Choices that matter: the impact of substitution models on machine learning-based species delimitation inference

Matheus M. A. Salles[1*], Fabricius M. C. B. Domingos[1]

[1]Departamento de Zoologia, Universidade Federal do Paraná, Curitiba/PR, Brazil

[*] Author for correspondence: Matheus Salles, Programa de Pós-Graduação em Zoologia, Universidade Federal do Paraná (Departamento de Zoologia, Setor de Ciências Biológicas, Centro Politécnico, Avenida Cel. Francisco H Santos, Jardim das Américas, 81531-980, Curitiba/PR, Brasil). Email: matheusmaciel.salles@gmail.com

ORCID Matheus: https://orcid.org/0000-0002-1177-9844

ORCID Fabricius: https://orcid.org/0000-0003-2069-9317

Abstract

The choice of nucleotide substitution models is a cornerstone of phylogenetic inference, influencing the accuracy of the estimated evolutionary parameters and, by extension, demographic and species delimitation model selection. With the growing adoption of machine learning methods trained on simulated data, it remains unclear how the substitution model used during simulation training influences classifier performance and robustness when applied to empirical data, usually characterized by pervasive genomic heterogeneity. To address this gap, we conducted a controlled simulation study to evaluate the impact of substitution-model misspecification on supervised machine learning inference. We trained supervised classifiers on data simulated under three common substitution models (JC69, HKY, and GTR) and evaluated their performance in selecting the correct delimitation model from test datasets featuring a mixture of substitution processes across loci, a realistic scenario mimicking genomic heterogeneity. Our results demonstrate that classifiers trained under a single, simplistic substitution model generalized effectively to mixed-model test data, consistently identifying the true demographic model with high posterior probability (mean probability > 0.86 even using 100 SNPs), with highest performance plateauing beyond 600–800 SNPs. Notably, the differences in accuracy among classifiers trained under JC69, HKY, or GTR were minimal, indicating that the demographic signal captured by the site frequency spectrum predominates over substitution-model artifacts within the tested parameter space. However, this robustness is context-dependent. We caution that some extreme, though realistic, evolutionary scenarios (such as deep divergence, strong among-site rate variation, or protein-coding data) likely exceeds the conditions tested here and may severely degrade classifier performance. Furthermore, robust model selection does not imply accurate parameter estimation, as branch lengths and evolutionary rates remain sensitive to model misspecification. We conclude that for many practical applications in species delimitation, faster and computationally efficient training under simple models can be sufficient, provided it is coupled with rigorous validation, model-adequacy assessment, and an awareness of the limitations imposed by complex genomic data. Our findings offer a pragmatic framework for

integrating phylogenetic model selection with modern ML workflows, balancing computational efficiency with biological rigor.

Keywords: supervised learning; random forest; site frequency spectrum; phylogenetics; simulation.

Introduction

The selection of nucleotide substitution models is a critical step in phylogenetic analysis, directly influencing the accuracy of varied evolutionary inferences, including but not limited to tree topology, branch lengths, and divergence times (Yang 1996; Naser-Khdour et al. 2019). These mathematical models approximate the complex process of sequence evolution by parameterizing factors such as transition-transversion bias, equilibrium base frequencies, rate heterogeneity across sites, and codon-position effects (see Reviews in Sullivan & Joyce 2005; Yang & Rannala 2012). Due to the super-exponential growth in possible tree topologies with increasing taxa and the computational cost of likelihood calculations, commonly used substitution models can be interpreted as necessarily simplified abstractions of biological reality.

The simplest models (e.g., JC-69: Jukes & Cantor 1969) assume uniform rates and state sets across sites, treating sequence evolution as an unconstrained random process. While tractable and parameter-sparse, such models are often biologically inadequate, as ignoring rate heterogeneity can produce biased estimates (Yang 1996). Consequently, most models incorporate subsets of higher complexity, such as substitution biases and among-site rate variation modeled via a discrete gamma distribution (Sullivan & Joyce 2005; Yang & Rannala 2012; Arenas 2015). Therefore, an inevitable trade-off exists between biological realism, statistical identifiability, and computational tractability.

Biological complexity further extends beyond rate variation to include site-specific constraints on permissible states. Functional pressures (such as those acting on protein active sites, codon positions, or RNA structures) restrict the repertoire of acceptable substitutions (Echave et al. 2016). This challenge is further compounded, for example, by pervasive heterogeneity in evolutionary processes across the genome. For instance, functionally constrained regions evolve under distinct

patterns compared to more flexible regions, creating heterogeneity that a single model cannot capture (Xia 2000). Furthermore, stationarity, reversibility, and homogeneity (SRH) violations frequently occur in specific genomic contexts (e.g., third codon positions, viral and mitochondrial genes), and these violations are genuine features of the data rather than artifacts of sampling noise (Naser-Khdour et al. 2019). This implies that different genomic partitions can yield statistically distinct phylogenetic patterns, underscoring the importance of partition-specific model assessment.

