PLOS Computational Biology

RESEARCH ARTICLE

Testing the validity and adequacy of linguistic phylogenetic analyses

Benedict King*

Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Saxony, Germany

* benedict_king@eva.mpg.de

Abstract

Bayesian phylogenetics has become a standard tool in historical linguistics, and for the most part implements models borrowed from evolutionary biology. Not enough work has been done to validate the analysis set-up that has become standardised in phylolinguistics, which consists of binary data with ascertainment bias, data partitions with correlated cognate count and rate, the binary covarion substitution model, and the uncorrelated lognormal branch rate model. Here I perform a set of simulation-based calibration studies to test a typical phylolinguistic analysis in the software BEAST2. Although the analysis can correctly recover the parameters of the substitution model, complications arise due to the combination of ascertainment bias and partitions of unequal length and rate. Reweighting the partition-specific rates by the number of cognates, as is the default behaviour, leads to poorly calibrated posteriors. An alternative approach, where each meaning is assumed to come from a set of cognates of equal size, behaves correctly in simulations and is found to fit better to empirical data. I also assess the adequacy of the covarion substitution model through posterior predictive simulations. The covarion is found to fall short of approximating the true process of lexical evolution, likely due to the prevalence of semantic shift and the non-independence of cognate substitutions in real data. This work highlights the importance of thorough testing of models and their implementation in phylolinguistics, as well as the need for further research on improving models of lexical evolution.

Author summary

Linguists use Bayesian inference to construct language trees. Given the prevalence of these methods, it is vital that they are tested, so that we can have confidence in the results. Bayesian inference should produce results that are "well-calibrated", for example predictions with an 80% probability should be true 80% of the time. A correctly implemented analysis will be well-calibrated by design, with any deviation from good calibration indicative of problems in the

OPEN ACCESS

Citation: King B (2026) Testing the validity and adequacy of linguistic phylogenetic analyses. PLoS Comput Biol 22(5): e1014312. https://doi.org/10.1371/journal.pcbi.1014312

Editor: Jordan Douglas, University of Auckland, NEW ZEALAND

Received: December 16, 2025

Accepted: May 10, 2026

Published: May 20, 2026

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1014312

Data availability statement: All data and code supporting this study are available at https://doi.org/10.6084/m9.figshare.30870404.

Funding: BK is supported by the Max Planck Institute for Evolutionary Anthropology (https://www.eva.mpg.de). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The author has declared that no competing interests exist.

analysis set-up or software bugs. Here I perform a set of calibration experiments for typical phylolinguistic analyses. The analyses are well-calibrated, but only after a change to the default model set-up for handling meaning-specific rates. I also test the adequacy of linguistic phylogenetic models, which refers to how accurately the models reflect real-world processes. In several quantifiable ways, data simulated under the model differs significantly from the real data. This is likely driven by the non-independent evolution and widespread semantic shift that characterises real lexical data. This work highlights that although the implementation of phylolinguistic models is valid, in that the correct parameter estimates are returned for a given model and dataset, more work is needed to test if the models themselves are appropriate.

Introduction

Bayesian phylogenetic analysis of linguistic data is prevalent in the historical linguistics literature [1–30]. A set of standard models, implemented in the software package BEAST2 [31], has emerged for phylogenetic analysis of lexical data [32]. The data for a phylolinguistic analysis is usually cognate-coded basic vocabulary data [33]. Lexemes, specifically the canonical word form, are collected for each language in a given set of meanings. A list of the most stable meanings is chosen (a Swadesh list or variations thereof), which typically has around 150–200 items. These lexemes are then cognate-coded, so that each item is assigned to a class based on common descent.

Two main possibilities exist for how the cognate coded lexical data is structured for analysis: multistate and binarised. In a multistate analysis a dataset of N meanings is represented as N phylogenetic characters with k_i states, where k_i represents the number of cognate classes in the meaning i. This multi-state approach has only occasionally been applied to linguistic data [3,19,34]. Instead, the standard approach in phylolinguistics is binarised data (Table 1). Each cognate is treated as a character with two states: absent (0) and present (1), and treated independently from other cognates in the same meaning. There will then be $\sum_i^N k_i$ characters. The data consists of N meaning partitions, within which there will be a varying number of cognate sets. The number of cognate sets in a meaning, and therefore the length of the corresponding data partition, is dependent on the stability of the meaning. The faster the rate of lexical replacement, the more cognate sets. This can be observed in Table 1, where the stability of the numeral *two* means all the lexemes are cognate, whereas lexemes for the meaning *belly* fall into four different cognate sets.

It follows from the way the data is collected that there are no cognate sets absent for every language. This is known as ascertainment bias, and must be corrected for during the likelihood calculation [35]. The ascertainment bias correction involves a re-normalisation of the likelihood function to account for the fact that all-absent cognates are unobservable. For a given pattern x (a pattern refers to a unique configuration of absent and present states across languages) the likelihood is normalised thus:

Table 1. Example of binarised cognate data for two meanings and five Slavic languages.

concept	two	belly			
cognate	*du̯o-	*terh$_1$-	*b^{h}reu̯s-	*g^{w}i̯eh$_3$-	*k^{w}eh$_2$l-
Slovene	1	1	0	0	0
Polish	1	0	1	0	0
Czech	1	0	1	0	0
Ukrainian	1	0	0	1	0
Rusyn	1	0	0	0	1

https://doi.org/10.1371/journal.pcbi.1014312.t001

$$\Pr(x \mid \text{observable}) = \frac{\Pr(x)}{1 - \Pr(\text{unobservable})}$$

From this equation it is apparent that when the likelihood of an all-absent cognate is high (for example due to a low evolutionary rate), the renormalisation procedure increases the likelihood of the observed patterns due to the smaller denominator. Failure to apply this correction leads to over-estimation of the substitution rate [36]. In contrast to morphological phylogenetics, where the correction is for invariant characters [37], for cognate data ascertainment bias correction is only for cognates that are *absent* in all languages. Cognates that are present in all languages, although invariant, can and are observed: because phylolinguistics draws upon a fixed set of meanings, occasionally all lexemes in a particular meaning are cognate. Note that for empirical datasets with missing data, the ascertainment bias correction is slightly modified, since the probability of an all-absent increases as the proportion of missing data increases [7].

A handful of different models have been applied to the analysis of binarised lexical data, namely the binary [1,35], covarion [38,39] and Pseudo-Dollo [40] models. Where multiple models have been compared, the binary covarion has usually been the best-fitting model [3,4,10,11,14,18,20,24,28,30], but see [8]. As it is also by far the most commonly used model for linguistic data, it is the focus of the present study.

The binary covarion model divides absent and present states into additional fast and slow states, making a total of four states: absent-fast, present-fast, absent-slow and present-slow. The parameters of the covarion model are α, which determines the rate of the slow state relative to the fast state, s, the switching rate between fast and slow states and the equilibrium frequencies π_{obs} and π_h. Equilibrium frequencies determine the relative frequencies of the states after the process runs for an infinite length of time. The frequencies π_{obs} of the observed states and the hidden states (π_h) are combined to provide the equilibrium frequencies of the four states. The full unnormalised matrix of instantaneous transition rates (from row to column) between the states is then given by:

$$Q = \begin{pmatrix} - & \pi_{obs}[1]\pi_h[0] & s\pi_{obs}[0]\pi_h[1] & 0 \\ \pi_{obs}[0]\pi_h[0] & - & 0 & s\pi_{obs}[1]\pi_h[1] \\ s\pi_{obs}[0]\pi_h[0] & 0 & - & \alpha\pi_{obs}[1]\pi_h[1] \\ 0 & s\pi_{obs}[0]\pi_h[0] & \alpha\pi_{obs}[0]\pi_h[1] & - \end{pmatrix}$$

The implementation of the binary covarion model in BEAST2 has three different modes. The Q matrix above demonstrates the fully time-reversible set-up, which includes the frequencies of the hidden states within the Q matrix. In the default "beast" mode, the hidden frequencies are excluded from the Q matrix, which is then only time-reversible if the fast and slow hidden states have equal frequencies (0.5, 0.5). Finally, in the original "Tuffley-Steel" mode, the parameterisation includes separate switch rates for fast and slow modes, and the α parameter is fixed at a value of 0. For this study, I used the default set-up, prevalent within phylolinguistics, which is the standard "beast" parameterisation with hidden frequencies set at (0.5, 0.5).

