# Phylodynamics modeling with unsupervised Bayesian neural networks

Gabriele Marino, Ugnė Stolz, Daniele Silvestro, Cecilia Valenzuela Aguï, Tanja Stadler

November 13, 2025

**Abstract**

Phylodynamic modeling quantifies the parameters that govern the growth of phylogenetic trees. Current extensions using generalized linear models (GLMs) allow the integration of external covariates, such as host traits or mobility data, but are limited to linear and additive effects. To overcome this limitation, we introduce a Bayesian neural network (BNN) approach within a Markov chain Monte Carlo (MCMC) phylodynamic framework. Rather than sampling parameters directly, the method samples BNN weights, enabling the network to learn functional mappings from predictors to rates of speciation, extinction, migration, or transmission. Simulations show that this framework accurately recovers predictor–parameter relationships, avoids overfitting, and performs robustly across a range of dynamics. Using explainable AI techniques, we show that our framework is able to correctly identify the predictors with the greatest influence, providing both interpretability and precision. We further validate the method on empirical datasets, estimating COVID-19 migration rates during its early spread in Europe and inferring speciation and extinction dynamics in platyrrhine phylogenies, proving that our unsupervised BNN framework provides a flexible and powerful tool for modeling complex epidemiological and macroevolutionary processes.

## 1 Introduction

Phylodynamic modeling provides a powerful framework for linking evolutionary processes with population dynamics by leveraging genetic data to study how lineages diversify, spread, and change over time. This integrative approach has been extensively applied across both epidemiology and macroevolution. In epidemiology, phylodynamics enables the estimation of key epidemiological parameters, such as the basic reproduction number, $R_t$, which quantifies the average number of secondary infections caused by a single infected individual in a fully susceptible population. These methods have been instrumental in understanding the spread of pathogens including SARS-CoV-2, Ebola, Zika, and HIV [1, 2, 3, 4]. In macroevolution, phylodynamic approaches allow inference of speciation and extinction rates, offering insights into the diversification dynamics of clades [5, 6, 7].

Beyond simply estimating diversification rates, these methods can be used to test hypotheses regarding temporal variation in evolutionary rates, the impact of ecological factors on lineage dynamics, and trait-dependent diversification. For example, birth-death skyline (BDSKY) models permit the reconstruction of time-varying evolutionary processes in epidemiological systems [8], while Bayesian analysis of macroevolutionary mixtures (BAMM) allows the detection and quantification of heterogeneity in speciation and extinction rates across phylogenetic trees, accounting for both temporal and lineage-specific variation [9, 10]. Multi-type birth-death (MTBD) models have further extended these approaches by enabling the estimation of distinct evolutionary dynamics for different compartments or subpopulations within a single system [11, 12]. In macroevolutionary research, state-speciation-extinction (SSE) models have become widely used for exploring how trait variation influences diversification rates, providing a mechanistic link between organismal traits and lineage dynamics [13, 14]. Additionally, generalized linear models (GLMs) have been employed to integrate external non-genetic data into phylodynamic analyses, and are frequently used in epidemiology to model the effect of covariates such as host mobility or environmental factors on parameters like migration rates between subpopulations [15, 16, 17].

Despite these important advances, current approaches remain constrained in several ways. BAMM, BDSKY, and MTBD models are not inherently hypothesis-driven and cannot directly incorporate external covariates. SSE models are restricted to analyses based solely on trait data, while GLMs are

limited by their functional form and often fail to capture complex interactions between variables. To address these challenges, we introduce a new framework based on Bayesian neural networks (BNNs). Neural networks, a class of graphical models, have gained increasing attention due to their capacity as universal approximators, capable of modeling arbitrarily complex functions when provided with sufficient parameters [18]. Typically, they are employed in supervised learning, where mappings between inputs and outputs are inferred from training data. However, in phylodynamic contexts, generating comprehensive training datasets is impractical due to the vast diversity of possible evolutionary scenarios. While recent studies have explored BNNs in the analysis of fossil and non-phylogenetic data [19], we propose extending their application to phylogenetic and molecular inference, offering a flexible and powerful alternative to existing approaches.

Through extensive simulations, we demonstrate that our framework reliably recovers the true relationships between non-genetic covariates and phylodynamic parameters, avoids overfitting despite its large parameter space, and matches the performance of GLMs in linear settings while significantly outperforming them under nonlinear dynamics. Using tools from explainable AI, we further show that the model can identify which predictors meaningfully contribute to epidemiological or macroevolutionary dynamics, providing both accurate inference and interpretability. Finally, we validate the method on empirical datasets, estimating migration rates underlying COVID-19 early spread in Europe from viral sequence alignments and mobility data and inferring speciation and extinction dynamics across living and extinct platyrrhine species. The results of our analyses highlight the flixibility and robustness of our approach, unlocking new avenues to model complex real-world epidemiological and macroevolutionary dynamics.

# 2 Methods

## 2.1 The phylodynamic-MLP model

Phylodynamic inference aims to estimate the epidemiological or evolutionary parameters that generate an observed phylogenetic tree $T$. In birth-death models, these parameters typically include the transmission or birth rate, $\lambda$, the removal or death rate, $\mu$, and the sampling rate, $\psi$ [12]. The posterior distribution over these parameters is given by

$$P(\theta \mid T) \propto P(T \mid \theta)\, P(\theta),$$

where $\theta = (\lambda, \mu, \ldots)$ represents the set of phylodynamic parameters. The likelihood $P(T \mid \theta)$ describes the probability of observing the tree topology and branch lengths under the specified birth-death process. Phylodynamic inference thus seeks to uncover the parameters that most plausibly generated the observed tree, while naturally accounting for uncertainty through the posterior distribution.

A multilayer perceptron (MLP) [18] is a type of feedforward artificial neural network that models complex, nonlinear relationships between inputs and outputs. An MLP consists of an input layer, one or more hidden layers, and an output layer. Each layer applies a linear transformation followed by a nonlinear activation function. Formally, for a layer $l$ with input vector $\mathbf{x}^{(l-1)}$, weights $\mathbf{W}^{(l)}$, and biases $\mathbf{b}^{(l)}$, the output is

$$\mathbf{x}^{(l)} = f\big(\mathbf{W}^{(l)}\mathbf{x}^{(l-1)} + \mathbf{b}^{(l)}\big),$$

where $f(\cdot)$ is a nonlinear activation function such as ReLU or sigmoid. By stacking multiple layers, the MLP can approximate arbitrary continuous functions, allowing it to model complex mappings from input data to target outputs. Classically, multilayer perceptrons (MLPs) are employed in supervised learning, where the network is trained to map input data to known output targets by minimizing a loss function, such as mean squared error for regression or cross-entropy for classification. In the supervised setting, the network weights and biases are treated as parameters that are optimized during training.

In this work, we adopt an unsupervised use of MLPs to integrate external, non-phylogenetic data into phylodynamic analyses. The MLP maps these input features (hereafter referred to as *predictors* or *covariates*) to a subset of the phylodynamic parameters, denoted $\theta_{\text{MLP}}$, while priors are placed on the network weights $\mathbf{W}$ and biases $\mathbf{b}$. The remaining parameters, $\theta_{\text{std}}$, are not predicted by the MLP and are instead sampled directly from their standard priors. This separation allows the model to exploit external covariates for parameters that can be informed by additional data, while retaining classical Bayesian sampling for the remaining parameters.

To ensure that the MLP produces parameter values within biologically or mathematically meaningful ranges, the choice of output activation functions can be tailored to the specific parameter being modeled. For instance, parameters that are inherently bounded between zero and one, such as probabilities, can be mapped through a sigmoid or logistic transformation. More generally, sigmoidal-like functions with user-specified upper and lower bounds allow parameters to be constrained to finite intervals, while exponential or softplus activations can enforce strict positivity. By aligning the output activation function with the natural support of each parameter, the model avoids generating implausible values and improves the efficiency of posterior exploration.

As in standard phylodynamic analyses, Markov chain Monte Carlo (MCMC) [20] is used to jointly sample from the posterior

$$P(\mathbf{W}, \mathbf{b}, \theta_{\text{std}} \mid T) \propto P(T \mid \theta_{\text{MLP}}(\mathbf{W}, \mathbf{b}), \theta_{\text{std}})\, P(\mathbf{W}, \mathbf{b})\, P(\theta_{\text{std}}),$$

with operators acting on both the network parameters and the standard prior-sampled parameters. Each MCMC iteration proposes new values for $\mathbf{W}, \mathbf{b}$, and $\theta_{\text{std}}$, which are then combined into the full set of phylodynamic parameters and evaluated under the tree likelihood. This approach propagates uncertainty from both the network and standard parameters, providing a fully Bayesian integration of external data into phylodynamic inference.

