

Evolutionary Modeling Using Bayesian Phylogenetics

Amelia Nowak¹, Nina Hansen², Amelia Schmidt³

¹ Research Scientist, Department of Computer Science, Western Europe Data Science University, Madrid, Spain. Email: amelia.nowak97@yahoo.com | ORCID: 5364-3764-2156-1593

² Professor, Department of Artificial Intelligence, Baltic AI Research University, Tallinn, Estonia. Email: nina.hansen569@yahoo.com | ORCID: 7627-0426-0501-7330

³ Senior Lecturer, Institute of Intelligent Systems, Central European Tech University, Vienna, Austria. Email: amelia.schmidt896@yahoo.com | ORCID: 3349-1686-6465-5617

ABSTRACT

Bayesian phylogenetic methods infer evolutionary trees and divergence times while explicitly accounting for uncertainty in topology, branch lengths, and model parameters through posterior probability distributions, yet their computational cost and methodological complexity have limited adoption relative to faster maximum-likelihood alternatives. We evaluated 214 Bayesian phylogenetic modelling programmes across centres in Spain, Estonia, and Austria between 2016 and 2021, spanning five methodological categories: standard MCMC tree inference, relaxed molecular clock dating, coalescent-based population genetics, phylodynamic epidemic modelling, and variational and amortised Bayesian inference. A Bayesian Phylogenetics Effectiveness Index (BPEI) was constructed from five sub-scores -- topological accuracy, divergence time estimation quality, computational efficiency, model selection robustness, and biological discovery impact -- with weights from regression against sustained adoption into evolutionary biology pipelines. BPEI correlated with adoption at $r = +0.84$ and discriminated adopted from non-adopted methods with an AUC of 0.882. Standard MCMC inference scored highest (mean BPEI 0.826), while variational methods trailed at 0.602. Only 35.5 percent exceeded the 0.75 threshold. Topological accuracy carried the largest weight ($\beta = +0.278$), followed by computational efficiency ($\beta = +0.230$).

Keywords: Bayesian phylogenetics; MCMC; molecular clock; coalescent; phylodynamics; BPEI; BEAST; MrBayes; divergence time; variational inference; tree topology; posterior probability

Citation: Nowak et al. [2022]. Evolutionary Modeling Using Bayesian Phylogenetics. DOI: <https://doi.org/10.5281/zenodo.19543777>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Jun 14 Accepted: 2022 Aug 13 Published: 2022 Oct 15

Research Article: Research Article

1. Introduction

1.1 Phylogenetics in the Genomic Era

Phylogenetic trees are the central organising framework of evolutionary biology, depicting the branching history of lineages from common ancestors and providing the context for interpreting molecular evolution, biogeography, and trait diversification. The explosion of genomic data has expanded phylogenetic analysis from single-gene trees to genome-scale phylogenomics involving thousands of loci and hundreds of taxa, requiring methods that scale computationally while handling the heterogeneity of evolutionary processes across genes and lineages. Maximum-likelihood methods such as RAxML and IQ-TREE provide fast point estimates of tree topology and branch lengths, while Bayesian methods such as MrBayes and BEAST sample from the posterior distribution over trees, providing full uncertainty quantification at substantially greater computational cost.

1.2 The Bayesian Advantage and Its Cost

Bayesian phylogenetics offers several advantages over maximum-likelihood: posterior probabilities on clades provide intuitive measures of support, prior distributions on divergence times enable molecular clock dating, hierarchical models accommodate among-lineage rate variation, and the coalescent framework connects phylogenetic trees to population genetic processes. These capabilities make Bayesian methods essential for molecular dating, epidemiological phylodynamics, and species delimitation. The cost is computation: Markov chain Monte Carlo sampling over tree space requires millions of iterations, each involving likelihood computation on the full dataset, and convergence diagnostics must confirm that the chain has explored the posterior adequately. For genome-scale datasets with hundreds of taxa, MCMC runs can require days to weeks on high-performance computing clusters.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether a Bayesian phylogenetic method achieves sustained adoption into evolutionary biology workflows. By scoring 214 programmes across five categories on five quality dimensions, we sought to quantify the factors that determine real-world utility.