As mentioned above, the number of cognate sets in a meaning partition is dependent on the rate of lexical replacement. Higher rates means more cognate sets. One way of modelling this is to apply different rate multipliers to each meaning partition. These rates are hereafter referred to as "partition rates", but are often called "mutation rates" within the BEAST2 software. The partition rates are drawn from a Dirichlet distribution, which, if unweighted, has a mean of 1 across the rates. However, partitions have different numbers of cognates, necessitating a reweighting procedure. Reweighting ensures that the overall mean rate across observed cognates is 1, taking into account that each partition rate is applied to a different number of cognates. The default reweighting procedure, implemented in BEAST2 analyses generated by the BEAUti software, sets the weights of the delta exchange operator equal to the number of cognates in each partition. However, the cognate sets are an ascertained sample, with all-absent cognates effectively filtered out. Because of this, the separate operations of ascertainment bias correction and partition rate reweighting may interact, the consequences of which are unexplored.

Three alternative configurations of partition rates were compared in [3]: all meanings given the same rate (no additional free parameters), an individual rate applied to every meaning partition (169 free parameters), and 8 different rates applied to meanings in bins according to the number of cognates (7 additional free parameters). The latter, binned model was found to be the best fitting.

In summary, Bayesian phylogenetic analysis of linguistic data combines the following features:

- Binary presence/absence data
- The binary covarion substitution model
- Ascertainment bias (all-absent cognates are not observed)
- Meaning partitions consisting of variable numbers of cognates
- Different rates applied to meaning partitions
- Correlation of partition length and partition rate

Recently there has been an increased focus on improved testing and validating of evolutionary biology software [41], including Bayesian phylogenetic models [42,43]. Because linguistic phylogenetic models borrow methodology from various sources within evolutionary biology, the particular combination of model aspects detailed above have not been thoroughly tested as a combination. Indeed, the *babel* package in BEAST2, which contains a number of features for analysing linguistic data specifically, has no associated publication.

Simulation-based calibration [44] is emphasised as a central method for validating phylogenetic models [43]. An implementation of a Bayesian model is well-calibrated when predictions with probability X% are correct X% of the time [45]. The procedure involves sampling parameters from their prior probability distributions, simulating data from these parameters, and analysing the simulated datasets under the same model. The results are then checked for coverage: for example the correct value for a parameter should lie within the 95% highest posterior density interval 95% of the time.

A second test is Rank Uniformity Validation (RUV). It has been shown that any particular draw of parameter values from the prior distribution is also a draw from the posterior distribution conditional on data generated by that same prior draw [46]. In other words, in SBC, the true value used to generate the data (the prior draw), and the parameter estimates produced when that data is analysed (the corresponding posterior draws), should come from the same distribution provided the analysis is correctly implemented. It follows that when the posterior draws are ordered by rank, the true value has an equal probability of falling between any two of the posterior draws. The rank of the true value within the corresponding posterior draws should follow a uniform distribution (S1 Fig). Rank Uniformity can also be visualised using an empirical cumulative distribution (ECDF) plot [47]. The rank of the true values within their respective posterior distributions, normalised to a 0–1 scale (the Probability Integral Transform, or PIT), is on the x-axis. The y axis is the ECDF, and shows cumulative

proportion of PIT values equal to or less the value on the x axis. Under rank uniformity, the ECDF plot should follow the diagonal. An ECDF difference plot, depicting departures from the expected value, more clearly shows any violations of rank uniformity. Examples of ECDF difference plots showing various kinds of departure from RUV are shown in S2 Fig.

Beyond validating the implementation of phylolinguistic analyses, there is also the question of model adequacy. Whereas validation ensures correctness of the computation in the inference machinery, adequacy refers to the how well the model represents the true process. It is possible that an analysis can be valid, in that the correct parameter estimates for a given model and dataset are estimated, but nevertheless inadequate if the model is mismatched to the underlying process. In phylolinguistics, model adequacy is a question of how accurately the covarion model approximates the true process of lexical evolution.

The absolute adequacy of phylogenetic models can be assessed using posterior predictive simulations [48–50]. In posterior predictive simulation, simulations are performed drawing on parameters from the posterior distribution, i.e., those estimated from the empirical data. If the model adequately describes the true process, these simulated datasets should "look like" the empirical data, and this is quantified using one or more metrics. A large discrepancy between values for a metric calculated from the posterior predictive datasets compared to the empirical data indicates model inadequacy.

Here I assess the validity of the standard approach to Bayesian inference of lexical data using simulation-based calibration. I also perform posterior predictive simulations to assess the adequacy of the covarion model to describe the process of lexical evolution.

Materials and methods

Simulation-based calibration

I set up a typical phylogenetic model for linguistic data (Fig 1). The model used the binary covarion substitution model, in which the parameters are the relative rate of the slow state (α_{bcov}), the switch rate between fast and slow states (s_{bcov}), the stable frequencies of the present and absent states (π_{obs}), and the stable frequencies of the hidden (fast and slow) states (π_{hidden}). To represent different data partitions with different stabilities, I drew three rates (m_i) from a Dirichlet distribution and simulated 3 partitions (D_i) of 10000 cognates using these rates. The large number of simulated cognates was necessary due to the large proportion of all-absent cognates in the simulated data. The tree of 30 tips was drawn from a Yule model with birth rate λ, and the branch rates from lognormal distribution [51]. The clock rate (μ) was fixed to 0.05, so that the root age and tree length were identifiable.

To perform simulation-based calibration (SBC) I generated 100 simulated datasets using the software LinguaPhylo [42]. Simulated data differs from empirical data because all-absent cognates are present. All-absent cognates in the simulations primarily result when simulations begin in the absent state at the root, and no transitions to the present state occur along the tree. When the transition rate from absent to present is low (either due to a slow clock rate or a low stable frequency of the present state), a correspondingly larger proportion of the 10,000 simulated cognate sets are all-absent. I re-analysed simulated datasets, using the same model and priors used to generate the data, in BEAST2.7.8 [31]. To confirm that the MCMC chains had run sufficiently long, I calculated effective sample size (ESS) scores for each free parameter in each simulation replicate using the R package sns [52]. ESS scores were over 200 for all parameters in all simulation replicates.

Three versions of the SBC were performed to test the ascertainment bias correction, based on different treatments of all-absent cognate classes and the weighting of the 3 partition rates.

1. The simulated data remained unaltered, all-absent cognates retained.
2. All-absent cognates removed and ascertainment bias correction implemented. Partition rates not reweighted.
3. All-absent cognates removed and ascertainment bias correction implemented. Partition rates re-weighted (the default option in BEAST2). I applied the same weighting to the Dirichlet distribution prior from which the partition rates are drawn, and the partition rate operator during reanalysis of simulated data.