To address heterogeneity, partitioned analyses and codon models allow different substitution processes for distinct loci or codon classes (Lanfear et al. 2012). However, this approach introduces a new trade-off: too few partitions risk underfitting, while excessive partitioning can lead to overfitting and increased parameter error (Lanfear et al. 2012; Gupta & Vadde 2023). Furthermore, although site-specific preferences can, in principle, be inferred (Meyer & Haeseler 2003), accurate estimation of such specific substitution models requires extensive data to overcome phylogenetic correlations and the large number of involved parameters (Puller et al. 2020). Model selection criteria (e.g., AIC, BIC, among many others) are consequently used to identify an optimal balance between fit and complexity by grouping sites that evolve under similar processes (Posada & Crandall 2001; Ripplinger et al. 2010).

This general trade-off between model complexity and biological realism becomes even more pronounced when considering different types of molecular data. In particular, sensitivity to substitution model choice is greater for protein data than for nucleotide data, mainly due to the larger amino acid state space (20 vs. 4), which increases the influence of model assumptions on inference accuracy. Specifically, empirical evidence for nucleotide data suggests that relatively complex models (e.g., GTR+I+G) can often function as reliable defaults (Abadi et al. 2019). In contrast, the prevailing view in protein phylogenetics is that explicit model selection is essential. The consequences are especially acute for protein evolution and ancestral sequence reconstruction (ASR), where inferences under best-fitting models yield superior topologies and branch lengths (Del Amparo & Arenas 2022, 2023). Conversely, inappropriate models, particularly those with divergent amino

acid exchangeability matrices, can introduce systematic biases that propagate across internal nodes of a tree, ultimately compromising downstream biological interpretations (Del Amparo & Arenas 2022).

Naturally, as these biases influence the estimation of phylogenetic relationships, their effects extend to downstream analyses such as species delimitation and demographic inference. Failure to account for model heterogeneity can bias essential evolutionary parameters (such as divergence times, effective population sizes, and migration rates), thus potentially compromising the distinction of species limits (Momigliano et al. 2021; Tiley et al. 2023). Because substitution models shape the likelihood surface from which gene trees and branch lengths are estimated, the uncertainty in these trees directly influences coalescent-based estimates of population parameters and species boundaries. Consequently, misspecified models (for example, ignoring across-site rate variation, compositional heterogeneity, heterotachy, codon structure in coding loci, or differences among loci that merit separate partitions) can produce systematic errors in tree topology and branch-length scaling, increase gene-tree estimation error, and thereby alter inferred levels of genealogical discordance that delimitation methods interpret as evidence for or against lineage independence. Contemporary species delimitation methods, such as those available in BEAST (Baele et al. 2025) and BPP (Flouri et al. 2018), allow integration of substitution models within the multispecies coalescent, but they commonly require user choices about partitioning and model parametrization; if those choices are not tested or if model uncertainty is not propagated, downstream delimitation and demographic inferences may be imprecise.

Building upon this foundation, machine learning (ML) has emerged as a powerful tool for navigating complex model spaces in species delimitation and demographic inference (Salles & Domingos 2025). These approaches, specially supervised ones, are typically trained on a vast amount of simulations (Schrider & Kern 2018). Furthermore, ML applications used in species delimitation commonly learn to map patterns in multidimensional feature vectors, composed of population genetics summary statistics, or to perform model selection of demographic or delimitation models

(Salles & Domingos 2025). A critical, yet often underappreciated, assumption of such a framework is that the joint distribution of features in the training data (simulations) must match that of the empirical data to which the classifier is applied. This is because the learned decision boundaries are intrinsically tied to the data parameter space on which the model was trained. Consequently, the substitution model chosen for simulation becomes a fundamental pillar of the entire ML pipeline.

This assumption, however, is challenged by the inherent properties of genomic data. Genomic datasets are characterized by pervasive heterogeneity in evolutionary processes, meaning that a single, uniform substitution model is biologically implausible for most genome-scale analyses. If an ML classifier is trained on alignments simulated under a simplistic or uniform model (e.g., JC69, or GTR applied homogeneously), the distribution of its input features will likely reflect this artificial homogeneity. Consequently, when such a classifier is applied to empirical data comprising distinct partitions (e.g., exons, introns) that evolved under divergent substitution processes, the resulting feature vectors are likely to be drawn from a different, distorted distribution. This distributional shift causes the classifier to operate in regions of feature space it was not trained on, potentially leading to poorly calibrated probabilities, misclassification, and reduced predictive performance. Ultimately, if this error propagates forward, parameter estimation and model selection may become biased, potentially leading to incorrect inferences in species delimitation. Therefore, the accuracy of ML-based inferences is not only a function of the evolutionary model being tested but is also tightly linked to the biological realism of the substitution models used during the training phase.