## 2.2 Simulation settings

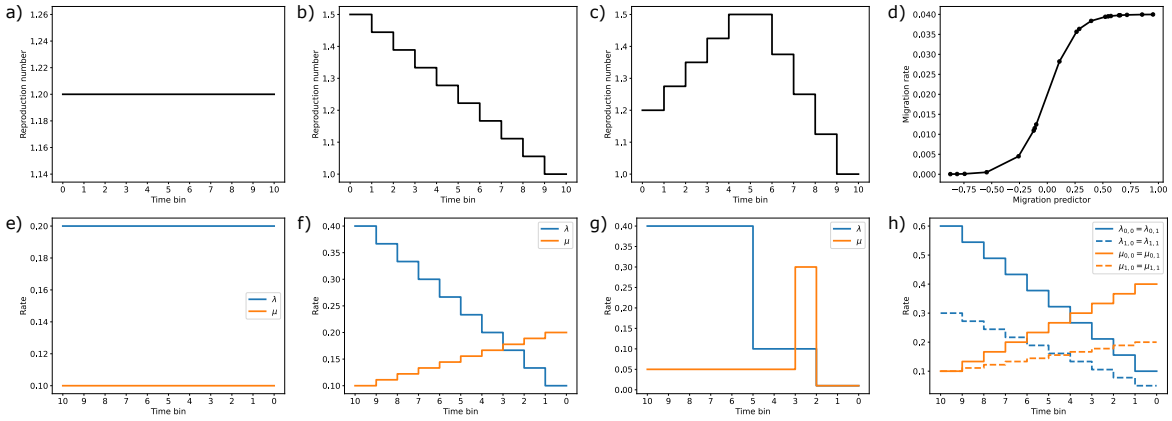


Figure 1: The different scenarios considered in the simulation studies. (a) Epi-1 — constant reproduction number; (b) Epi-2 — linearly decreasing reproduction number; (c) Epi-3 — non-monotonic reproduction number; (d) Epi-4 — migration rates between five populations predicted by a continuous covariate; (e) FBD-1 — constant speciation and extinction rates; (f) FBD-2 — linearly decreasing speciation rate and linearly increasing extinction rate; (g) FBD-3 — non-linear speciation rate with a spike in the extinction rate; (h) FBD-4 — linearly decreasing speciation rate and linearly increasing extinction rate across four groups of species defined by two binary traits (only one trait affects diversification).

We benchmarked our approach on **eight simulation scenarios**, grouped into two classes of multitype birth-death models [12]: epidemiological processes (with removal after sampling) and fossilized birth-death (FBD) processes [21] (with ancestor sampling and complete present-day sampling). Figure 1 graphically illustrates the inference targets for each scenario with respect to their predictors.

All epidemiological scenarios (a-d) share: the same become-uninfectious rate $\delta = \mu + \psi = 0.07$, the same sampling proportion $s = \frac{\psi}{\mu+\psi} = 0.15$, and the same simulation time $T = 250$. Scenarios *a*-*c* are characterized by a unique population, and the inference task is to estimate a piece-wise constant (skyline) reproduction number $R_t = \frac{\lambda}{\mu+\psi}$ using time as predictor. $R_t$ is constant in scenario *a*, linearly decreasing in scenario *b*, and non-monotonic in scenario *c*. Scenario *d* is characterized by five populations, with constant reproduction numbers for each of them (respectively, 0.8, 1.0, 1.2, 1.4 and 1.6), and the task is to infer the 20 pairwise migration rates using a continuous covariate

as migration predictor. To generate these rates, the migration predictors were first sampled from a uniform distribution $\mathcal{U}(-1, 1)$ and then mapped into target migration rates using a sigmoid function.

All FBD scenarios (e-h) assume the same constant ancestor sampling rate $\psi = 0.2$ and simulation time $T = 35$, and the inference task is to estimate skyline speciation and extinction rates across 10 time bins. Predictors are given by time and, for scenario $h$, by two additional binary traits. Scenario $e$ assumes constant speciation and extinction rates; scenario $f$ features a linearly decreasing speciation rate and a linearly increasing extinction rate; scenario $g$ models nonlinear decreases in the speciation rate, while the extinction rate exhibits a sharp upward spike in the second half of the simulation. Scenario $h$ involves four different groups of species, distinguished by the combination of values of two binary traits. However, the second of these traits does not affect evolutionary dynamics, meaning that the four populations can be further grouped into two pairs that share identical speciation and extinction rates. For each pair, the speciation rate decreases linearly while the extinction rate increases linearly.

To assess the robustness of our approach to misspecifications in the predictor set—particularly with respect to the inclusion of irrelevant predictors—we incorporated an additional randomly generated predictor into all simulation studies. This predictor was constructed randomly and independently of the inference target, and resampled for each tree. In scenario $d$, it consisted of a uniformly distributed random value in $[1, -1]$ for each of the 20 population pairs, whereas for all other (skyline) scenarios, it was defined as a random walk with standard normal increments across the time bins.

## 2.3 Inference settings

For each scenario, we generated 100 phylogenetic trees with tip counts ranging from 200 to 500, and performed inference on each of them using the BDMM-Prime package [22] implemented in BEAST2 [23]. Each experiment was repeated with multiple inference approaches: a predictor-agnostic model, a GLM implementation [17], and three different Bayesian MLP architectures. The MLPs used in our analyses had hidden layer sizes of $(3, 2)$, $(16, 8)$, and $(32, 16)$ neurons, with ReLU activations in the hidden layers and a sigmoid activation in the output layer. For analyses requiring inference of multiple parameters (e.g., speciation and extinction rates in FBD scenarios), a separate MLP was used for each target parameter, so that the output layer of each MLP consisted of a single neuron. The size of the input layer varied depending on the experiment and was set equal to the number of predictors characterizing that experiment. In accordance with best practices for GLMs and MLPs, we normalized each predictor to the range $[0, 1]$ prior to inference.

All MCMC runs were executed with a chain length of 10,000,000 steps. For predictor-agnostic and GLM models, we used uniform priors: $\mathcal{U}(1, 5)$ for reproduction numbers, $\mathcal{U}(0, 0.05)$ for migration rates, and $\mathcal{U}(0, 2)$ for speciation and extinction rates. For the MLPs, we instead applied a normal prior $\mathcal{N}(0, 1)$ to the weights, while constraining the output via a sigmoid output activation function so that predictions remained within the same bounds as the uniform priors defined for predictor-agnostic and GLM models. For all analyses, we discarded the first 10% of samples as burn-in and assessed convergence by verifying that the effective sample size (ESS) for each parameter exceeded 200.

## 2.4 Evaluation metrics

To compare the performance of the different inference approaches (predictor-agnostic, GLM, and MLP), we computed the posterior median of each target parameter from every MCMC run, along with the corresponding 95% credible intervals (CIs). Model performance within each scenario was evaluated based on three complementary metrics: the mean absolute error (MAE) between the posterior median estimates and the true parameter values across the 100 simulated trees, the average width of the 95% CIs, and the empirical coverage of these intervals.

Formally, let $\boldsymbol{\theta}_p = (\theta_{p_1}, \ldots, \theta_{p_K})$ denote the vector of true values for parameter $p$, where $K$ is the number of elements in the vector (e.g., time bins or population pairs), and let $\hat{\boldsymbol{\theta}}_p^i = (\hat{\theta}_{p_1}^i, \ldots, \hat{\theta}_{p_K}^i)$ be the corresponding posterior median estimates obtained from tree $i$, with $i = 1, \ldots, N$ and $N = 100$ simulated trees per scenario. The mean absolute error (MAE) for each parameter and each tree was computed as

$$\mathrm{MAE}_p^i = \frac{1}{K} \sum_{k=1}^{K} \left| \hat{\theta}_{p_k}^i - \theta_{p_k} \right|,$$

and the overall MAE for a given parameter in a given scenario was then obtained by averaging across all trees:

$$\mathrm{MAE}_p = \frac{1}{N}\sum_{i=1}^{N}\mathrm{MAE}_p^i.$$

The width of the 95% credible interval (CI) for element $k$ of parameter $p$ in tree $i$ was defined as

$$\mathrm{CI}_{p_k}^i = \hat{\theta}_{p_k}^{i,97.5\%} - \hat{\theta}_{p_k}^{i,2.5\%}.$$

The overall CI width for parameter $p$ in a given scenario was then obtained by first averaging over all elements of the parameter vector for each tree and then across all trees:

$$\mathrm{CI}_p = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{K}\sum_{k=1}^{K}\mathrm{CI}_{p_k}^i\right).$$

Coverage was defined as the proportion of true parameter values contained within their respective 95% credible intervals. For each element $k$ of parameter $p$ in tree $i$, let

$$C_{p_k}^i = \mathbb{I}\left(\hat{\theta}_{p_k}^{i,2.5\%} \le \theta_{p_k} \le \hat{\theta}_{p_k}^{i,97.5\%}\right),$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. The overall coverage for parameter $p$ was then obtained by averaging over all elements and all trees:

$$\mathrm{Coverage}_p = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{K}\sum_{k=1}^{K}C_{p_k}^i\right).$$

MAE quantifies overall estimation accuracy, $\mathrm{CI}_{\mathrm{width}}$ reflects the precision of posterior uncertainty, and coverage measures the calibration of the credible intervals. Together, these three criteria provide a comprehensive assessment of both point estimates and uncertainty quantification across models and scenarios.