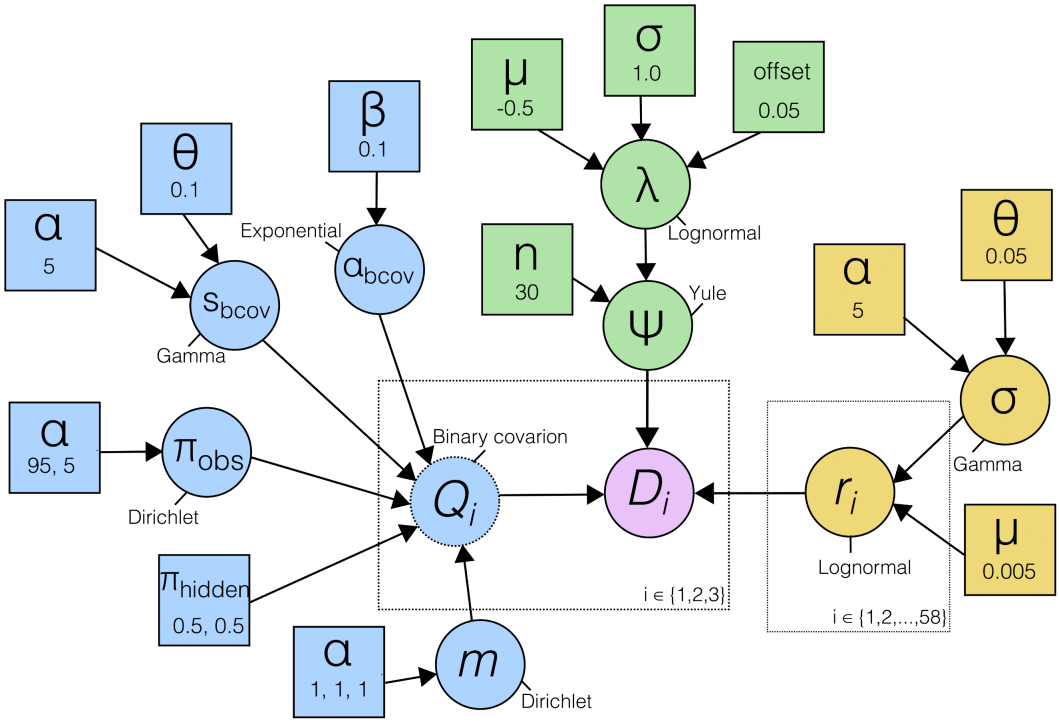

Fig 1. Graphical model of the analysis used for Simulation-Based Calibration. In blue, the parameters of the covarion substitution model, in green the tree model and in yellow the relaxed clock model. $i \in \{1, 2, 3\}$ and $i \in \{1, 2, ..., 58\}$ refer to iteration over partition rates and branches, respectively.

https://doi.org/10.1371/journal.pcbi.1014312.g001

The outcomes of the Bayesian analyses were assessed for both coverage and Rank Uniformity Validation, following best practice guidelines [43]. To assess if clade posterior probabilities were well-calibrated, I first generated for each simulation replicate a list of all clades found in the posterior, then calculated their posterior probability and whether or not they were present in the true tree. Results for all 100 simulation replicates were then pooled and calibration assessed using stable reliability diagrams [53].

Model-testing on empirical data

I tested three different model setups on a dataset of Indo-Iranic languages taken from the IE-CoR dataset [54]. The analysis set-up followed that of [3]. The 170 meaning classes were divided into five partitions based on the number of cognate sets, with bins of 1–5, 6–10, etc. Each of these partitions had a separate partition rate parameter (sometimes referred to as the mutation rate). These parameters were drawn from a weighted Dirichlet distribution and a weighted delta exchange operator was applied to propose new values.

Three different weighting strategies were compared.

1. Weighting by the number of cognate sets in each partition. This is the default mode in BEAST2.
2. Unweighted: i.e., all partitions have equal weights
3. Weighting by the number of meanings in each partition

I calculated the marginal likelihood of these three models using path sampling [55], with 30 power posteriors estimated using parallel computation [56].

Posterior predictive simulations

I used a sample from the IE-CoR dataset of the extant languages of the Germanic, Italic and Celtic subgroups (35 languages, 1251 cognate sets). This sample contains no missing data, therefore facilitating the comparison of the empirical and simulated datasets. I extracted the subtree of these languages from the summary tree in [3]. This tree was fixed during the analysis, and the values of the binary covarion and relaxed clock parameters were estimated in BEAST2.7.8.

I simulated 1001 posterior predictive datasets, each corresponding to a set of parameter estimates from the log file of the BEAST2 analysis. For comparison of simulated and empirical datasets I calculated 8 metrics, designed to capture key aspects of language divergence, the composition of languages (in terms of cognates), and the distribution of cognates across languages. For languages I calculated the maximum, minimum and mean pairwise distance between languages, mean proportion of present cognates and variance in the number of present cognates. For cognates I calculated the number of prevalent cognates (prevalent defined as greater than 80% of the language sample), number of singletons (cognates present in one language, autapomorphies) and variance in cognate prevalence. For each metric I calculated the mid-point two-tailed p-value [57], denoted as p_B. A value of 1 would indicate perfect model fit: the metric for the empirical data falls in the centre of the distribution over the posterior predictive sample. A value of <0.05 indicates a significant difference between the empirical and posterior predictive datasets.

Results and discussion

The binary covarion model performs well, but ascertainment bias with multiple partitions remains a difficult problem

Validation simulation study 1, in which the simulated data were not modified to remove all-absent cognates, performed well on both coverage and RUV (S3 and S4 Figs). In simulation study 2, ascertainment columns were removed resulting in partitions of varying length, but the Dirichlet prior and delta exchange operator were not reweighted during the reanalysis of simulated data. This analysis also performed correctly. Parameters of the binary covarion substitution model, the partition-specific rates, the branch rate parameters and the tree height and length were all well-calibrated (Figs 2 and 3). Clade posterior probabilities were also well-calibrated (i.e., nodes with 60% support were in the true tree 60% of the time, etc.), as shown by the reliability diagram (Fig 4).

For simulation study 3, ascertainment columns were removed, and the Dirichlet prior and delta exchange operator on partition rates reweighted according to the number of cognates in each partition. This model fails coverage and RUV checks (Fig 5). Partition rates and the birth rate are underestimated, whereas the tree height and length are overestimated. Nevertheless, estimates for binary covarion parameters, branch rate model parameters, and clade posterior probabilities remain well-calibrated.

That analysis 3 should fail calibration checks is not surprising. Reweighting the Dirichlet distribution prior for reanalysis of simulated data means that there is a small difference between the model used to simulate the data and the model used to analyse it. More importantly, reweighting the delta exchange operator prevents the analysis finding the exactly correct values for the 3 partition rates, which sum to 1 under an unweighted rather than a weighted Dirichlet distribution.

The problem when it comes to analyses of empirical data is that the number of cognate sets before the removal of ascertainment columns is not known. Therefore an exact replica of our well-calibrated analysis 2 is not possible. More to the point, the number of cognate sets including all-absent cognates is not meaningful for empirical data, since there is no such thing as a cognate set that is absent in all languages, and no set number of these.

Nevertheless, the data is analysed *as if* all-absent cognate sets exist. One way to model this is to assume that for each meaning there is a fixed number of cognate sets, some of which are absent in all languages, and the number of observed cognate sets is a function of the rate for that meaning. On an empirical dataset of Indo-Iranic languages, this analysis