In this context, the present study evaluates the effect of substitution-model heterogeneity on ML-based inference of species limits and demographic models. Using controlled multi-locus simulations and supervised classifiers, we compare uniform *versus* partition-aware model selection, assess how partitioning schemes (that is, how finely the data are subdivided into partitions with distinct substitution models) trades off model fit and classifier performance across a range of SNP densities, and characterize the direction and magnitude of biases that arise when among-locus heterogeneity is ignored. Finally, we integrate these results into practical recommendations for

combining phylogenetic model-selection and ML workflows in species-delimitation and demographic inference.

Material & Methods

Overview and software

All training and test datasets were generated with *popai v1.0* (https://github.com/SmithLabBio/popai), a Python-based framework for simulating population genomic data under user-specified demographic and mutational models, with an emphasis on species delimitation. *popai* orchestrates simulation runs through configuration files that specify the phylogenetic background, migration matrix, population assignment, and parameter space. The software natively accommodates key evolutionary processes, including divergence, secondary contact, and divergence with ongoing migration. Global keys in the configuration file define elements such as random seed, number of replicates, substitution model, prior distribution for mutation rates, flags for symmetric versus asymmetric migration, and options to enable secondary contact or divergence-with-gene-flow models. *popai* generates simulation commands for the underlying coalescent and sequence simulators, logs metadata for each replicate, and preserves the complete configuration, ensuring reproducibility and traceability of all experimental steps.

Demographic models

We evaluated five competing demographic scenarios (Fig. 1), which included divergence-only, divergence with secondary contact, and divergence with gene flow between non-sister populations. The species tree topology, with three populations (A, B, and C), was fixed with 10 diploid individuals sampled per population and 20 loci per dataset. Effective population sizes were kept constant across all populations. Migration was modeled through a Boolean migration matrix that specified eligible population pairs. We restricted rates to symmetric values in the range $[10^{-5}, 10^{-4}]$ per generation and limited each model to a single migration event. Secondary contact models were

activated via a flag, introducing migration at the midpoint between the most recent divergence and the present, persisting until the present. In contrast, divergence-with-gene-flow models, in which migration begins immediately after divergence and ceases halfway to the next split, were not considered in this study. Mutation rates followed a uniform prior $U(5\times10^{-9}, 5\times10^{-7})$ substitutions/site/generation.