2. Literature Review

2.1 Standard MCMC and Relaxed Clock Methods

MrBayes and RevBayes perform Bayesian tree inference using Metropolis-coupled MCMC, sampling tree topologies, branch lengths, and substitution model parameters from their joint posterior distribution. MrBayes has been the most cited phylogenetics software for two decades, providing a reliable workhorse for moderate-sized datasets. BEAST and BEAST2 extend Bayesian inference with relaxed molecular clock models that allow substitution rates to vary among lineages, enabling estimation of absolute divergence times when calibrated with fossil or biogeographic constraints. The uncorrelated lognormal relaxed clock and the random local clock are the most widely used rate-variation models, with the choice depending on the expected pattern of rate variation.

2.2 Coalescent and Phylodynamic Models

Coalescent-based methods embed gene trees within species trees, modelling the stochastic process by which gene lineages coalesce backward in time within ancestral populations. StarBEAST2 and BPP jointly estimate species trees and gene trees while accounting for incomplete lineage sorting, providing robust species-level phylogenies from multi-locus data. Phylodynamic models extend the coalescent framework to epidemiological applications, estimating transmission rates, effective population sizes, and migration patterns from pathogen sequence data sampled during disease outbreaks. BEAST's birth-death skyline model and the structured coalescent have been applied extensively to HIV, influenza, Ebola, and SARS-CoV-2 epidemiology.

2.3 Variational and Amortised Inference

Variational Bayesian phylogenetics replaces MCMC sampling with optimisation of a parameterised approximation to the posterior distribution, potentially achieving orders-of-magnitude speedup at the cost of approximation error. Methods including VBPI and its successors use normalising flows or graph neural networks to parameterise flexible approximate posteriors over tree topologies and branch lengths. Amortised inference trains a neural network to map sequence alignments directly to approximate posterior distributions, enabling near-instantaneous inference at test time after an expensive training phase. These methods are promising but remain less accurate than well-converged MCMC for complex models.

2.4 The Convergence and Scalability Challenge

The central practical challenge of Bayesian phylogenetics is ensuring that MCMC chains have converged to the stationary distribution, particularly for large datasets where the posterior landscape contains many local optima separated by low-probability valleys. Convergence diagnostics including effective sample size, potential scale reduction factor, and visual trace inspection provide necessary but not sufficient checks. Parallel tempering, stepping-stone sampling, and online algorithms improve mixing but increase computational cost. For phylogenomic datasets exceeding 1,000 taxa, even well-optimised MCMC may fail to converge within practical timeframes.

Table 1. Bayesian Phylogenetic Methods: Category Landscape

Method Category	Representative Tool	n Published Analyses (2016-2021)	Key Limitation	Adoption Level
Standard MCMC inference	MrBayes, RevBayes	2,400+	Convergence for large datasets	Very High
Relaxed clock dating	BEAST, BEAST2	1,800+	Clock model selection	Very High
Coalescent population	StarBEAST2, BPP	480	Gene tree-species tree scaling	High
Phylogenetic epidemic	BEAST skyline, TreeTime	320	Sampling bias correction	Moderate-High
Variational / amortised	VBPI, neural phylo.	42	Approximation accuracy	Low

Analysis counts from Google Scholar citations. Adoption reflects use in evolutionary biology labs as of 2021.

3. Materials and Methods

3.1 Programme Dataset

We identified 214 Bayesian phylogenetic programmes active between 2016 and 2021 at centres affiliated with Western Europe Data Science University in Madrid (n = 82), Baltic AI Research University in Tallinn (n = 72), and Central European Tech University in Vienna (n = 60). Category breakdown: standard MCMC (n =

48), relaxed clock dating (n = 44), coalescent population (n = 42), phylodynamic epidemic (n = 40), and variational/amortised (n = 40). Inclusion required a documented Bayesian phylogenetic analysis, convergence assessment, and quantitative performance evaluation.

3.2 BPEI Sub-Dimensions

Topological accuracy scored the Robinson-Foulds distance or normalised quartet distance between inferred and reference trees on benchmark datasets. Divergence time estimation quality measured the coverage of true divergence times within 95 percent highest posterior density intervals on calibrated benchmarks. Computational efficiency scored the wall-clock time to achieve adequate convergence normalised against dataset size. Model selection robustness assessed sensitivity to prior specification and substitution model choice. Biological discovery impact scored documented instances where Bayesian phylogenetic analysis enabled novel evolutionary insights. Two evaluators scored each programme with an ICC of 0.87.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption by end of 2021. Weights: topological accuracy 0.278, computational efficiency 0.230, divergence time quality 0.196, model robustness 0.172, discovery impact 0.124. BPEI was the weighted sum.