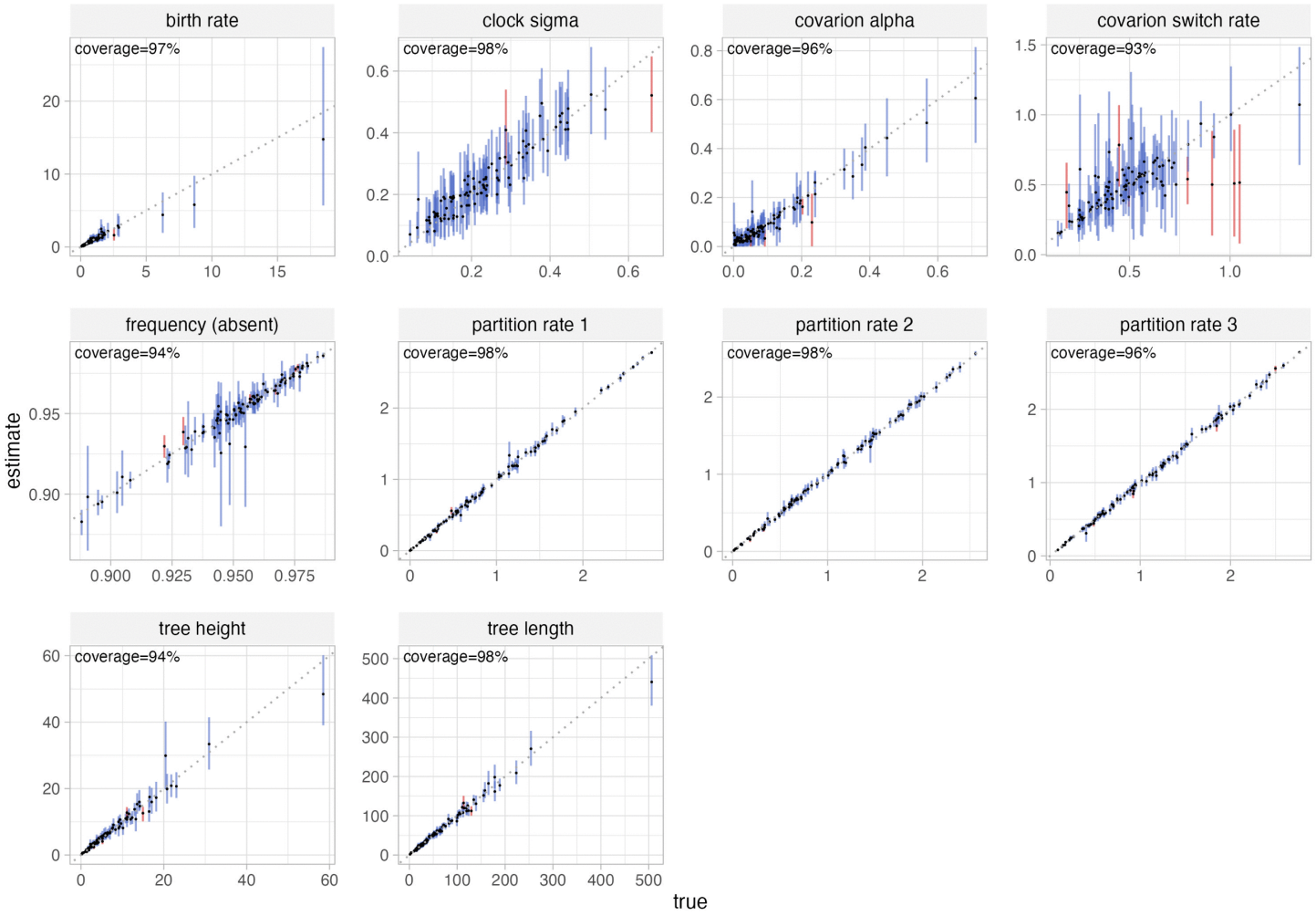

Fig 2. Coverage plot of validation simulation study 2. Ascertainment cognates (absent in all languages) were removed in this implementation, but partition rates were not reweighted. Each panel shows the true value from which data were simulated on the x axis, and the 95% highest posterior density (HPD) interval of the estimate of the parameter from the simulated data. The dotted diagonal lines represent x = y. Lines are blue when the true value is within the HPD, otherwise red. Coverage refers to the percentage of HPDs which contain the true value, which should be approximately 95% for a well-calibrated analysis.

https://doi.org/10.1371/journal.pcbi.1014312.g002

set-up, in which the Dirichlet prior and delta exchange operator are reweighted by the number of meanings, outperforms both a typical set-up in which reweighting follows the number of cognates, and an unweighted set-up (Table 2).

Comparison of parameter estimates from the meaning-weighted and cognate-weighted analyses shows differences in the partition rate estimates and the clock rate, mirroring the results found in the SBC tests (S5 and S6 Figs). Most notably, the clock rate is lower in the meaning-weighted analysis. This is not surprising, since the cognate-rich meanings no longer have an outsized effect on the weighted average of the partition rates. The clock rate under a meaning-weighted model set-up requires reinterpretation. The units correspond to expected changes per cognate per unit time *in a meaning with average rate*.

The results raise questions about how previously published linguistic phylogenies, which use the default partitioning set up, have been affected. SBC tests show that problems are confined to the partition rates, birth rate, tree height and tree length (the latter three are expected to correlate). In the empirical analyses, only the clock rate differs between the

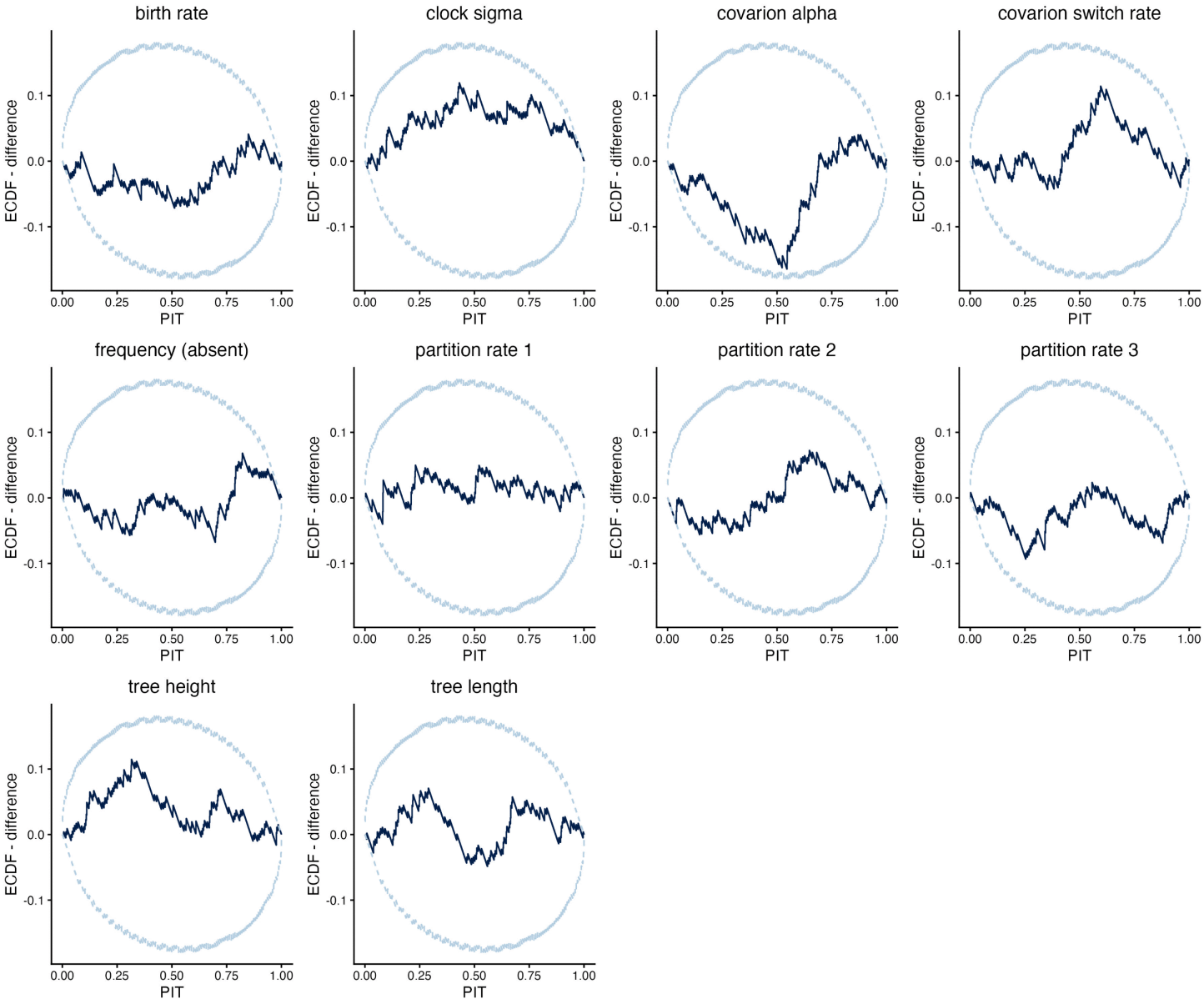

Fig 3. ECDF difference plot of validation simulation study 2. Ascertainment cognates were removed in this implementation, but partition rates were not reweighted. An ECDF plot shows the empirical cumulative density curve of the PIT scores for the true values within the posterior estimates from simulated data. The ECDF difference plot shows deviation from the expected ECDF curve under uniformity. Since the plot remains within the 95% confidence bands for all parameters, there is no significant deviation from uniformity, and the analysis is considered well-calibrated. Further information on interpretation is in the main text and S1 and S2 Figs.