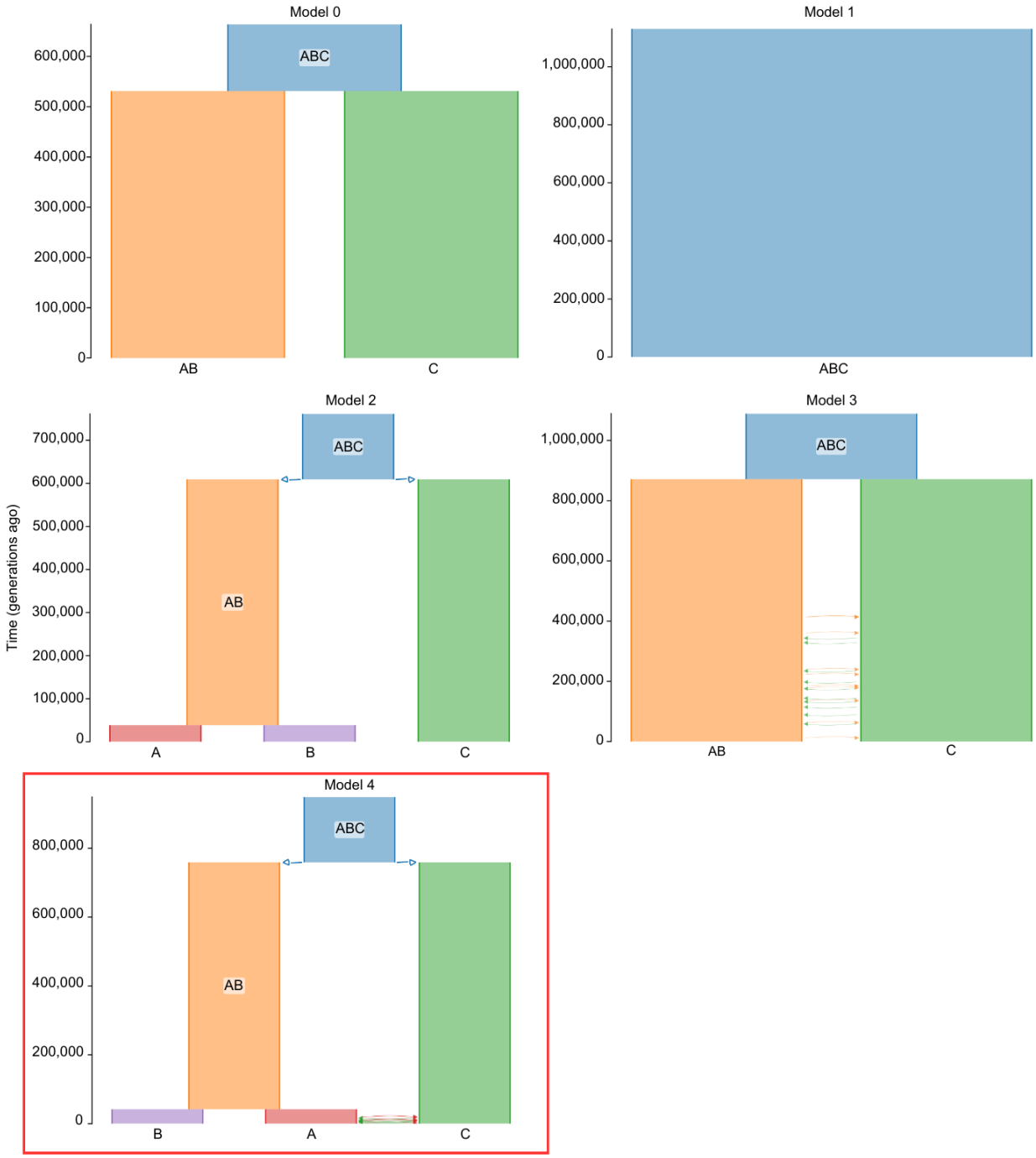

Figure 1. Five competing demographic models tested in the present study. Models include scenarios with divergence without gene flow, divergence with secondary contact between sister populations, and divergence with asymmetric gene flow between non-sister populations. The model used to generate the empirical test datasets (assumed as the true one) is highlighted in red. Time is given in generations before the present.

Simulation design

Training data for the five demographic scenarios (Fig. 1) were generated under three substitution models (JC69, HKY, and GTR). Sequence lengths ranged from 500 to 3,500 bp in 500 bp increments, corresponding approximately to 100, 300, 600, 800, 1,000, 1,200, and 1,400 segregating sites (SNPs). These SNP counts were used directly to build the site frequency spectrum (SFS). We here define a *regime* as a specific combination of substitution model and matrix size (i.e., one substitution model applied to one alignment length & SNP count). For each regime, we simulated the five fixed demographic scenarios, resulting in a total of 3 substitution models × 7 alignment sizes × 5 demographic scenarios = 105 training datasets. All simulations used the same tree topology (Fig. 1), so the only sources of variation across regimes were the substitution model and alignment length. Importantly, we did not perform mixed-model training: each training dataset was generated under a single substitution model.

Test datasets were generated independently from the training sets and we assumed Model 4 (Fig. 1) as the true demographic scenario. For each SNP matrix size (100-1,400), we simulated 100 independent replicates per substitution model. Unlike the training sets, test datasets consisted of heterogeneous substitution regimes: each dataset included 20 loci drawn from a fixed mixture of models (e.g., 7 loci JC69, 7 loci HKY, 6 loci GTR). This design mimics realistic heterogeneity across loci. Because all data were fully simulated, no missing genotypes were present and down-projection was unnecessary.

Classifier training and evaluation

The folded SFS (based on minor allele frequencies) was used as the main summary statistic. For each test replicate, 10 SFS were generated, preserving stochastic variance across replicates. These were used as input features for classification. For each substitution model regime, a separate Random Forest (RF) classifier was trained using the corresponding datasets. Features included SFS bins exported from *popai*. Each trained classifier was then evaluated against the 100 independent test

datasets per SNP size. Evaluation mostly considered model-mismatched conditions (e.g., training under JC69, testing with mixed JC69/HKY/GTR). Probability outputs across demographic models were recorded for each replicate. Performance metrics included mean, median, standard deviation, amplitude, and 95% confidence intervals, which were estimated from the 100-test set resamples.

The full workflow of dataset generation, classifier training, and evaluation is summarized in Figure 2.