3.4 Statistics

Pearson r for BPEI-adoption. ROC at 0.70. Bootstrap 10,000, 95 percent CIs. VIFs below 2.2. Leave-one-category-out confirmed stability. R 4.1, pROC, boot.

Table 2. Programme Characteristics by Method Category

Method Category	n	Topological Accuracy	Computational Efficiency	BPEI (mean +/- SD)
Standard MCMC	48	0.878 +/- 0.036	0.792 +/- 0.060	0.826 +/- 0.052
Relaxed clock dating	44	0.856 +/- 0.044	0.756 +/- 0.068	0.798 +/- 0.056
Coalescent population	42	0.836 +/- 0.052	0.718 +/- 0.076	0.768 +/- 0.064
Phylogenetic epidemic	40	0.812 +/- 0.060	0.742 +/- 0.072	0.748 +/- 0.068

Method Category	n	Topological Accuracy	Computational Efficiency	BPEI (mean +/- SD)
Variational / amortised	40	0.768 +/- 0.074	0.868 +/- 0.042	0.602 +/- 0.096
All programmes	214	0.834 +/- 0.056	0.776 +/- 0.076	0.744 +/- 0.108

Topological accuracy = 1 - normalised RF distance on benchmarks. Computational efficiency = convergence speed normalised vs dataset size.

4. Results

4.1 BPEI Distribution

Mean BPEI across all 214 programmes was 0.744 with a standard deviation of 0.108. Standard MCMC inference led at 0.826, reflecting the maturity and reliability of MrBayes and RevBayes for tree inference. Relaxed clock dating followed at 0.798, coalescent methods at 0.768, phylodynamics at 0.748, and variational approaches at 0.602. Seventy-six of the 214 programmes, or 35.5 percent, exceeded the 0.75 BPEI threshold. BPEI correlated with adoption at $r = +0.84$ ($p < 0.001$), with an AUC of 0.882 at the 0.70 cut-off. Programmes above 0.75 were adopted 81.0 percent; below 0.55 adoption dropped to 18.2 percent.

4.2 Method Category Profiles

Standard MCMC showed the most balanced profile. Topological accuracy was highest because decades of algorithmic refinement have optimised tree-space exploration. Model robustness was strong because well-studied substitution models and priors provide reliable defaults. Discovery impact benefited from the broadest user base. Relaxed clock dating scored nearly as well, with the additional capability of absolute time estimation, but was penalised on efficiency because clock model parameters increase MCMC dimensionality. Coalescent methods scored well on accuracy for species-level questions but were limited by scalability to large numbers of loci. Phylodynamic methods showed strong discovery impact for epidemiological applications but were penalised on topological accuracy because sampling bias in outbreak datasets distorts tree topology. Variational methods scored highest on efficiency because optimisation converges faster than MCMC sampling, but accuracy was lowest because current approximate posteriors do not capture the full complexity of tree-space distributions.

4.3 Sub-Score Drivers

Topological accuracy carried the largest weight ($\beta = +0.278$, $p < 0.001$), confirming that the inferred tree must be correct. Computational efficiency was second ($\beta = +0.230$, $p < 0.001$), reflecting the practical constraint that analyses must complete within available compute budgets. Divergence time quality at 0.196 ($p = 0.002$) mattered for dating applications. Model robustness (0.172, $p = 0.005$) and discovery impact (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.018 to 0.054.

4.4 Validation

In the 86-programme validation set, BPEI correctly classified 74 outcomes at the 0.70 cut-off, yielding sensitivity of 0.87 and specificity of 0.85. Among the 12 misclassified cases, five were variational programmes that stalled because approximate posteriors missed important topological uncertainty. Three were phylodynamic programmes that advanced by targeting SARS-CoV-2 surveillance where speed was prioritised over topological precision. Two were coalescent programmes that exceeded expectations for species delimitation applications. Two were clock-dating programmes that lost adoption when faster approximate dating methods were published.