To assess the computational efficiency and sampling performance of each inference framework, we computed the mean effective sample size per hour. For each model and scenario, we first estimated the effective sample size (ESS) of all inference targets obtained from the posterior samples. The mean ESS across all parameters was then divided by the total wall-clock time (in hours) required for the corresponding inference run. This procedure was repeated across all replicate phylogenetic trees, yielding the mean ESS/hour averaged over all replicates for each scenario and model combination. This measure allows for direct comparison of sampling efficiency between models, independently of absolute runtime or dimensionality of the parameter space.

### 2.4.1 Interpretability analysis

To better understand and interpret the behavior of multilayer perceptrons (MLPs), which are often considered "black-box" models due to their complex and non-linear structure, we employed two explainable AI (xAI) techniques [24]: partial dependence plots (PDPs) [25] and Shapley values (SHAP) [26]. These methods allow us to quantify the influence of individual predictors on model outputs and to visualize their effects across the predictor space. In particular, we used them to assess the robustness of MLPs to irrelevant input predictors, as well as to gain general insights into how the models leverage relevant features in the inference task.

PDPs show the average effect of a single predictor $x_S$ (or a subset of predictors) on the model's output by marginalizing over the distribution of all other features $x_C$. For a prediction function $f(\mathbf{x})$, the partial dependence of feature subset $S$ is defined as:

$$PD_S(x_S) = \mathrm{E}_{\mathbf{x}_C}\left[f(x_S, \mathbf{x}_C)\right],$$

where $C$ is the complement of $S$, and $p(\mathbf{x}_C)$ is the marginal distribution of the remaining features.

PDPs are particularly useful because they provide an interpretable, model-agnostic visualization of how a predictor influences the prediction on average, even in the presence of complex nonlinear interactions captured by models such as MLPs. If the predictor is relevant, the PDP typically exhibits

a structured relationship (e.g., monotonic trend, threshold effect, or nonlinear curve). Conversely, if the predictor is irrelevant, the PDP should remain approximately flat, indicating that varying this feature does not systematically alter the model output. This makes PDPs well-suited for diagnosing spurious dependencies and detecting overfitting when irrelevant predictors are included in the input space. However, PDPs also have important limitations, as they implicitly assume independence between features when marginalizing, which may lead to misleading plots if predictors are correlated.

SHAP addresses these issues by offering a game-theoretic framework for feature attribution. For a model $f$ and input instance $\mathbf{x}$, the Shapley value for feature $i$ is:

$$\phi_i(f, \mathbf{x}) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \Big[ f_{S \cup \{i\}}(\mathbf{x}_{S \cup \{i\}}) - f_S(\mathbf{x}_S) \Big],$$

where $F$ is the full feature set and $S$ a subset excluding $i$, and the term $f_S(\mathbf{x}_S)$ is defined through a conditional expectation:

$$f_S(\mathbf{x}_S) = \mathrm{E}_{\mathbf{x}_{F \setminus S}} \left[ f(\mathbf{x}_S, \mathbf{x}_{F \setminus S}) \mid \mathbf{x}_S \right],$$

that is, the model output averaged over the distribution of the missing features given the known ones.

SHAP values provide local, prediction-level explanations, but in practice, feature importances are often summarized by averaging the absolute Shapley values across all instances in the input dataset:

$$\Phi_i(f) = \frac{1}{|X|} \sum_{j=1}^{|X|} |\phi_i(f, \mathbf{x}^{(j)})|,$$

where $|X|$ is the number of samples. This provides a global quantification of each feature's contribution, complementing the local and marginal insights offered by PDPs, and facilitates a more comprehensive evaluation of how predictors influence MLP outputs across scenarios and simulated trees.

In our study, each MLP has multiple configurations due to the weights sampled during the MCMC process. To compute PDPs and SHAP values, we first subsampled 100 MCMC samples per run, yielding 100 distinct MLP models per tree. For each input feature, we computed the PDP values for each model by discretizing the feature space into a grid of 10 bins, and then aggregated them into a single PDP per tree by taking the median value at each grid point. Similarly, SHAP values were computed for each of the 100 subsampled MLP models, then marginalized so that the feature contributions sum to 1, and finally aggregated by computing the median feature importance per tree.

### 2.5 Empirical study

## 3 Results

### 3.1 Simulation-based Evaluation

Our implementation consistently matched or exceeded the performance of state-of-the-art phylodynamic models across a diverse range of simulation scenarios, as measured by all considered evaluation metrics. Table 1, Table 2, and Table 3 summarize the overall performance of the different inference methods across all scenarios and inference targets, in terms of mean absolute error (MAE), coverage, and average credible interval width, respectively. The Bayesian MLP approach reliably produced accurate parameter estimates with well-calibrated uncertainty, while remaining computationally feasible despite the additional inference cost.

In scenario Epi-1 (Figure 2a), the inference target is a constant skyline reproduction number $R_t$. The smallest MLP model tested MLP-3/2 (two hidden layers with 3 and 2 neurons), achieved the best performance, with a MAE of 0.032 and a high coverage of 0.98. It also demonstrated the highest precision, reporting the smallest average credible interval width of 0.29. The GLM model followed, with slightly lower performance (MAE: 0.052, coverage: 0.96, average CI width: 0.35), while the larger MLP architectures—MLP-16/8 and MLP-32/16—performed comparably to the GLM, with MAEs of 0.058 and 0.065, coverages of 0.99 and 0.98, and average CI widths of 0.51 and 0.67, respectively. The predictor-agnostic model performed the worst, with a MAE of 0.11; although coverage remained acceptable (0.93), its average credible interval width was very large (0.96). A similar pattern emerged for scenario FBD-1 (Figure 3a), which shares the same constant dynamics as the inference target.

| Mean Absolute Error (MAE) | | | | | | |
|---|---|---|---|---|---|---|
| **Scenario** | **Var.** | PA | GLM | MLP-3/2 | MLP-16/8 | MLP-32/16 |
| Epi-1 | $R_t$ | 0.11 | 0.052 | **0.032** | 0.058 | 0.065 |
| Epi-2 | $R_t$ | 0.048 | **0.0064** | 0.038 | 0.0075 | 0.017 |
| Epi-3 | $R_t$ | 0.055 | 0.12 | 0.084 | 0.054 | **0.036** |
| Epi-4 | $M$ | 0.011 | 0.0085 | 0.0068 | **0.0031** | 0.0045 |
| FBD-1 | $\lambda$ | 0.12 | 0.032 | **0.018** | 0.025 | 0.023 |
| | $\mu$ | 0.077 | 0.0051 | **0.0039** | 0.017 | 0.027 |
| | Avg. | 0.098 | 0.019 | **0.011** | 0.021 | 0.025 |
| FBD-2 | $\lambda$ | 0.075 | 0.030 | 0.026 | 0.021 | **0.020** |
| | $\mu$ | 0.057 | 0.0088 | 0.022 | **0.0031** | 0.0072 |
| | Avg. | 0.066 | 0.020 | 0.024 | **0.012** | 0.013 |
| FBD-3 | $\lambda$ | 0.061 | 0.16 | **0.011** | 0.015 | 0.025 |
| | $\mu$ | 0.097 | 0.053 | **0.020** | 0.025 | 0.024 |
| | Avg. | 0.079 | 0.11 | **0.015** | 0.020 | 0.025 |
| FBD-4 | $\lambda$ | 0.13 | 0.025 | 0.021 | 0.013 | **0.011** |
| | $\mu$ | 0.1754 | 0.0160 | 0.0261 | 0.012 | **0.0070** |
| | Avg. | 0.153 | 0.021 | 0.023 | 0.013 | **0.0090** |

Table 1: Mean absolute error (MAE) across all scenarios for all inference frameworks considered (PA is the predictor-agnostic model). Bold indicates the best, underlined indicates the second-best.