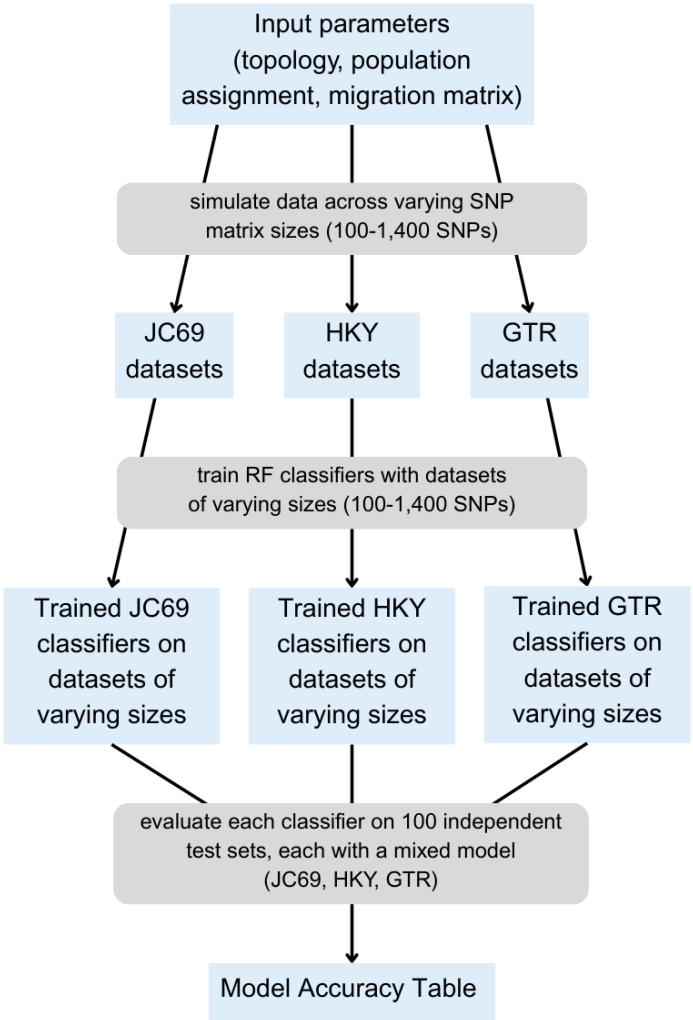

Figure 2. Workflow of the experimental design. Input files were processed with *popai* to generate multilocus datasets of varying SNP sizes. Then, training datasets simulated under different substitution regimes were used to train RF classifiers. Each classifier was evaluated on 100 independent test datasets per SNP size, simulated under the true demographic model and including a fixed mixture of substitution models.

Reproducibility

All simulations were executed with fixed random seeds to ensure reproducibility. Complete configuration files, scripts, together with input and output data are archived in Zenodo [https://doi.org/10.5281/zenodo.17274456], including: 1. The species tree file (tree.nex); 2. The population assignment file (populations.txt), 3. The migration matrix file (migration.txt), and 4. Simulation configuration files for each substitution model × alignment length combination. Together with step-by-step execution instructions, these resources allow full replication of the workflow, from simulation to classifier training and evaluation.

Results

Trained classifiers consistently identified the correct demographic model with the highest posterior probability across all simulated regimes. That is, for every dataset used for testing, the correct scenario (Model 4, as previously described) was always ranked as the most probable, demonstrating strong classification performance. Mean probabilities of correct model assignment, along with their 95% confidence intervals, are summarized in Table 1. Overall, accuracy increased with the number of SNPs included in the alignment, although performance stabilized beyond approximately 600-800 SNPs. For instance, under the JC69 training regime, the mean probability of correctly inferring the model increased from 0.8693 (95% CI: 0.7100-0.9780) with 100 SNPs to 0.9234 (95% CI: 0.8120-0.9800) with 800 SNPs, with comparable values at larger matrix sizes. A similar pattern was observed for HKY and GTR training regimes, both of which exhibited slightly higher accuracies for larger alignments, reaching mean probabilities above 0.93 with 1,200 SNPs. Importantly, performance differences among training regimes (JC69, HKY, and GTR) were minimal, suggesting that the classifiers were robust to the specific substitution model used during training. For example, with 1,200 SNPs, mean probabilities were 0.9264 (JC69), 0.9380 (HKY), and 0.9318 (GTR), with overlapping confidence intervals.

Table 1. Mean probability of correct model selection and corresponding 95% confidence intervals (in brackets) across different training regimes (JC69, HKY, and GTR) and matrix sizes (100-1,400 SNPs). For every test set, the correct evolutionary scenario (Model 4) was always inferred with the highest posterior probability.

