcd = 0.1$. The maximum rates are: $q_{max} = 0.2$ and $\lambda_{max} = 0.3$. So, $N_Q = 0.2/0.1 = 2$, $N_\Lambda = 2$, which means that $N_h = 2$, resulting in EHE having two hidden states for each initial state, as illustrated in Eq. 29 and Table S1.

The algorithm outlined above constructs the EHE model with the minimal number of hidden states. However, if one selects a different common denominator (other than the $lcd$) and applies the same algorithm, it will generate an EHE model with a greater number of hidden states.

## The Structure of a Congruence Class

To derive parametric forms for congruent models, we classify HEs into three main types based on the number of parameters in a HE model compared to the original SSE. This classification focuses on counting parameters located in the off-diagonal blocks (referred to as off-diagonal parameters) of the $Q$ matrix in the expanded model, while ignoring the total number of parameters in the model. This distinction is crucial because the parameters in the diagonal blocks are unidentifiable and can be set to any values, as they do not affect the RWR (see the section "Lumpability of a basic CTMC"). For example, setting the diagonal $q$'s in the $Q_{CID8}$ from Eq. (28) to 0 or any other separate parameter will not impact lumpability.

The three main types of HEs are as follows: equivalent expansion, super-expansion, and sub-expansion. These types generate models with the same, greater, and fewer off-diagonal parameters, respectively.

We also define two additional categories of HE models that are crucial for analyzing SSE dynamics. The Waiting-Time Preserving (WTP) hidden expansion does not alter the expected waiting times in $Q$ between the original model and HE. These waiting times are indicated by the negative main-diagonal entries in the rate matrix. Conversely, the opposite category is a non-Waiting-Time Preserving (nWTP) expansion, which modifies the expected waiting times in $Q$ between the original and HE models. For instance, $Q_{CID8}$ in Eq. (28) is nWTP, but if we change the $q$'s in the diagonal blocks to 0, it becomes WTP.

The parametric forms for equivalent and super-expansion models can be directly derived from $M_0$ using the lumpability conditions. The total number of parameters in these models remains the same (e.g., WTP equivalent type models) or increases (e.g., nWTP equivalent or super-expansion models). Super-expansion models have the potential for an unlimited number of parameters since there is no upper bound to maintain lumpability.

On the other hand, the parametric forms for sub-expansion models depend on the specific parameter values of $M_0$. This implies that each point in the parameter space of $M_0$ corresponds to a unique sub-expansion model. The total number of parameters in sub-expansion models can vary: (1) it can be fewer than in $M_0$ (e.g., WTP models), ranging down to two parameters (EHE), indicating that these models are nested within $M_0$; or (2) it can be the same as or greater than in $M_0$ (e.g., nWTP models).

## Are there other congruence classes?

Below, we show that CTMCs have other types of asymptotically unidentifiable congruent models besides those reviewed here, based on the lumpability property. In order to illustrate this, we use a CTMC alone for ancestral character state reconstruction that is a submodel of the SSE without diversification.

Certain congruent models can be constructed given that $e^X e^Y = e^{(X+Y)}$, if the matrices commute: $XY = YX$. For example, imagine a homogeneous CTMC that we denote as $\hat{Q}$ with all transitions equal to one (i.e, $\hat{q} = 1$); it evolves over a tree branch of length $\tau$, with probability $P(\hat{Q}, \tau) = e^{\hat{Q}\tau}$. This CTMC is congruent to a heterogeneous CTMC consisting of two rate matrices $Q_1$ and $Q_2$, in which the parameters ($q_1$ and $q_2$, respectively) are different between the matrices but equal within them following this relationship $q_1 + q_2 = 2$; for example, the rates can be: $q_1 = 0.5$ and $q_2 = 1.5$; apparently, $Q_1$ and $Q_2$ commute. The heterogeneous CTMC evolves under $Q_1$ from the start ($\tau = 0$) to the half-time ($t = \tau/2$), and under $Q_2$ from the half-time to $\tau$; its probability is $P(Q_1, Q_2, \tau) = e^{(Q_1+Q_2)\frac{\tau}{2}}$. Clearly, the probabilities are the same for these two processes $P(\hat{Q}, \tau) = P(Q_1, Q_2, \tau)$, and hence the likelihoods are the same too. There are infinite ways to construct a congruent heterogeneous process from a homogeneous one, following the properties of matrix commutativity. Obviously, in such trivial cases, the heterogeneous CTMC will always have more parameters than the homogeneous CTMC, which does not pose any problem of model selection since the homogeneous model can always be chosen due to its simplicity. Even if the true process was heterogeneous, we would never be able to detect it. We must assume that the homogeneous CTMC represents the average rate of the true process that we consider the simplest description. Therefore, this trivial congruence is less interesting than the HE congruence at this moment.

Things become more complicated if the matrices do not commute (i.e., $XY \neq YX$). Then, the homogeneous rate matrix ($\hat{Q}$) may or may not exist. This issue is related to the multiplicative closure of Markov models (2) that has been studied using interesting methods of Lie group theory (3, 4). These methods might be promising for studying model congruence in general. In this type of congruence, the problem of model selection may arise if it is possible to show that some homogeneous process ($\hat{Q}$) can be represented by a set of rate matrices that have fewer parameters in total than $\hat{Q}$. In that case, this would make the model selection ambiguous. We do not know any such examples, but this issue requires further research.

## References


1. S Nee, RM May, PH Harvey, The reconstructed evolutionary process. *Philos. Transactions Royal Soc. London. Ser. B: Biol. Sci.* **344**, 305–311 (1994).
2. JG Sumner, et al., Is the general time-reversible model bad for molecular phylogenetics? *Syst. biology* **61**, 1069–1074 (2012).
3. JG Sumner, J Fernández-Sánchez, PD Jarvis, Lie markov models. *J. theoretical biology* **298**, 16–31 (2012).
4. MD Woodhams, J Fernández-Sánchez, JG Sumner, A new hierarchy of phylogenetic models consistent with heterogeneous substitution rates. *Syst. biology* **64**, 638–650 (2015).