| Coverage | | | | | | |
|---|---|---|---|---|---|---|
| **Scenario** | **Var.** | PA | GLM | MLP-3/2 | MLP-16/8 | MLP-32/16 |
| Epi-1 | $R_t$ | 0.93 | 0.96 | 0.98 | 0.99 | 0.98 |
| Epi-2 | $R_t$ | 0.95 | 0.98 | 0.98 | 0.99 | 0.98 |
| Epi-3 | $R_t$ | 0.96 | 0.75 | 0.91 | 0.98 | 0.98 |
| Epi-4 | $M$ | 0.91 | 0.35 | 0.79 | 0.97 | 0.90 |
| FBD-1 | $\lambda$ | 0.90 | 0.90 | 0.97 | 0.98 | 0.98 |
| | $\mu$ | 0.97 | 0.96 | 0.97 | 0.97 | 0.93 |
| | Avg. | 0.94 | 0.93 | 0.97 | 0.97 | 0.95 |
| FBD-2 | $\lambda$ | 0.93 | 0.87 | 0.95 | 0.96 | 0.97 |
| | $\mu$ | 0.96 | 0.96 | 0.95 | 0.98 | 0.98 |
| | Avg. | 0.94 | 0.91 | 0.95 | 0.97 | 0.97 |
| FBD-3 | $\lambda$ | 0.96 | 0.23 | 0.85 | 0.87 | 0.92 |
| | $\mu$ | 0.97 | 0.32 | 0.73 | 0.73 | 0.78 |
| | Avg. | 0.96 | 0.28 | 0.79 | 0.80 | 0.85 |
| FBD-4 | $\lambda$ | 0.95 | 0.91 | 0.96 | 0.98 | 0.98 |
| | $\mu$ | 0.89 | 0.90 | 0.92 | 0.98 | 0.97 |
| | Avg. | 0.92 | 0.91 | 0.94 | 0.98 | 0.97 |

Table 2: Coverage across all scenarios for all inference frameworks considered (PA denotes the predictor-agnostic model).

| Average 95% CI Width | | | | | | |
|---|---|---|---|---|---|---|
| **Scenario** | **Var.** | PA | GLM | MLP-3/2 | MLP-16/8 | MLP-32/16 |
| Epi-1 | $R_t$ | 0.96 | 0.35 | 0.29 | 0.51 | 0.67 |
| Epi-2 | $R_t$ | 0.89 | 0.34 | 0.36 | 0.50 | 0.63 |
| Epi-3 | $R_t$ | 1.1 | 0.42 | 0.44 | 0.62 | 0.76 |
| Epi-4 | $M$ | 0.041 | 0.014 | 0.021 | 0.022 | 0.020 |
| FBD-1 | $\lambda$ | 0.44 | 0.14 | 0.11 | 0.17 | 0.22 |
| | $\mu$ | 0.45 | 0.11 | 0.093 | 0.13 | 0.15 |
| | Avg. | 0.44 | 0.12 | 0.10 | 0.15 | 0.18 |
| FBD-2 | $\lambda$ | 0.35 | 0.13 | 0.14 | 0.18 | 0.21 |
| | $\mu$ | 0.39 | 0.10 | 0.10 | 0.14 | 0.17 |
| | Avg. | 0.37 | 0.12 | 0.12 | 0.16 | 0.19 |
| FBD-3 | $\lambda$ | 0.40 | 0.17 | 0.16 | 0.20 | 0.22 |
| | $\mu$ | 0.39 | 0.071 | 0.074 | 0.088 | 0.095 |
| | Avg. | 0.40 | 0.12 | 0.12 | 0.14 | 0.16 |
| FBD-4 | $\lambda$ | 0.68 | 0.17 | 0.20 | 0.25 | 0.30 |
| | $\mu$ | 0.79 | 0.14 | 0.1775 | 0.24 | 0.28 |
| | Avg. | 0.73 | 0.16 | 0.19 | 0.24 | 0.29 |

Table 3: Average 95% credible interval width across all scenarios for all inference frameworks considered (PA denotes the predictor-agnostic model).

| **Scenario** | **PA** | **GLM** | **MLP-3/2** | **MLP-16/8** | **MLP-32/16** |
|---|---|---|---|---|---|
| Epi-1 | 39878.25 | 30543.81 | 12750.74 | 6630.82 | 4586.71 |
| Epi-2 | 58356.51 | 49470.21 | 13622.51 | 11485.42 | 8598.30 |
| Epi-3 | 54439.66 | 43323.65 | 10997.51 | 11408.65 | 7643.71 |
| Epi-4 | 162.19 | 175.93 | 137.84 | 106.11 | 79.70 |
| FBD-1 | 57313.24 | 41452.32 | 12060.29 | 11601.46 | 7513.56 |
| FBD-2 | 36272.02 | 23972.37 | 5120.69 | 5309.66 | 3614.67 |
| FBD-3 | 29514.58 | 20281.02 | 1919.92 | 1770.44 | 1835.22 |
| FBD-4 | 71.49 | 61.05 | 8.96 | 10.92 | 11.32 |

Table 4: Mean effective sample size (ESS) per hour across all scenarios and inference frameworks considered (PA denotes the predictor-agnostic model), averaged over all inference targets and all replicate trees for each scenario.

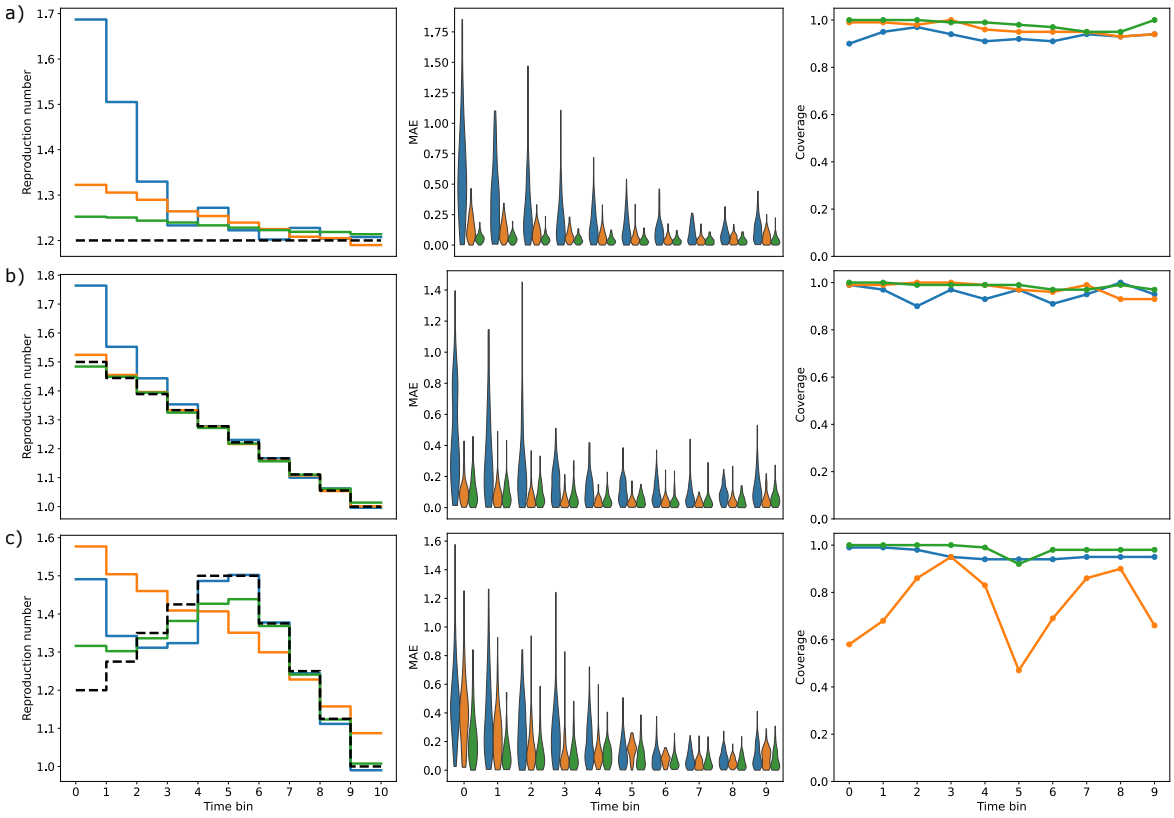


Figure 2: Results for the epidemiological skyline scenarios: (a) Epi-1; (b) Epi-2; (c) Epi-3. For each inference framework—predictor-agnostic model (blue), GLM (orange), and the best-performing MLP in terms of MAE (Epi-1: MLP-3/2; Epi-2: MLP-16/8; Epi-3: MLP-32/16, shown in green)—the panels display, from left to right: the median reproduction number estimates across all replicate trees; the distribution of the mean absolute error (MAE) across trees; and the coverage of the 95% credible intervals.