Substitution model used during training ↓ / matrix size →	100 SNPs	300 SNPs	600 SNPs	800 SNPs	1,000 SNPs	1,200 SNPs	1,400 SNPs
JC69	0.8693, [0.7100, 0.9780]	0.9088, [0.8100, 0.9800]	0.9239, [0.8160, 0.9860]	0.9234, [0.8120, 0.9800]	0.9170, [0.8320, 0.9720]	0.9264, [0.8340, 0.9780]	0.9207, [0.8400, 0.9820]
HKY	0.8680, [0.7100, 0.9660]	0.9114, [0.8000, 0.9840]	0.9197, [0.8000, 0.9800]	0.9277, [0.8520, 0.9800]	0.9276, [0.8460, 0.9800]	0.9380, [0.8660, 0.9840]	0.9247, [0.8500, 0.9760]
GTR	0.8669, [0.6980, 0.9660]	0.9134, [0.8100, 0.9820]	0.9295, [0.8300, 0.9800]	0.9304, [0.8520, 0.9860]	0.9248, [0.8460, 0.9760]	0.9318, [0.8720, 0.9760]	0.9299, [0.8420, 0.9820]

Across classifiers from all SNP sizes and training regimes, the demographic model most frequently assigned the second-highest probability (and thus the primary source of classification confusion) was consistently Model 2. The mean probability assigned to this alternative model decreased with increasing SNP counts, falling from approximately 0.089 with 100 SNPs to around 0.046 with 1,200 SNPs, which aligns with the observed narrowing of confidence intervals and improved discriminatory power with larger datasets.

Discussion

We used supervised classifiers based on simulated data to evaluate how substitution-model assumptions influence the selection of species delimitation models. In our controlled framework (training separate classifiers under single substitution models and testing on independent datasets that deliberately mixed substitution models) classifiers were consistently able to recover the true model with high posterior probability. Even with only 100 SNPs the mean posterior for the correct model exceeded ~0.86, and performance rose with increasing SNP counts before reaching a plateau near ~600-800 SNPs. Confidence intervals narrowed with more SNPs, indicating greater precision as phylogenetic signals increased.

Two aspects of these results are particularly informative. First, classifier performance improved with additional SNPs but exhibited diminishing returns beyond a moderate data size; this suggests there is a practical "sweet spot" where additional data yield marginal gains. Second, differences among classifiers trained under JC69, HKY and GTR were small across the SNP range we explored. For example, at 1,200 SNPs mean correct-model posteriors were 0.9264 (JC69), 0.9380 (HKY) and 0.9318 (GTR) with overlapping confidence intervals. Within our simulation design these results indicate that delimitation signals (or demographic ones) encoded in the SFS and related summaries dominated the idiosyncratic differences induced by substitution models, so that even classifiers trained under simpler models often generalized well to heterogeneous test sets.

Furthermore, the pattern of model misassignment offers a critical biological interpretation. Across all SNP sizes and training regimes, Model 2 (divergence between three species, without gene flow) was consistently identified as the second-most probable scenario. This systematic confusion is phylogenetically meaningful. Although Model 2 lacks gene flow, it shares an identical topology with the true model (Model 4, divergence with ongoing migration), which likely results in similar site frequency spectrum signatures that require substantial data to disentangle. The consistent decrease in probability assigned to Model 2 as SNP number increased demonstrates that larger datasets progressively sharpen the discrimination between these topologically equivalent but demographically distinct histories. In contrast, the remaining models attracted negligible support, underscoring their clear distinguishability from the true scenario.

These findings are useful but must be interpreted within the limits of our experimental design. First, all results derive from simulations in which the evolutionary scenario, including species topology, sampling scheme, locus number, among many other factors, were fixed by design. Furthermore, the test sets were independent but drawn from the same family of delimitation scenarios. Real empirical data often violate simplifying assumptions in ways not captured by our simulations. For instance, Naser-Khdour et al. (2019) have shown that model violation in phylogenetic analysis is common and heterogeneous across genomic partitions, and that distinct partitions from a single

dataset can produce statistically discordant trees. Such partition-level violations alter the distribution of summary statistics used by ML classifiers and therefore may degrade predictive performance if not identified and handled.

Second, several forms of model misspecification that we did not explore in this study can materially influence classifier performance and therefore limit the scope of our conclusions. In particular, molecular phenomena such as strong among-site rate heterogeneity, highly skewed equilibrium base frequencies, deep sequence divergence (saturation), and time-dependent shifts in site preferences, may all change the distribution of site patterns and hence the SFS and other summary statistics potentially used as inputs to the classifiers. These changes can move empirical feature vectors outside the parameter space sampled by our training simulations, producing miscalibration and misclassification. Work on site-specific models (Puller et al. 2020) shows that estimating per-site preferences can improve local fit when many moderately diverged sequences are available, but identifiability problems and residual branch-length errors persist, a reminder that even more realistic models do not always recover true divergence times or eliminate bias. Likewise, protein datasets (Del Amparo & Arenas 2023), with a larger state space and more complex exchangeability structure, are substantially more sensitive to model choice than the SNP data we simulated. Taken together, these points imply that the robustness we observed for classifiers trained under simple nucleotide models should not be extrapolated to datasets exhibiting extreme heterogeneity or deep divergence. For such cases, strategies like model-adequacy screening, targeted simulations covering those regimes, and validation (e.g., mixed-model training or posterior-predictive checks) before applying classifiers to empirical data, may still be necessary.