https://doi.org/10.1371/journal.pcbi.1014312.g003

different partitioning set-ups, whereas birth rate, tree height and tree length remain unchanged. In the SBC tests the clock rate is fixed so problems will manifest as increased tree height. In the empirical data, where clock rate and tree height are estimated, only the clock rate and partition rates differ between alternative partitioning set-ups. Since the clock rate and partition rates are rarely the parameters of interest on which the success or failure of a particular hypothesis rests, the results here do not call into question the main conclusions of previous phylolinguistic studies. Since the partition rates in

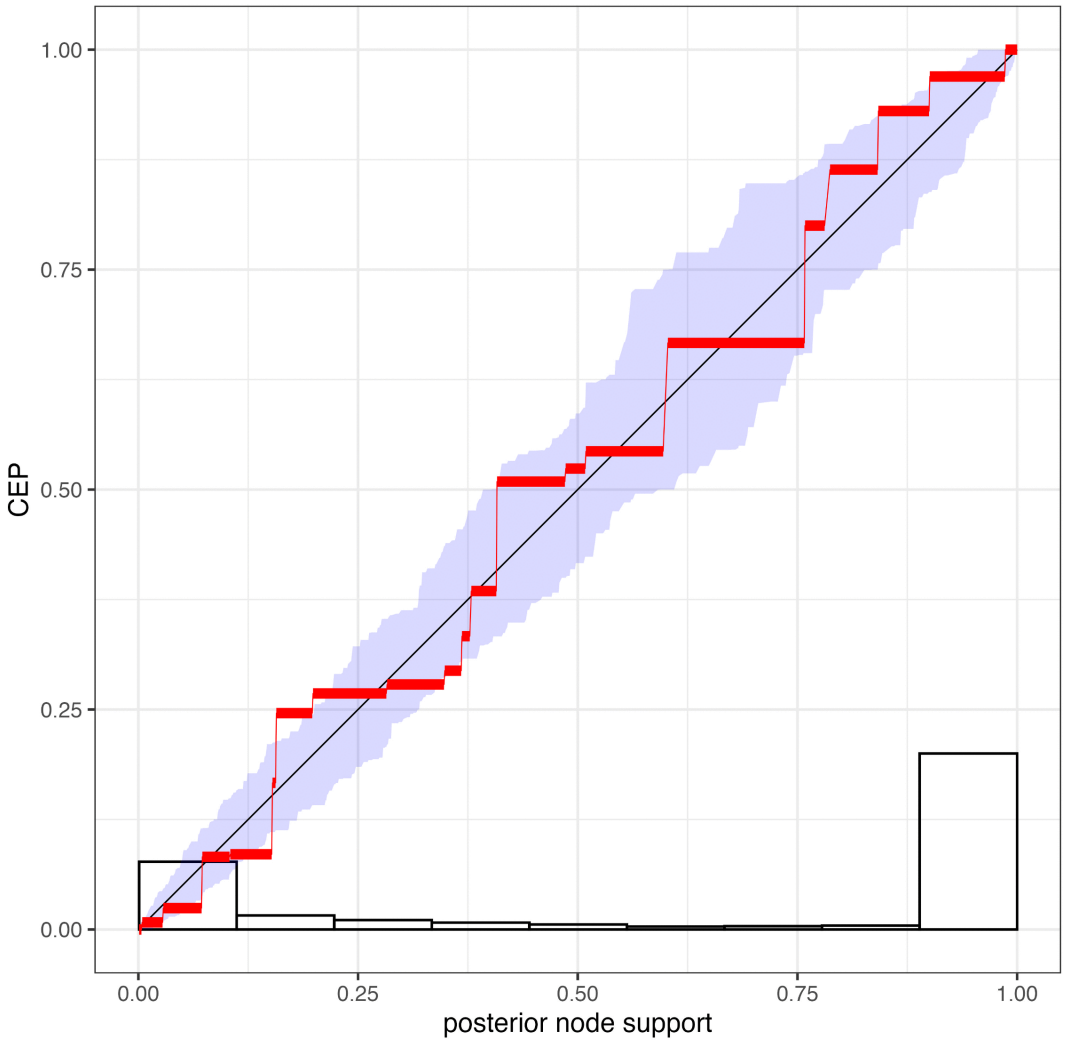

Fig 4. Stable reliability diagram showing that node support values are well-calibrated in validation simulation study 2. The x-axis shows the posterior node probability, and the y-axis shows the conditional event probability (CEP), the probability that a node is observed in the true tree given it's posterior support value. The stepped nature of the graph is because of binning of nodes with a range of posterior probabilities. The purple area represents 90% confidence bands. The histogram shows the frequency of nodes with different support values. Essentially, this diagram shows that nodes with a posterior probability of X% are true X% of the time.

https://doi.org/10.1371/journal.pcbi.1014312.g004

the SBC test are drawn from distributions similar to those found in empirical analyses, problems should not necessarily increase on larger datasets than those tested.

The binary covarion model does not fully describe all aspects of lexical evolution

Of the eight metrics used to compare the posterior predictive and empirical datasets, four show a significant difference ($p_B < 0.05$), suggesting model misspecification (Fig 6). These differences can be attributed to two main causes, violation of the assumption of stability, and non-independent evolution of cognate sets.

Violation of the assumption of stability is evidenced in the lower proportion of present cognates in the simulated data than the empirical data. Arguably, the stable frequency of the present state for any given cognate is even more extreme