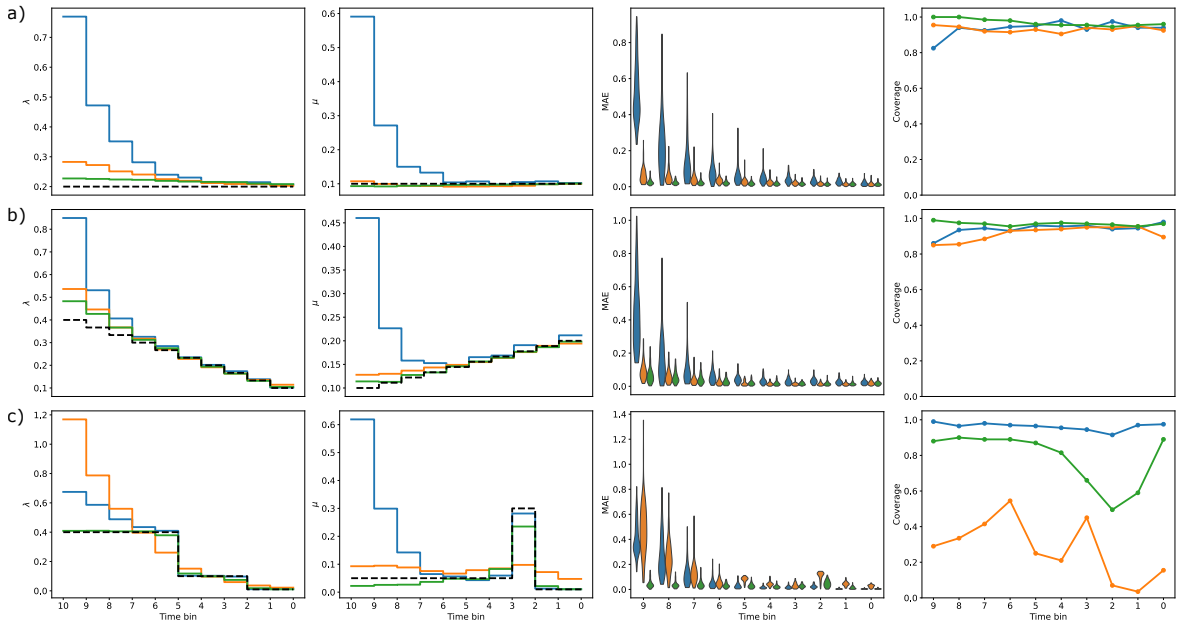


Figure 3: Results for the birth–death skyline scenarios: (a) FBD-1; (b) FBD-2; (c) FBD-3. For each inference framework—predictor-agnostic model (blue), GLM (orange), and the best-performing MLP in terms of MAE (FBD-1: MLP-3/2; FBD-2: MLP-16/8; FBD-3: MLP-3/2, shown in green)—the panels show, from left to right: the median speciation rate ($\lambda$) estimates across all replicate trees; the median extinction rate ($\mu$) estimates across all replicate trees; the distribution of the mean absolute error (MAE) across trees for both rates combined; and the mean coverage of the 95% credible intervals for the aggregated rates.

Here, the smallest MLP again outperformed all other models, followed by the GLM and the larger MLP architectures, which achieved performances similar to the GLM. The predictor-agnostic model again showed poor accuracy, coupled with very wide credible intervals despite reasonable coverage.

Figure 2b corresponds to scenario Epi-2, in which the reproduction number $R_t$ decreases linearly over time. In this setting, the GLM achieved the best performance, with a MAE of 0.0064, coverage of 0.98, and an average credible interval width of 0.34. The MLP-16/8 closely followed, showing a MAE of 0.0075, coverage of 0.98, and average CI width of 0.36. The larger MLP architectures also performed well, with MLP-32/16 achieving a MAE of 0.017 and coverage of 0.98, though with a wider average CI of 0.63. The smallest MLP, MLP-3/2, showed a higher MAE of 0.038, coverage of 0.98, and a narrow average CI width of 0.36. The predictor-agnostic model again performed the worst, with a MAE of 0.048, coverage of 0.95, and a very large average CI width of 0.89. For scenario FBD-2, which also featured linear dynamics (Figure 3b), the best performance was achieved by MLP-16/8 and MLP-32/16, with MAEs of 0.012 and 0.013, respectively. These were followed closely by the GLM (MAE: 0.020) and the smallest MLP-3/2 (MAE: 0.024). In terms of coverage, all MLP models outperformed the GLM, achieving values above 0.95, whereas the GLM reported 0.91. Once more, the predictor-agnostic model exhibited the poorest performance, with a MAE of 0.066, coverage of 0.94, and a large average CI width of 0.37. Similar patterns were observed for scenario FBD-4, which also featured linear dynamics. In this scenario, the MLP-32/16 and MLP-16/8 models achieved the best performance in terms of both MAE and coverage, while maintaining reasonable credible interval widths. The GLM, although close to the MLPs in terms of MAE and acceptable in absolute terms, was the most certain model, with the narrowest average credible interval width of 0.1558 (compared to 0.1875, 0.2443, and 0.2901 for MLPs of increasing sizes). However, this came at the cost of lower coverage (0.9066), while the MLP models achieved higher coverages of 0.9414, 0.9767, and 0.9743, respectively. As in previous scenarios, the predictor-agnostic model performed the worst, exhibiting a very high MAE of 0.1529 and wide credible intervals of 0.7325.

Figure 2c presents the results for scenario Epi-3, in which the reproduction number $R_t$ follows a nonlinear trajectory, increasing during the first half of the simulation and decreasing thereafter. In this

setting, the largest MLP architecture achieved the best performance, with a MAE of 0.036 and coverage of 0.98. It was closely followed by the MLP-16/8 and the predictor-agnostic model, which obtained similar MAEs of 0.054 and 0.055, with good coverage (0.98 and 0.96, respectively). The smallest MLP model underperformed, with a MAE of 0.084 and coverage of 0.91, yet still outperformed the GLM, which yielded a high MAE of 0.12 and low coverage of 0.75. In terms of average credible interval width, the MLPs reported reasonably sized intervals, ranging from 0.44 to 0.76 with increasing model size, whereas the predictor-agnostic model showed larger uncertainty with a width of 1.1. In scenario Epi-4, where the task was to estimate migration rates between five subpopulations given a nonlinear sigmoidal predictor, MLPs of intermediate and larger size again performed best. The MLP-16/8 achieved the lowest MAE of 0.0031 and coverage of 0.97, followed by the MLP-32/16 with MAE of 0.0045 and coverage of 0.90. The smallest MLP exhibited worse performance (MAE 0.0068, coverage 0.79) but still outperformed the GLM, which had MAE 0.0085 and very poor coverage of 0.35. Credible interval widths were reasonable for all GLM and MLP models (0.014 for GLM; 0.20–0.22 for MLPs), whereas the predictor-agnostic model had higher MAE (0.011) but acceptable coverage (0.91) and the widest interval (0.041). Finally, Figure 3c shows results for scenario FBD-4, characterized by a nonlinearly decreasing speciation rate and a spiking extinction rate. In this nonlinear setting, the MLP models clearly outperformed the other approaches in terms of MAE, with values of 0.015, 0.020, and 0.025 for models of increasing size. The predictor-agnostic model achieved a high MAE of 0.079, which was still lower than the GLM (0.11). However, coverage was low for all MLP models, falling below 0.90 (0.79 for MLP-3/2, 0.80 for MLP-16/8, and 0.85 for MLP-32/16), primarily due to the contribution of the extinction rate estimation (coverage 0.73, 0.73, and 0.78, respectively). In contrast, the GLM had very poor coverage (0.28), whereas the predictor-agnostic achieved the highest coverage (0.96) but at the cost of substantial uncertainty, with an average CI width of 0.40 compared to $\leq 0.16$ for all other models.

Table 4 reports the mean effective sample size (ESS) per hour obtained for each inference framework across all simulated scenarios. This metric summarizes the sampling efficiency of each model by combining its effective sample size across parameters with the corresponding computational time. Across all scenarios, the predictor-agnostic (PA) model consistently achieved the highest ESS per hour, followed by the GLM. For example, in the Epi scenarios, ESS/hour for PA ranged from 162.19 in Epi-4 to 58,356.51 in Epi-2, consistently exceeding the GLM, which itself remained substantially more efficient than any of the MLP architectures. The MLP models generally showed significantly lower ESS/hour. In most scenarios, smaller MLPs converged faster than larger ones—for instance, in Epi-1, MLP-3/2 reached 12,750.74 compared with 6,630.82 and 4,586.71 for MLP-16/8 and MLP-32/16, respectively. However, FBD-4 was an exception: in this case, convergence improved with increasing MLP size, with the largest model (MLP-32/16) achieving the highest ESS/hour (11.32) among the MLPs. Overall, the pattern shows that PA is the most efficient sampler, followed by the GLM, while MLP architectures are generally slower, with smaller networks typically converging faster except in specific scenarios such as FBD-4.