It is also important to stress that robustness in model selection does not imply robustness in parameter estimation. Our experiments targeted model-choice accuracy (a classification task), but parameter inference (e.g., divergence times, migration rates, substitution rates) might be more sensitive to substitution-model misspecification because these estimates depend directly on branch-length scaling and substitution–time mapping. Potentially, even when model choice is correct,

incorrectly specified substitution models may bias branch lengths, distort SFS summaries, and mislead parameter estimates. Thus, successful model choice should be treated as necessary but not sufficient, with follow-up sensitivity analyses under alternative models and posterior predictive calibration to evaluate estimator reliability.

Finally, the reproducibility and stability of classification accuracy across many independent replicates reinforce the robustness of our pipeline, built upon frameworks such as *popai*. Supervised learning approaches based on summary statistics like the SFS are now well established in evolutionary inference and species delimitation (Schrider & Kern 2018; Smith & Carstens 2020). Our results extend this literature by showing that, within the tested ranges, substitution-model differences alone do not undermine classifier accuracy. This finding supports a pragmatic strategy: training classifiers under simpler substitution models can be computationally more efficient without substantially compromising performance, provided that empirical safeguards (such as partition-level adequacy tests, calibration diagnostics, and validation on more realistic simulations) are employed.

This point becomes particularly relevant as we consider that the computational burden of data simulation escalates rapidly with increasing model complexity. Integrating realistic features such as heterogeneous substitution processes, heterotachy, or among-locus heterogeneity can impose prohibitive runtime and memory costs, especially for genome-wide datasets or diverse demographic scenarios. This trade-off is fundamental to supervised machine learning, where simulated data are paramount. When robust simulators exist (as in population genetics) they permit the generation of virtually unlimited training data, a capability bounded chiefly by computational resources (Korfmann et al. 2023). In line with this, we have documented comprehensive runtime benchmarks across our simulation regimes; these data, which clearly illustrate the associated computational trade-offs, will be presented in the final manuscript. Collectively, these factors underscore the critical need to balance biological realism with practical feasibility in the construction of training datasets for supervised inference.

Future directions

Several avenues can deepen and generalize our findings. A first step is to implement mixed-model training, where substitution models are randomly assigned across training replicates. This would test whether robustness extends to heterogeneous training regimes and whether demographic signal truly overrides substitution-model artifacts. Expanding demographic heterogeneity in training (e.g., sampling a wider range of divergence times, migration rates, population sizes, bottlenecks, and expansions) would also test robustness under more realistic evolutionary scenarios. Likewise, testing across alternative tree topologies and sampling schemes would help confirm that robustness is not tied to a particular design.

Improving interpretability is another priority. Decision-tree algorithms like RF allow feature-importance analyses, which could identify which joint-SFS bins or summary statistics most influence classification decisions, and link them to population-genetic expectations (e.g., excess rare variants under growth or migration). Benchmarking against other approaches, such as DIYABC-RF (Collin et al., 2021), under the same simulations would provide context on accuracy, sensitivity to misspecification, and computational cost. Broader comparisons, including alternative machine learning algorithms, could reveal distinct strengths (higher accuracy, better interpretability, or improved scalability) relative to RF. Finally, constructing formal learning curves with larger SNP matrices (e.g., 2,000-5,000 SNPs) would quantify marginal gains of additional data and help define sequencing targets for empirical studies.

Conclusions

Within the controlled parameter space explored in this study, supervised classifiers trained under single substitution models generalized well to mixed-model test data and offered a computationally efficient route to demographic/delimitation model selection. The reproducibility of posterior probabilities across many replicates supports their practical reliability. At the same time, model violation in empirical datasets, deep divergence, and site-specific complexity impose real

information and identifiability limits. Robust model choice does not guarantee accurate parameter estimation, underscoring the need for empirical adequacy checks, calibration diagnostics, and validation under more complex simulation regimes. For future applied work, we recommend pairing efficient training under simple substitution models with explicit adequacy testing, targeted validation under complex models, and interpretability analyses. This dual strategy may balance computational efficiency with empirical rigor, supporting stronger and more reliable biological inferences.