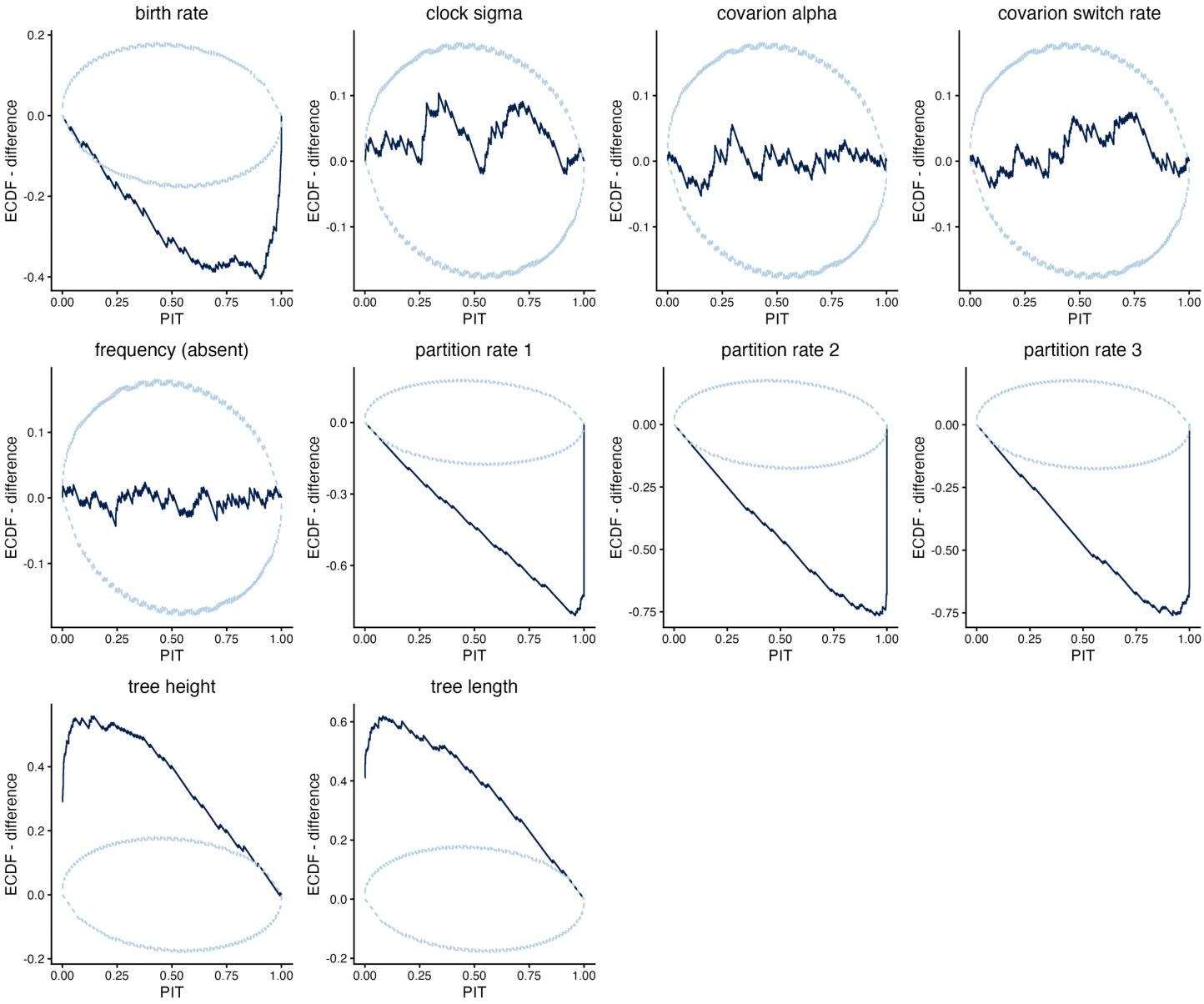

Fig 5. ECDF difference plot of validation simulation study 3. Ascertainment cognates were removed in this implementation, and partition rates were reweighted. The plots reveal systematic underestimation of birth rate and partition rates, and overestimation of tree height and tree length.

https://doi.org/10.1371/journal.pcbi.1014312.g005

Table 2. Bayes factor comparison of three different ways of weighting the parition-specific rate multipliers (mutation rates). Bayes factors compare each model to the best-supported model. Weighting by the number of meaning classes in the partition is the best performing model, whereas weighting by the number of cognate sets performs the worst.

model	log marginal likelihood	Bayes Factor vs best
reweighted by number of cognates	-15401.4	0.0116
reweighted by number of meanings	-15396.94	1
equal weighting	-15399.4	0.089

https://doi.org/10.1371/journal.pcbi.1014312.t002

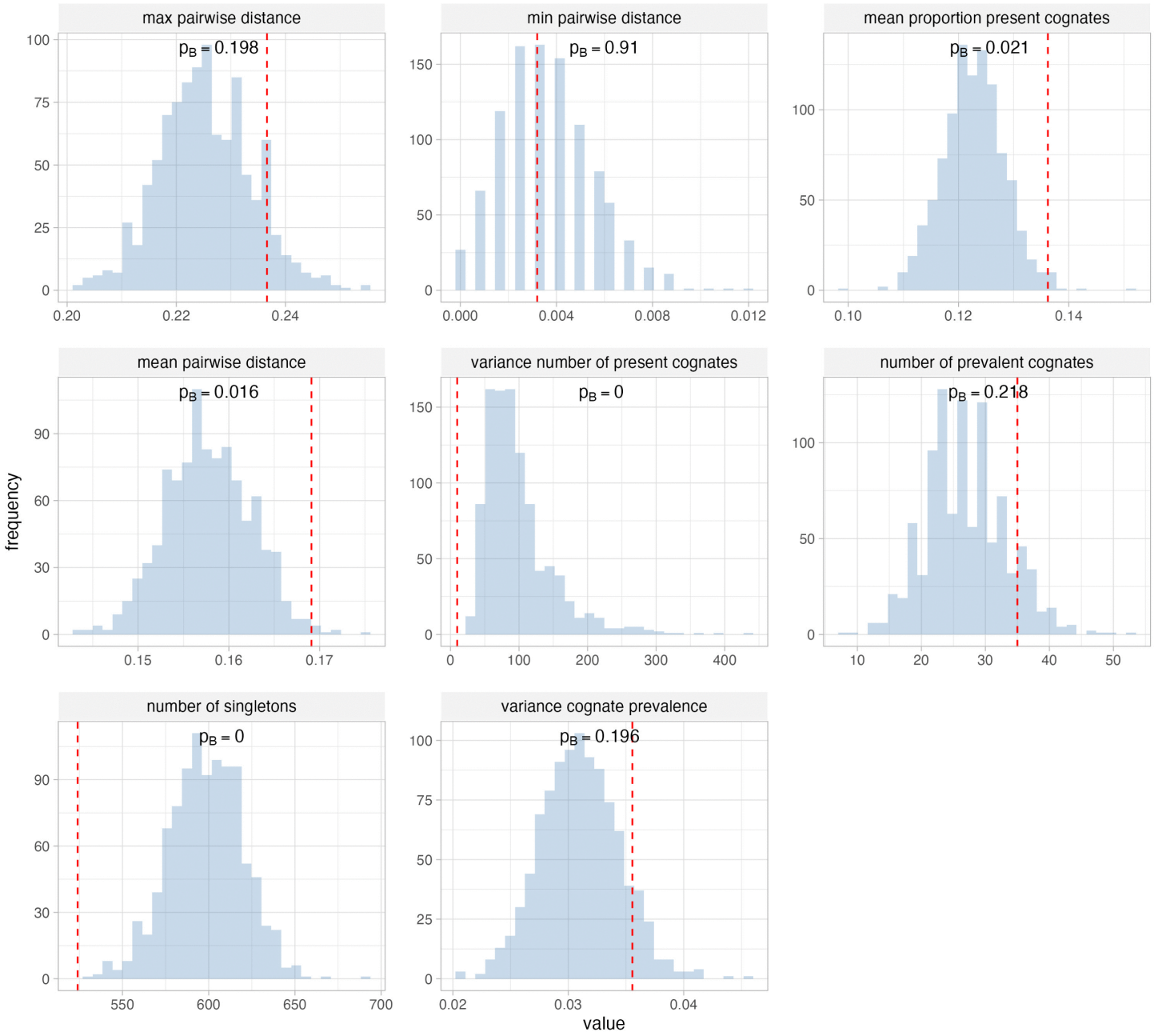

Fig 6. Comparing eight metrics calculated on the posterior predictive datasets (blue histogram) and the empirical data (red line). p_B is the posterior predictive p value.

https://doi.org/10.1371/journal.pcbi.1014312.g006

than that estimated by the analysis, and is in fact zero. When cognates are truly lost (rather than undergoing semantic shift) this process is not reversible. Over very long timescales any given cognate will eventually be lost, although many can persist for thousands of years.

The prevalence of semantic shift in the evolution of the lexicon could explain both the higher mean pairwise distance and the lower number of singletons in the empirical data. Frequent semantic shifts within a language subgroup could

inflate the number of cognate differences between closely related languages, but most of these "innovations" would be of cognate sets that occur elsewhere on the tree rather than new cognate sets unique to a single language.

Perhaps the least surprising result is the difference in the variance in the number of present cognates between the empirical and posterior predictive datasets. The IE-CoR dataset contains a single lexeme per meaning class per language, with few exceptions. Therefore cognate sets are not truly independent within a meaning class, despite being modelled as such. An alternative way of modelling the dataset is the multistate approach [3,19,34], where each meaning is modelled as a single character with multiple states, one for each cognate set. Most promising in this regard is the multistate model with infinite state space [34]. However, this has yet to be implemented on large-scale datasets or in mainstream phylogenetic software.

Model inadequacy, as shown by posterior predictive simulations, raises the question of how the estimation of the parameters of interest is affected. In a phylolinguistic analysis the parameters of interest are usually node ages and tree topology. Node age estimates could conceivably be affected by semantic shift, for example if cognate differences between closely related languages are inflated more relative to the differences between distantly languages. Node age estimates may also be affected by the non-independent evolution of cognates. This is demonstrated by previous implementations of a multistate model, which produced younger (arguably unrealistically young) node age estimates when compared to the binary covarion model [3,19,34]. Future work to explore how the inadequacy of the covarion model affects node age estimates is needed.