## 3.2 Interpretation of Bayesian MLPs

Figure 4 presents the results of the interpretability analyses conducted on the Bayesian MLP models for two distinct inference tasks from the simulation study: estimating migration rates ($M$) in scenario Epi-4, and estimating speciation ($\lambda$) and extinction ($\mu$) rates in scenario FBD-4. The analyses employed partial dependence plots (PDPs) and Shapley values (SHAP).

In scenario Epi-4, the task involved inferring 20 migration rates among five population pairs, using two continuous predictors as MLP inputs. The first predictor was relevant, exhibiting a sigmoidal relationship with the target migration rate, whereas the second was irrelevant, randomly drawn from a uniform distribution. The PDP for the relevant predictor clearly captured the expected sigmoidal trend, while the PDP for the random predictor was essentially flat, indicating no meaningful influence on the model output (Figure 4a). This pattern was corroborated by the SHAP analysis (Figure 4b), where the relevant predictor displayed strong feature attribution, and the random predictor contributed negligibly to the predictions.

Scenario FBD-4 focused on estimating speciation ($\lambda$) and extinction ($\mu$) rates for four classes of species, each defined by combinations of two binary traits. The predictors included time, the two binary traits (one relevant and one irrelevant), and an additional continuous predictor modeled as a random walk with standard normal increments across time bins. For speciation rate predictions,

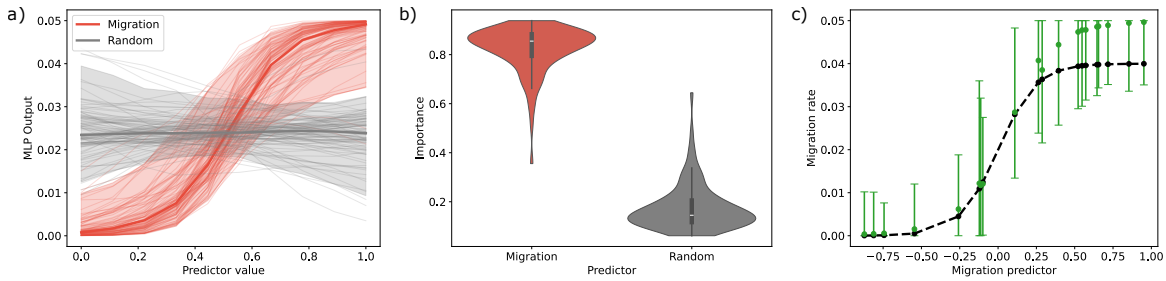


Figure 4: Results of the MLP-16/8 model for scenario Epi-4, with two continuous predictors—time (relevant, red) and a random predictor (irrelevant, gray). (a) Distributions of partial dependence plots (PDPs) across replicate trees for both predictors. (b) Distributions of SHAP feature importances across trees. (c) Reconstruction of the relationship between the relevant predictor and the inferred migration rates, showing the median estimates across trees with corresponding 95% credible intervals.

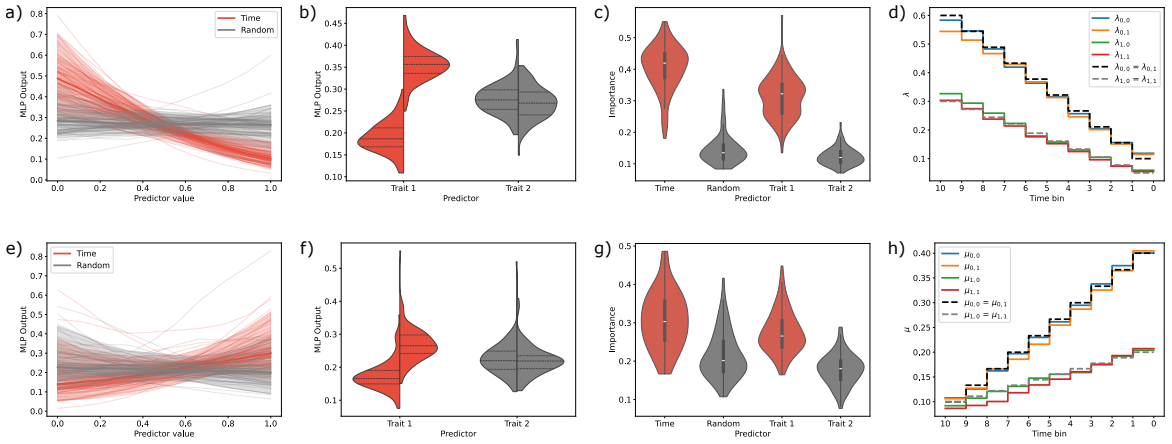


Figure 5: Results of the MLP-32/16 model for scenario FBD-4. The analysis includes four predictors: time (relevant, shown in red), a random continuous predictor (irrelevant, gray), and two binary traits: trait 1 (relevant), and trait 2 (irrelevant). The first column shows the distributions of partial dependence plots (PDPs) across replicate trees for the two continuous predictors (time and random). The second column reports PDPs for the two binary traits, where for each violin plot, the left and right halves correspond to trait values of 0 and 1, respectively. The third column presents the distributions of SHAP feature importances across trees. The fourth column displays the median predicted rates through time. The top row refers to the speciation rate ($\lambda$), and the bottom row to the extinction rate ($\mu$).

the PDPs with respect to time and the random walk predictor (Figure 5a) and with respect to the binary traits (Figure 5b) revealed that the MLP effectively captured the decreasing temporal trend and the influence of the relevant trait, while remaining largely insensitive to the irrelevant and random predictors. This interpretation is supported by the SHAP analysis (Figure 5c), which identified time and the relevant trait as the most influential features, with the others showing minimal contribution. The median predicted speciation rates across all species classes (Figure 5d) further demonstrate that the MLP accurately recovered the distinct evolutionary dynamics associated with each class.

For the extinction rate ($\mu$), analogous patterns were observed (Figure 5e-h). The PDPs for the relevant trait and time predictor exhibited clear trends, while those for the irrelevant and random predictors remained flat (Figure 5e-f). The SHAP values (Figure 5g) again highlighted time and the relevant trait as the primary drivers of the model's predictions, albeit with slightly less pronounced feature attribution than for speciation. Despite this reduced clarity, the MLP successfully produced accurate extinction rate estimates across all populations (Figure 5h).

### 3.3 Results on Empirical Data

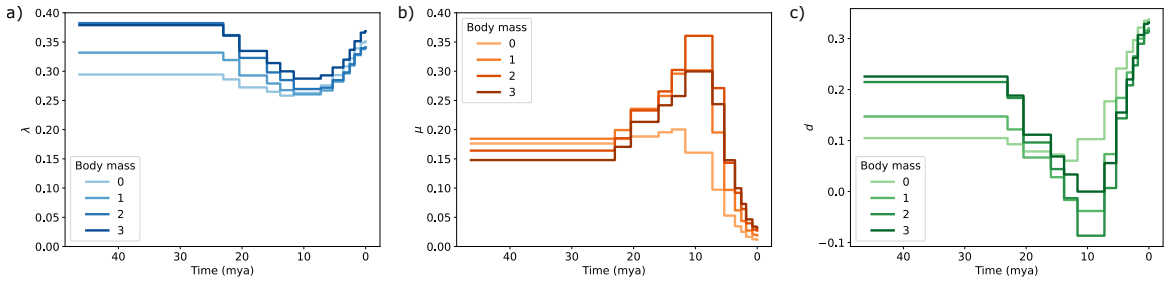


Figure 6: Results for the Platyrrhine empirical dataset. The panels display, from left to right, the median estimates across all replicate trees for (a) the speciation rate ($\lambda$); (b) the extinction rate ($\mu$); and (c) the diversification rate ($d = \lambda - \mu$).

Figure 6 summarizes the results obtained by applying our Bayesian MLP model to the empirical dataset of Platyrrhine primates. The analysis aimed to estimate the speciation rate ($\lambda$), extinction rate ($\mu$), and diversification rate ($d = \lambda - \mu$) over time, across groups of species classified by a four-level trait representing body mass (0: small, 1: medium-small, 2: medium-large, 3: large). Both time and body mass were included as predictors of the evolutionary rates.

For a representative replicate tree, the estimated body masses for each branch are shown in Figure 7, while Figure 8 displays the median branch-specific estimates of the evolutionary rates, and Figure 9 presents the marginal estimates with their associated uncertainty intervals.

Overall, the model captures a mixed influence of time and body mass on the evolutionary dynamics. The speciation rate ($\lambda$) remains relatively constant across both predictors, showing only minor temporal fluctuations and limited dependence on body mass. In contrast, the extinction rate ($\mu$) exhibits a stronger interaction between time and body mass: larger-bodied species experience a pronounced spike in extinction around 10 million years ago, followed by a rapid decline toward the present, whereas smaller-bodied species show a more gradual and later decrease. Consequently, the diversification rate ($d$) also reflects this interaction—larger-bodied species undergo a temporary reduction in diversification around 10 million years ago, while smaller-bodied species maintain a relatively stable rate. After this period, all body mass groups exhibit an upward trend in diversification approaching the present.