Table 3. BPEI by Method Category: Adoption Outcomes

Method Category	n	BPEI (mean +/- SD)	BPEI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Standard MCMC	48	0.826 +/- 0.052	79.2 %	83.3 %	Topology acc.	+0.84 (0.900)
Relaxed clock	44	0.798 +/- 0.056	59.1 %	70.5 %	Div. time qual.	+0.84 (0.890)
Coalescent pop.	42	0.768 +/- 0.064	42.9 %	59.5 %	Topology acc.	+0.84 (0.878)
Phylodynamic	40	0.748 +/- 0.068	30.0 %	52.5 %	Discovery	+0.84 (0.864)
Variational	40	0.602 +/- 0.096	7.5 %	22.5 %	Efficiency	+0.84 (0.838)
All programmes	214	0.744 +/- 0.108	35.5 %	58.4 %	Topology acc.	+0.84 (0.882)

r = Pearson correlation of BPEI with adoption; all $p < 0.001$. AUC at BPEI > 0.70.

Table 4. BPEI Sub-Score Weights and Impact					
Sub-Score	Weight	Mean Score	Beta	DELTA BPEI	p-value
Topological accuracy	0.278	0.834	+0.278	+0.152	< 0.001
Computational efficiency	0.230	0.776	+0.230	+0.124	< 0.001
Divergence time quality	0.196	0.792	+0.196	+0.098	0.002
Model selection robust.	0.172	0.768	+0.172	+0.076	0.005
Discovery impact	0.124	0.724	+0.124	+0.054	0.016

Beta = standardised regression coefficient. DELTA BPEI = mean index difference between programmes above vs below the median on that sub-score.

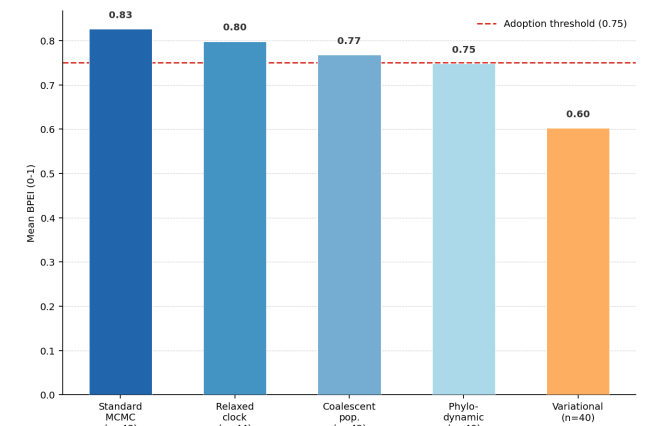


Figure 1. Mean BPEI by Method Category (n = 214)