Conclusion

The standard set-up for a phylogenetic analysis of linguistic data, combining ascertainment bias, partitioned rates and the covarion model has been validated. The BEAST2 software returns well-calibrated estimates of the covarion model substitution parameters and stationary frequencies. A caveat concerns the combination of ascertainment bias with partitions of varying length, a consequence of partition length being a direct consequence of the partition rate. Weighting the partition rates by the number of cognates, as is the default in BEAST2, results in the estimates of the partition rates and tree height/clock rate no longer being well-calibrated. Bayes Factors comparisons on empirical data confirm that assuming that each meaning comes has the same number of total cognate sets (including those that are absent in all languages), with partitions weighted by the number of meanings rather than the number of cognates, results in better fit. Weighting partitions by the number of meanings rather than the number of cognates should be the recommended practice in phylolinguistics going forward. An explicit weighted dirichlet distribution prior on the partition rates is also recommended, because an explicit prior distribution is lacking in the current BEAST2 default set-up.

Posterior predictive tests of model adequacy reveal misspecification of current phylogenetic models when it comes to capturing key aspects of lexical evolution. The prevalence of parallel semantic shift and the non-independent evolution of cognate sets are two major explanations for this. This highlights the need for more work on modelling lexical evolution in a realistic way. Researchers performing phylolinguistic analyses should be aware that treating each cognate set as independent, and the inability of current models to account for semantic shift, may distort results.

Supporting information

S1 Fig. Example of a PIT histogram, ECDF plot and ECDF difference plot for a well-calibrated model.
(JPG)

S2 Fig. Examples of ECDF difference plots for various examples of incorrectly implemented analyses.
(JPG)

S3 Fig. Coverage plot of validation simulation study 1. Ascertainment cognates are not removed in this implementation.
(JPG)

S4 Fig. ECDF difference plot of validation simulation study 1. Ascertainment cognates are not removed in this implementation.
(JPG)

S5 Fig. Parameter estimates under three weighting schemes for partition-specific rates. From an analysis on Indo-Iranic languages.
(JPG)

S6 Fig. Estimates of the partition-specific rate multipliers under three weighting schemes for partition-specific rates. From an analysis on Indo-Iranic languages.
(JPG)

Acknowledgments

I thank Alexei Drummond and Walter Xie for help with Lphy and members of the Department of Linguistic and Cultural Evolution for discussions.

Author contributions

Conceptualization: Benedict King.

Data curation: Benedict King.

Formal analysis: Benedict King.

Investigation: Benedict King.

Methodology: Benedict King.

Validation: Benedict King.

Visualization: Benedict King.

Writing – original draft: Benedict King.

Writing – review & editing: Benedict King.

References

1. Gray RD, Atkinson QD. Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature. 2003;426(6965):435–9. https://doi.org/10.1038/nature02029 PMID: 14647380
2. Auderset S, Greenhill SJ, DiCanio CT, Campbell EW. Subgrouping in a 'dialect continuum': a Bayesian phylogenetic analysis of the Mixtecan language family. J Lang Evol. 2023;8(1):33–63. https://doi.org/10.1093/jole/lzad004
3. Heggarty P, Anderson C, Scarborough M, King B, Bouckaert R, Jocz L, et al. Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. Science. 2023;381(6656):eabg0818. https://doi.org/10.1126/science.abg0818 PMID: 37499002
4. Bouckaert R, Lemey P, Dunn M, Greenhill SJ, Alekseyenko AV, Drummond AJ, et al. Mapping the origins and expansion of the Indo-European language family. Science. 2012;337(6097):957–60. https://doi.org/10.1126/science.1219669 PMID: 22923579
5. Sagart L, Jacques G, Lai Y, Ryder RJ, Thouzeau V, Greenhill SJ, et al. Dated language phylogenies shed light on the ancestry of Sino-Tibetan. Proc Natl Acad Sci U S A. 2019;116(21):10317–22. https://doi.org/10.1073/pnas.1817972116 PMID: 31061123
6. Wu M-S, Bodt TA, Tresoldi T. Bayesian phylogenetics illuminate shallower relationships among Trans-Himalayan languages in the Tibet-Arunachal area. LTBA. 2022;45(2):171–210. https://doi.org/10.1075/ltba.21019.wu
7. Chang W, Cathcart C, Hall D, Garrett A. Ancestry-constrained phylogenetic analysis supports the indo-european steppe hypothesis. Language. 2015;91(1):194–244. https://doi.org/10.1353/lan.2015.0005
8. Robbeets M, Bouckaert R. Bayesian phylolinguistics reveals the internal structure of the Transeurasian family. J Lang Evol. 2018;3(2):145–62. https://doi.org/10.1093/jole/lzy007
9. Robbeets M, Bouckaert R, Conte M, Savelyev A, Li T, An D-I, et al. Triangulation supports agricultural spread of the Transeurasian languages. Nature. 2021;599(7886):616–21. https://doi.org/10.1038/s41586-021-04108-8 PMID: 34759322