## 4 Discussion

In this study, we have introduced a novel Bayesian neural network (BNN) framework for phylodynamic inference that effectively integrates non-genetic covariates to model complex evolutionary dynamics. Our model uses multilayer perceptrons (MLPs) within a Markov chain Monte Carlo (MCMC) setting to unsupervisedly learn functional mappings from predictors to key parameters such as migration, speciation, and extinction rates, overcoming the limitations of traditional generalized linear models

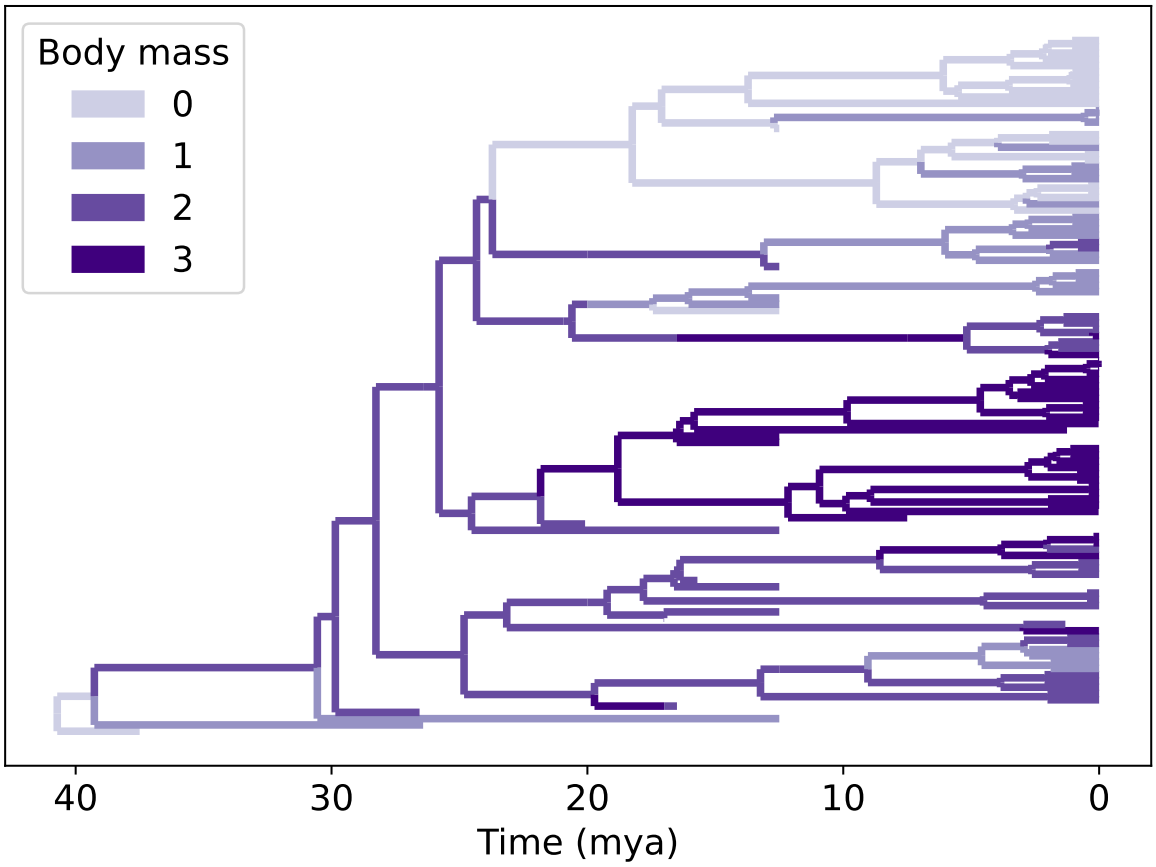


Figure 7: Estimated body masses for each branch in a representative Platyrrhine tree, computed as the median estimated body mass per branch.

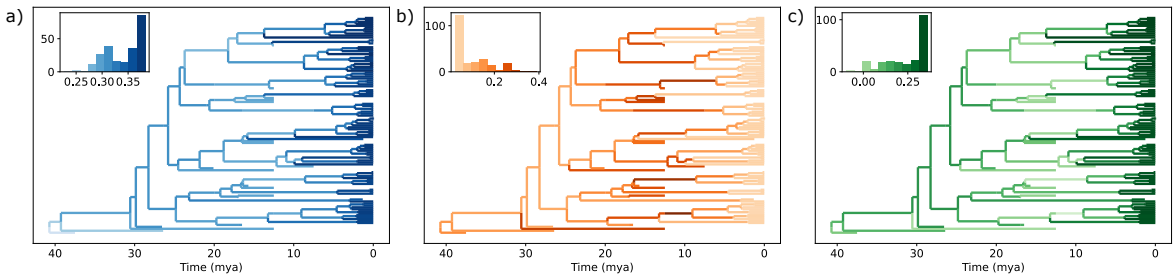


Figure 8: Median estimates for each branch for a representative Platyrrhine tree, showing (a) the speciation rate ($\lambda$); (b) the extinction rate ($\mu$); and (c) the diversification rate ($d = \lambda - \mu$). Each panel includes an inset histogram illustrating the distribution of the corresponding rate across the tree.

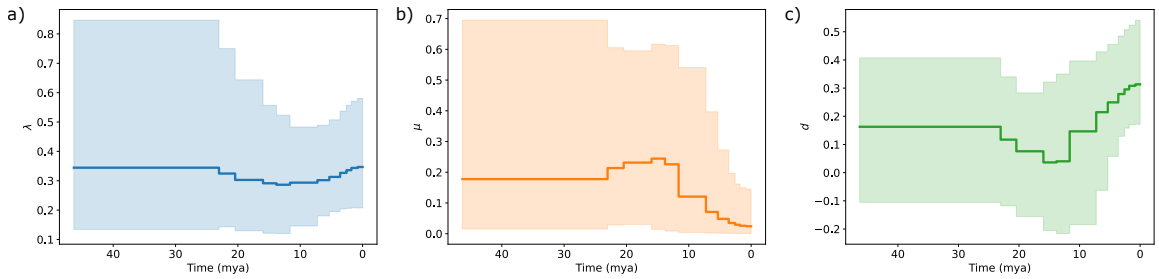


Figure 9: Marginal estimates and 95% credible intervals for (a) the speciation rate ($\lambda$); (b) the extinction rate ($\mu$); and (c) the diversification rate ($d = \lambda - \mu$) for a representative Platyrrhine tree.

(GLMs) that are restricted to linear and additive effects. We benchmarked our approach through extensive simulations, including four epidemiological and four macroevolutionary scenarios, with a diverse range of predictor-parameter relationships, evaluating the performance of three different MLP architectures against predictor-agnostic and GLM models in terms of estimation accuracy and precision. We analysed our results by grouping scenarios that shared similar dynamics (constant, linear, nonlinear) and found that our BNN framework consistently matches or exceeds the performance of state-of-the-art phylodynamic models across all scenarios.

In scenarios characterized by constant predictor-parameter relationships (Epi-1 and FBD-1), our smallest MLP architecture outperformed all other models, achieving the lowest mean absolute error (MAE), highest coverage, and narrowest credible intervals. Its performance even exceeded that of the GLM—despite the latter being well suited to linear settings and the MLP having a larger parameter space. Larger MLP architectures performed comparably well, including the largest network tested, demonstrating a notable robustness to overfitting in simple settings, even when the network size was suboptimal. A slight decrease in precision was observed as model complexity increased. In contrast, the predictor-agnostic model performed substantially worse in terms of MAE and credible interval width, while maintaining acceptable coverage across all scenarios. This pattern was consistently observed across all simulation studies and is largely attributable to the unconstrained parameterization of the predictor-agnostic model, which hinders accurate recovery of constant underlying processes and produces unstable estimates with systematically inflated errors. The resulting uncertainty was particularly pronounced in early time bins, where limited data make it difficult for the model to infer stable parameters, though this issue diminishes as phylogenetic trees accumulate more tips and branching events over time. When tested on linear dynamics (Epi-2, FBD-2, FBD-4), we observed similar trends: the MLPs performed on par with the GLM in Epi-2, and even outperformed it in FBD-2 and FBD-4, with the intermediate and larger architectures achieving the best results. These findings, together with those from the constant scenarios, are particularly noteworthy given that GLMs are specifically designed to model linear relationships. Yet, our Bayesian neural network (BNN) framework demonstrated comparable or superior performance, while flexibly adapting to the data without prior assumptions about functional form.