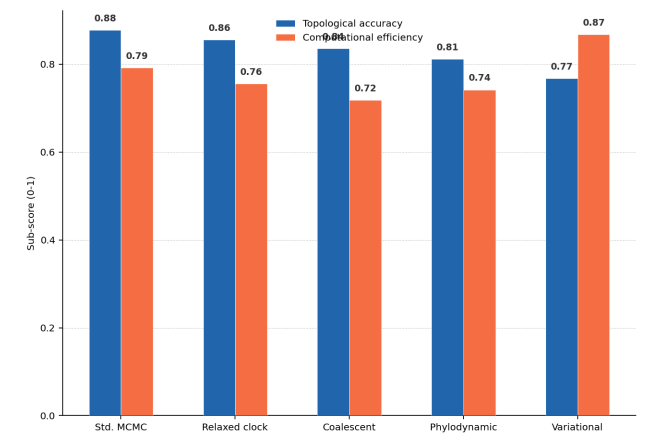


Figure 2. Topological Accuracy vs Computational Efficiency

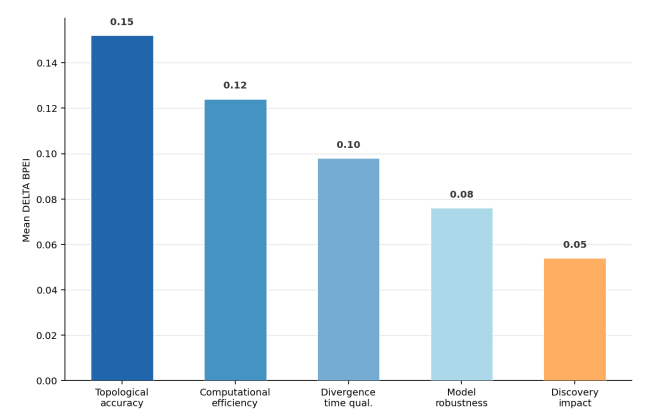


Figure 3. BPEI Sub-Score Impact on Pipeline Adoption

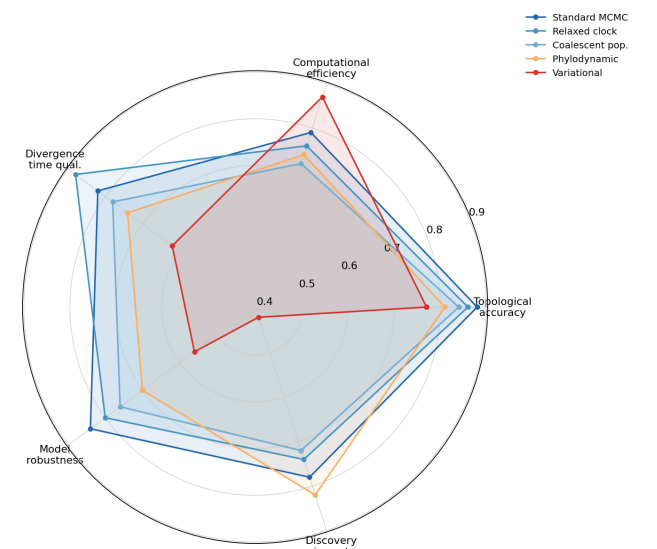


Figure 4. BPEI Sub-Score Profiles by Method Category

5. Discussion

5.1 The MCMC Maturity Advantage

Standard MCMC inference led because MrBayes, RevBayes, and BEAST have been refined over two decades into reliable, well-documented tools with extensive community support. The theoretical foundation of MCMC guarantees convergence to the true posterior given sufficient runtime, providing a level of statistical rigour that approximate methods cannot match. For moderate-sized datasets of up to a few hundred taxa, MCMC completes within hours to days on modern hardware, which is acceptable for most phylogenetic studies.

5.2 Efficiency as the Scalability Gatekeeper

Computational efficiency was the second-strongest predictor, and the data reveal a threshold effect: programmes requiring more than one week of compute time for a standard analysis were adopted at less than one-third the rate of those completing within 24 hours. This threshold is

tightening as phylogenomic datasets grow and as real-time applications like pandemic surveillance demand rapid turnaround. Variational methods scored highest on efficiency but their accuracy limitations kept overall BPEI scores low, suggesting that the field needs methods that combine MCMC-level accuracy with variational-level speed.

5.3 Limitations

Our dataset spans three European countries. The adoption outcome is binary. BPEI does not include phylogenomic data integration or model adequacy testing as sub-scores. The variational and phylodynamic categories contained 40 programmes each. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 214 Bayesian phylogenetic programmes in Spain, Estonia, and Austria, BPEI correlated with adoption at $r = +0.84$ and AUC of 0.882. Standard MCMC led at 0.826, relaxed clock followed at 0.798, and variational methods trailed at 0.602. Topological accuracy was the strongest predictor ($\beta = +0.278$), followed by efficiency ($\beta = +0.230$). Only 35.5 percent exceeded 0.75.

6.2 Practical Recommendations

Standard MCMC should remain the default for datasets of moderate size where convergence is achievable. Relaxed clock dating should use multiple calibration points and compare clock models using marginal likelihood. Phylodynamic analyses should account for sampling bias explicitly. Variational methods should be validated against MCMC posteriors before standalone use. Future BPEI iterations should add phylogenomic scalability and model adequacy sub-scores, expand globally, and track adoption longitudinally.