10. Lee S, Hasegawa T. Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages. Proc Biol Sci. 2011;278(1725):3662–9. https://doi.org/10.1098/rspb.2011.0518 PMID: 21543358
11. Kolipakam V, Jordan FM, Dunn M, Greenhill SJ, Bouckaert R, Gray RD, et al. A Bayesian phylogenetic study of the Dravidian language family. R Soc Open Sci. 2018;5(3):171504. https://doi.org/10.1098/rsos.171504 PMID: 29657761
12. King B, Greenhill SJ, Reid LA, Ross M, Walworth M, Gray RD. Bayesian phylogenetic analysis of Philippine languages supports a rapid migration of Malayo-Polynesian languages. Sci Rep. 2024;14(1):14967. https://doi.org/10.1038/s41598-024-65810-x PMID: 38942799
13. Bowern C, Atkinson Q. Computational phylogenetics and the internal structure of Pama-Nyungan. Language. 2012;88(4):817–45.
14. Bouckaert RR, Bowern C, Atkinson QD. The origin and expansion of Pama-Nyungan languages across Australia. Nat Ecol Evol. 2018;2(4):741–9. https://doi.org/10.1038/s41559-018-0489-3 PMID: 29531347
15. Grollemund R, Branford S, Bostoen K, Meade A, Venditti C, Pagel M. Bantu expansion shows that habitat alters the route and pace of human dispersals. Proc Natl Acad Sci U S A. 2015;112(43):13296–301. https://doi.org/10.1073/pnas.1503793112 PMID: 26371302
16. Koile E, Greenhill SJ, Blasi DE, Bouckaert R, Gray RD. Phylogeographic analysis of the Bantu language expansion supports a rainforest route. Proc Natl Acad Sci U S A. 2022;119(32):e2112853119. https://doi.org/10.1073/pnas.2112853119 PMID: 35914165
17. Birchall J, Dunn M, Greenhill SJ. A combined comparative and phylogenetic analysis of the Chapacuran language family. Int J Am Linguist. 2016;82(3):255–84. https://doi.org/10.1086/687383
18. Lee S. A sketch of language history in the Korean Peninsula. PLoS One. 2015;10(5):e0128448. https://doi.org/10.1371/journal.pone.0128448 PMID: 26024377
19. Kitchen A, Ehret C, Assefa S, Mulligan CJ. Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. Proc R Soc B Biol Sci. 2009;276(1668):2703–10. https://doi.org/10.1098/rspb.2009.0408 PMID: 19403539
20. Zhang M, Yan S, Pan W, Jin L. Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic. Nature. 2019;569(7754):112–5. https://doi.org/10.1038/s41586-019-1153-z PMID: 31019300
21. Michael L, Chousou-Polydouri N, Bartolomei K, Donnelly E, Meira S, Wauters V, et al. A Bayesian phylogenetic classification of Tupí-Guaraní. LIAMES: Ling Indig Am. 2015;15(2):193–221. https://doi.org/10.20396/liames.v15i2.8642301
22. Honkola T, Vesakoski O, Korhonen K, Lehtinen J, Syrjänen K, Wahlberg N. Cultural and climatic changes shape the evolutionary history of the Uralic languages. J Evol Biol. 2013;26(6):1244–53. https://doi.org/10.1111/jeb.12107 PMID: 23675756
23. Syrjänen K, Honkola T, Korhonen K, Lehtinen J, Vesakoski O, Wahlberg N. Shedding more light on language classification using basic vocabularies and phylogenetic methods: a case study of Uralic. Diachronica. 2013;30(3):323–52.
24. Oskolskaya S, Koile E, Robbeets M. A Bayesian approach to the classification of Tungusic languages. Diachronica. 2021;39(1):128–58. https://doi.org/10.1075/dia.20010.osk
25. Takahashi T, Onohara A, Ihara Y. Bayesian phylogenetic analysis of pitch-accent systems based on accentual class merger: a new method applied to Japanese dialects. J Lang Evol. 2023;8(2):169–91. https://doi.org/10.1093/jole/lzae004
26. Dhakal DN, List J-M, Roberts SG. A phylogenetic study of South-Western Tibetic. J Lang Evol. 2024;9(1–2):14–28. https://doi.org/10.1093/jole/lzae008
27. Huisman JLA, McLean B, Wu C-H. Combined lexical and phonotactic data resolve uncertainties in the evolutionary diversification of the Japonic language family. J Lang Evol. 2025;10(1):lzaf002. https://doi.org/10.1093/jole/lzaf002
28. Tao Y, Wei Y, Ge J, Pan Y, Wang W, Bi Q, et al. Phylogenetic evidence reveals early Kra-Dai divergence and dispersal in the late Holocene. Nat Commun. 2023;14(1):6924. https://doi.org/10.1038/s41467-023-42761-x PMID: 37903755
29. Ferraz Gerardi F, Wientzek T, Roksandic I, Gregorio de Souza J, Orphão de Carvalho F. A phylogenetic classification of the Je language family. Open Res Eur. 2025;5:29. https://doi.org/10.12688/openreseurope.19346.3 PMID: 40475316
30. Greenhill S, Haynie H, Ross R, Chira A, List J-M, Campbell L, et al. A recent northern origin for the Uto-Aztecan family. Language. 2023. https://doi.org/10.1353/lan.0.0276
31. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et al. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 2019;15(4):e1006650. https://doi.org/10.1371/journal.pcbi.1006650 PMID: 30958812
32. Hoffmann K, Bouckaert R, Greenhill SJ, Kühnert D. Bayesian phylogenetic analysis of linguistic data using BEAST. J Lang Evol. 2021;6(2):119–35. https://doi.org/10.1093/jole/lzab005
33. Greenhill SJ, Heggarty P, Gray RD. Bayesian phylolinguistics. In: The handbook of historical linguistics, vol. 2; 2020. p. 226-53.
34. Rönchen P, Wiklund T, Hammarström H. Likelihood calculation in a multistate model of vocabulary evolution for linguistic dating. Lang Dyn Change. 2024;14(1):1–41. https://doi.org/10.1163/22105832-bja10032
35. Felsenstein J. Phylogenies from restriction sites: a maximum-likelihood approach. Evolution. 1992;46(1):159–73. https://doi.org/10.1111/j.1558-5646.1992.tb01991.x PMID: 28564959
36. Gray RR, Tatem AJ, Johnson JA, Alekseyenko AV, Pybus OG, Suchard MA, et al. Testing spatiotemporal hypothesis of bacterial evolution using methicillin-resistant Staphylococcus aureus ST239 genome-wide data within a bayesian framework. Mol Biol Evol. 2011;28(5):1593–603. https://doi.org/10.1093/molbev/msq319 PMID: 21112962

37. Lewis PO. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst Biol. 2001;50(6):913–25. https://doi.org/10.1080/106351501753462876 PMID: 12116640
38. Tuffley C, Steel M. Modeling the covarion hypothesis of nucleotide substitution. Math Biosci. 1998;147(1):63–91. https://doi.org/10.1016/s0025-5564(97)00081-3 PMID: 9401352
39. Penny D, McComish BJ, Charleston MA, Hendy MD. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. J Mol Evol. 2001;53(6):711–23. https://doi.org/10.1007/s002390010258 PMID: 11677631
40. Bouckaert RR, Robbeets M. Pseudo Dollo models for the evolution of binary characters along a tree. BioRxiv. 2017:207571.
41. Darriba D, Flouri T, Stamatakis A. The state of software for evolutionary biology. Mol Biol Evol. 2018;35(5):1037–46. https://doi.org/10.1093/molbev/msy014 PMID: 29385525
42. Drummond AJ, Chen K, Mendes FK, Xie D. LinguaPhylo: a probabilistic model specification language for reproducible phylogenetic analyses. PLoS Comput Biol. 2023;19(7):e1011226. https://doi.org/10.1371/journal.pcbi.1011226 PMID: 37463154
43. Mendes FK, Bouckaert R, Carvalho LM, Drummond AJ. How to validate a Bayesian evolutionary model. Syst Biol. 2025;74(1):158–75. https://doi.org/10.1093/sysbio/syae064 PMID: 39506375
44. Talts S, Betancourt M, Simpson D, Vehtari A, Gelman A. Validating Bayesian inference algorithms with simulation-based calibration. arXiv:180406788 [Preprint]. 2018.
45. Dawid AP. The well-calibrated Bayesian. J Am Stat Assoc. 1982;77(379):605–10.
46. Cook SR, Gelman A, Rubin DB. Validation of software for bayesian models using posterior quantiles. J Comput Graph Stat. 2006;15(3):675–92. https://doi.org/10.1198/106186006x136976
47. Säilynoja T, Bürkner P-C, Vehtari A. Graphical test for discrete uniformity and its applications in goodness-of-fit evaluation and multiple sample comparison. Stat Comput. 2022;32(2). https://doi.org/10.1007/s11222-022-10090-6
48. Gelman A, Meng XL, Stern H. Posterior predictive assessment of model fitness via realized discrepancies. Stat Sin. 1996:733–60.
49. Bollback JP. Bayesian model adequacy and choice in phylogenetics. Mol Biol Evol. 2002;19(7):1171–80. https://doi.org/10.1093/oxfordjournals.molbev.a004175 PMID: 12082136
50. Brown JM. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. Syst Biol. 2014;63(3):334–48. https://doi.org/10.1093/sysbio/syu002 PMID: 24415681
51. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biol. 2006;4(5):e88. https://doi.org/10.1371/journal.pbio.0040088 PMID: 16683862
52. Mahani AS, Hasan A, Jiang M, Sharabiani MTA. Stochastic Newton Sampler: the R Package sns. J Stat Soft. 2016;74(Code Snippet 2). https://doi.org/10.18637/jss.v074.c02
53. Dimitriadis T, Gneiting T, Jordan AI. Stable reliability diagrams for probabilistic classifiers. Proc Natl Acad Sci U S A. 2021;118(8):e2016191118. https://doi.org/10.1073/pnas.2016191118 PMID: 33597296
54. Anderson C, Scarborough M, Jocz L, Kümmel MJ, Jügel T, Irslinger B, et al. The Indo-European Cognate Relationships dataset. Sci Data. 2025;12(1):1541. https://doi.org/10.1038/s41597-025-05445-3 PMID: 40897732
55. Lartillot N, Philippe H. Computing Bayes factors using thermodynamic integration. Syst Biol. 2006;55(2):195–207. https://doi.org/10.1080/10635150500433722 PMID: 16522570
56. Höhna S, Landis MJ, Huelsenbeck JP. Parallel power posterior analyses for fast computation of marginal likelihoods in phylogenetics. PeerJ. 2021;9:e12438. https://doi.org/10.7717/peerj.12438 PMID: 34760401
57. Höhna S, Coghill LM, Mount GG, Thomson RC, Brown JM. P3: Phylogenetic Posterior Prediction in RevBayes. Mol Biol Evol. 2018;35(4):1028–34. https://doi.org/10.1093/molbev/msx286 PMID: 29136211