Acknowledgements

We are especially grateful to André Luiz Gomes de Carvalho, Fernanda de Pinho Werneck, Marcio Roberto Pie and Renato José Pires Machado for their helpful comments on earlier versions of the text.

Author contributions

MMAS: study design, analyses, lead writing. FMCBD: study design, writing final draft. Both authors read, commented, and approved the final draft of the manuscript.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data underlying this article, including phylogenetic datasets, corresponding trees, input and output files for all analyses, and any other relevant supplementary files are available in Zenodo, at https://doi.org/10.5281/zenodo.17274456.

Funding

This work was supported by the Brazilian federal agency *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES) through a PhD scholarship awarded to MMAS. FMCBD acknowledges a productivity fellowship from CNPq (#311997/2025-2).

Generative AI in scientific writing

During the preparation of this work the authors used ChatGPT to review their own original translation of the text from Portuguese to English. After using this tool/service, both authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Abadi S, Azouri D, Pupko T, Mayrose I. 2019. Model selection may not be a mandatory step for phylogeny reconstruction. *Nat Commun.* 10:934.

Arenas M. 2015. Trends in substitution models of molecular evolution. *Front Genet.* 6:319.

Baele G, et al. 2025. BEAST X for Bayesian phylogenetic, phylogeographic and phylodynamic inference. *Nat Methods.* In press.

Collin FD, et al. 2021. Extending approximate Bayesian computation with supervised machine learning to infer demographic history from genetic polymorphisms using DIYABC random forest. *Mol Ecol Resour.* 21(8):2598–2613.

Del Amparo R, Arenas M. 2022. Consequences of substitution model selection on protein ancestral sequence reconstruction. *Mol Biol Evol.* 39(7):msac144.

Del Amparo R, Arenas M. 2023. Influence of substitution model selection on protein phylogenetic tree reconstruction. *Gene.* 865:147336.

Echave J, Spielman SJ, Wilke CO. 2016. Causes of evolutionary rate variation among protein sites. *Nat Rev Genet.* 17:109–121.

Flouri T, Jiao X, Rannala B, Yang Z. 2018. Species tree inference with BPP using genomic sequences and the multispecies coalescent. *Mol Biol Evol.* 35(10):2585–2593.

Gupta MK, Vadde R. 2023. Next-generation development and application of codon model in evolution. *Front Genet.* 14:1091575.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. *Mammalian protein metabolism.* New York: Academic Press. p. 21–132.

Korfmann K, Gaggiotti OE, Fumagalli M. 2023. Deep learning in population genetics. *Genome Biol Evol.* 15(2):evad008.

Lanfear R, Calcott B, Ho SYW, Guindon S. 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol Biol Evol.* 29(6):1695–1701.

Meyer S, von Haeseler A. 2003. Identifying site-specific substitution rates. *Mol Biol Evol.* 20(2):182–189.

Momigliano P, Florin AB, Merilä J. 2021. Biases in demographic modeling affect our understanding of recent divergence. *Mol Biol Evol.* 38(7):2967–2985.

Naser-Khdour S, et al. 2019. The prevalence and impact of model violations in phylogenetic analysis. *Genome Biol Evol.* 11(12):3341–3352.

Posada D, Crandall KA. 2001. Selecting the best-fit model of nucleotide substitution. *Syst Biol.* 50(4):580–601.

Puller V, Sagulenko P, Neher RA. 2020. Efficient inference, potential, and limitations of site-specific substitution models. *Virus Evol.* 6(2):veaa066.

Ripplinger J, Sullivan J. 2010. Assessment of substitution model adequacy using frequentist and Bayesian methods. *Mol Biol Evol.* 27(12):2790–2803.

Salles MMAS, Domingos FMCB. 2025. Towards the next generation of species delimitation methods: an overview of machine learning applications. *Mol Phylogenet Evol.* 108368.

Schrider DR, Kern AD. 2018. Supervised machine learning for population genetics: a new paradigm. *Trends Genet.* 34(4):301–312.

Smith ML, Carstens BC. 2020. Process-based species delimitation leads to identification of more biologically relevant species. *Evolution.* 74(2):216–229.

Sullivan J, Joyce P. 2005. Model selection in phylogenetics. *Annu Rev Ecol Evol Syst.* 36(1):445–466.

Tiley GP, et al. 2023. Estimation of species divergence times in presence of cross-species gene flow. *Syst Biol.* 72(4):820–836.

Xia X. 2000. Phylogenetic relationship among horseshoe crab species: effect of substitution models on phylogenetic analyses. *Syst Biol.* 49(1):87–100.

Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol.* 11(9):367–372.

Yang Z, Rannala B. 2012. Molecular phylogenetics: principles and practice. *Nat Rev Genet.* 13(5):303–314.