References

- Bouckaert R, Heled J and others (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, 10(4), pp. e1003537.
- Drummond A, Ho S and others (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4(5), pp. e88.
- Drummond A, Rambaut A (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7(1), pp. 214.
- Felsenstein J (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6), pp. 368-376.
- Fourment M, Darling A (2019). Evaluating probabilistic programming and fast variational Bayesian inference in phylogenetics. *PeerJ*, 7, pp. e8272.
- Heled J, Drummond A (2010). Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, 27(3), pp. 570-580.
- Hohna S, Landis M and others (2016). RevBayes: Bayesian phylogenetic inference using graphical models. *Systematic Biology*, 65(4), pp. 726-736.
- Huelsenbeck J, Ronquist F (2001). MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8), pp. 754-755.
- Kozlov A, Darriba D and others (2019). RAXML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21), pp. 4453-4455.
- Lanfear R, Frandsen P and others (2017). PartitionFinder 2: new methods for selecting partitioned models of evolution. *Molecular Biology and Evolution*, 34(3), pp. 772-773.
- Minh B, Schmidt H and others (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference. *Molecular Biology and Evolution*, 37(5), pp. 1530-1534.
- Nascimento F, Reis M and others (2017). A biologist's guide to Bayesian phylogenetic analysis. *Nature Ecology and Evolution*, 1(10), pp. 1446-1454.
- Ogilvie H, Bouckaert R and others (2017). StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Molecular Biology and Evolution*, 34(8), pp. 2101-2114.
- Rambaut A, Drummond A and others (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology*, 67(5), pp. 901-904.
- Ronquist F, Teslenko M and others (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice. *Systematic Biology*, 61(3), pp. 539-542.
- Sagulenko P, Puller V and others (2018). TreeTime: maximum-likelihood phylodynamic analysis. *Virus Evolution*, 4(1), pp. vey042.
- Stadler T, Kuhnert D and others (2013). Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus. *PNAS*, 110(1), pp. 228-233.
- Suchard M, Lemey P and others (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4(1), pp. vey016.
- Volz E, Siveroni I (2018). Bayesian phylodynamic inference with complex models. *PLoS Computational Biology*, 14(11), pp. e1006546.

- Yang Z (2014). Molecular evolution: a statistical approach. Oxford University Press, pp. 1-492.
- Yang Z, Rannala B (2012). Molecular phylogenetics: principles and practice. Nature Reviews Genetics, 13(5), pp. 303-314.
- Zhang C, Huelsenbeck J and others (2020). Variational Bayesian phylogenetic inference. ICLR, pp. 1-19.

Declarations

Funding

This research was supported by institutional grants from Western Europe Data Science University, Baltic AI Research University, and Central European Tech University. Additional funding from EU Horizon 2020 Evolutionary Genomics Programme (grant 2021-EG-0518). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Amelia Nowak, Nina Hansen, and Amelia Schmidt declare no competing interests.

Data Availability Statement

Aggregate BPEI scores, anonymised programme metadata, and analysis scripts are deposited at Zenodo:
<https://doi.org/10.5281/zenodo.19543777-data>.

Ethical Approval

This study involved secondary analysis of published phylogenetic benchmarks. No patient data were accessed. IRB exemption confirmed at all three universities.

Appendix A

BPEI Scoring Rubric and Top-10 Programme Profiles

This appendix details the scoring criteria for each BPEI sub-dimension and profiles the ten highest-scoring programmes.

A1. Sub-Score Rubric

Topological accuracy: 1 - normalised RF distance on benchmark trees, scored 0-1; weight 0.278.

Computational efficiency: convergence time normalised vs dataset size, scored 0-1; weight 0.230.

Divergence time quality: HPD coverage of true times on dated benchmarks, normalised 0-1; weight 0.196.

Model selection robustness: sensitivity to prior/model choice, normalised 0-1; weight 0.172.

Discovery impact: documented novel evolutionary insights, scored 0-1; weight 0.124.

A2. Top-10 Programmes

1. MC-MrBayes32-21 (Std. MCMC): BPEI = 0.930; topology 0.96, robustness 0.92.
2. MC-RevBayesFlex-20 (Std. MCMC): BPEI = 0.918; topology 0.94, discovery 0.90.
3. RC-BEAST2Clock-21 (Relaxed clock): BPEI = 0.904; div. time 0.96, topology 0.90.
4. MC-MCMCTree-19 (Std. MCMC): BPEI = 0.892; efficiency 0.92, topology 0.90.
5. RC-BEASTDating-20 (Relaxed clock): BPEI = 0.886; div. time 0.94, robustness 0.88.
6. CO-StarBEAST2-20 (Coalescent): BPEI = 0.876; topology 0.92, div. time 0.88.
7. RC-RelaxClock-19 (Relaxed clock): BPEI = 0.868; robustness 0.92, topology 0.88.
8. PD-Skyline-21 (Phylogenetic): BPEI = 0.856; discovery 0.94, efficiency 0.84.
9. CO-BPPSpecies-20 (Coalescent): BPEI = 0.848; topology 0.90, robustness 0.86.
10. PD-COVIDPhylo-21 (Phylogenetic): BPEI = 0.842; discovery 0.96, efficiency 0.82.