Under nonlinear dynamics in scenarios Epi-3 and Epi-4, the intermediate and larger MLP architectures clearly outperformed the GLM in terms of MAE and coverage, as the GLM was unable to capture nonlinear relationships due to its inherent linearity assumptions. The smallest MLP architecture showed somewhat reduced performance relative to its larger counterparts, likely due to its limited capacity to model complex nonlinear functions, though it still surpassed the GLM. Both intermediate and larger MLP models also outperformed the predictor-agnostic model, producing more precise estimates with narrower credible intervals and lower MAE, particularly in Epi-4. The final nonlinear scenario, FBD-3, presented a more challenging case, involving a nonlinearly decreasing speciation rate coupled with a spiking extinction rate. As a continuous model, the Bayesian MLP is not ideally suited to capturing such sharp, transient events: while it successfully identified the overall trend, it tended to smooth the spike, leading to reduced coverage. Nonetheless, increasing the model size improved performance, and all MLP architectures outperformed both the GLM and the predictor-agnostic models in terms of MAE. Overall, these results illustrate that Bayesian MLPs provide accurate and credible reconstructions of nonlinear dynamics, though they are less effective at modeling abrupt, spiking

events. In such cases, high coverage from more flexible models like BDSky may reflect uncertainty rather than precise inference, whereas the MLP framework offers a more balanced trade-off between predictive accuracy and credible uncertainty.

When selecting an MLP architecture, the complexity of the underlying parameter-predictor relationships should be taken into account. In our experiments, smaller architectures generally performed better in scenarios with simpler, more constant relationships, while larger networks were more effective at capturing nonlinear, continuously changing, or abruptly changing dynamics. These differences in predictive performance, however, were not decisive. The number of parameters also affects computational speed: MLP models are typically slower than the predictor-agnostic and GLM frameworks, and larger networks require more time to converge. Nonetheless, all tested MLP configurations were computationally feasible, and runtime differences were not prohibitive. Depending on the scenario, smaller MLPs tend to converge faster in simple settings, whereas larger architectures may be preferable when multiple predictors or more complex dynamics are involved. This underscores the importance of balancing model complexity, predictive performance, and computational efficiency when designing Bayesian neural networks.

Interpreting complex models is critical in evolutionary and phylodynamic inference, as understanding which predictors drive parameter estimates provides both biological insight and confidence in model outputs. Our analyses showed that the Bayesian MLP models accurately recovered relationships between relevant predictors and target parameters while remaining largely insensitive to irrelevant or random inputs. For instance, in scenario Epi-4, the model captured the expected sigmoidal relationship for migration rates, and in FBD-4, it effectively identified temporal trends and the effects of relevant traits on speciation and extinction rates. Importantly, we demonstrated that these models are interpretable using partial dependence plots (PDPs) and Shapley values (SHAP), which allow quantification of each predictor's influence on model outputs. The ability to distinguish relevant from irrelevant features is particularly valuable in high-dimensional or noisy settings, where spurious predictors could otherwise bias inference. By correctly attributing effects to biologically meaningful variables while remaining largely insensitive to random or irrelevant inputs, Bayesian MLPs combine predictive accuracy with transparency. This capability supports robust inference in complex scenarios, including nonlinear dynamics or multi-class evolutionary processes, where traditional parametric models may struggle to capture subtle or interacting effects. Overall, these findings highlight that Bayesian MLPs provide a flexible and interpretable framework, capable of integrating complex predictor information while maintaining clarity about which factors truly drive predictions.

Beyond phylodynamic applications, the proposed Bayesian MLP framework holds promise for a wide range of evolutionary inference tasks. The same model structure could be applied to non-phylodynamic settings such as estimating phylogenetic substitution rates, molecular clock parameters, or other evolutionary processes where parameter dynamics depend on external predictors. Its flexibility in capturing both linear and nonlinear relationships without strong prior assumptions makes it a compelling alternative to traditional parametric models, potentially enabling more accurate and data-driven reconstructions of complex evolutionary patterns.

# References


[1] Tanja Stadler, Denise Kühnert, David A Rasmussen, and Louis du Plessis. Insights into the early epidemic spread of ebola in sierra leone provided by viral sequence data. *PLoS currents*, 6:ecurrents–outbreaks, 2014.

[2] Tetyana I Vasylyeva, Louis Du Plessis, Andrea C Pineda-Peña, Denise Kühnert, Philippe Lemey, Anne-Mieke Vandamme, Perpétua Gomes, Ricardo J Camacho, Oliver G Pybus, Ana B Abecasis, et al. Tracing the impact of public health interventions on hiv-1 transmission in portugal using molecular epidemiology. *The Journal of Infectious Diseases*, 220(2):233–243, 2019.

[3] Marta Giovanetti, Nuno Rodrigues Faria, José Lourenço, Jaqueline Goes de Jesus, Joilson Xavier, Ingra Morales Claro, Moritz UG Kraemer, Vagner Fonseca, Simon Dellicour, Julien Thézé, et al. Genomic and epidemiological surveillance of zika virus in the amazon region. *Cell Reports*, 30(7):2275–2283, 2020.

[4] Torsten Seemann, Courtney R Lane, Norelle L Sherry, Sebastian Duchene, Anders Gonçalves da Silva, Leon Caly, Michelle Sait, Susan A Ballard, Kristy Horan, Mark B Schultz, et al. Tracking the covid-19 pandemic in australia using genomics. *Nature communications*, 11(1):4376, 2020.

[5] Hélène Morlon. Phylogenetic approaches for studying diversification. *Ecology letters*, 17(4):508–525, 2014.

[6] Tanja Stadler. How can we improve accuracy of macroevolutionary rate estimates? *Systematic biology*, 62(2):321–329, 2013.

[7] Sean Nee, Robert Mccredie May, and Paul H Harvey. The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344(1309):305–311, 1994.

[8] Tanja Stadler, Denise Kühnert, Sebastian Bonhoeffer, and Alexei J Drummond. Birth–death skyline plot reveals temporal changes of epidemic spread in hiv and hepatitis c virus (hcv). *Proceedings of the National Academy of Sciences*, 110(1):228–233, 2013.

[9] Daniel L Rabosky, Francesco Santini, Jonathan Eastman, Stephen A Smith, Brian Sidlauskas, Jonathan Chang, and Michael E Alfaro. Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. *Nature communications*, 4(1):1958, 2013.

[10] Daniel L Rabosky, Stephen C Donnellan, Michael Grundler, and Irby J Lovette. Analysis and visualization of complex macroevolutionary dynamics: an example from australian scincid lizards. *Systematic biology*, 63(4):610–627, 2014.

[11] Denise Kühnert, Tanja Stadler, Timothy G Vaughan, and Alexei J Drummond. Phylodynamics with migration: a computational framework to quantify population structure from genomic data. *Molecular biology and evolution*, 33(8):2102–2116, 2016.

[12] Tanja Stadler and Sebastian Bonhoeffer. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1614):20120198, 2013.

[13] Wayne P Maddison, Peter E Midford, and Sarah P Otto. Estimating a binary character's effect on speciation and extinction. *Systematic biology*, 56(5):701–710, 2007.

[14] Jeremy M Beaulieu and Brian C O'Meara. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. *Systematic biology*, 65(4):583–601, 2016.

[15] Philippe Lemey, Andrew Rambaut, Trevor Bedford, Nuno Faria, Filip Bielejec, Guy Baele, Colin A Russell, Derek J Smith, Oliver G Pybus, Dirk Brockmann, et al. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza h3n2. *PLoS pathogens*, 10(2):e1003932, 2014.

[16] Nicola F Müller, Gytis Dudas, and Tanja Stadler. Inferring time-dependent migration and coalescence patterns from genetic sequence and predictor data in structured populations. *Virus evolution*, 5(2):vez030, 2019.

[17] Cecilia Valenzuela Agüí. A comprehensive study of the phylodynamics of sars-cov-2 in europe. Master's thesis, ETH Zurich, 2021.

[18] Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

[19] Torsten Hauffe, Juan L Cantalapiedra, and Daniele Silvestro. Trait-mediated speciation and human-driven extinctions in proboscideans revealed by unsupervised bayesian neural networks. *Science Advances*, 10(30):eadl2643, 2024.

[20] Siddhartha Chib. Chapter 57 - markov chain monte carlo methods: Computation and inference. In James J. Heckman and Edward Leamer, editors, *Handbook of Econometrics, Volume 5*, volume 5 of *Handbook of Econometrics*, pages 3569–3649. Elsevier, 2001.

[21] Tracy A. Heath, John P. Huelsenbeck, and Tanja Stadler. The fossilized birth–death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences*, 111(29):E2957–E2966, 2014.

[22] Tim Vaughan and Tanja Stadler. Bayesian phylodynamic inference of multi-type population trajectories using genomic data. *Molecular Biology and Evolution*, 42:msaf130, 2025. (BDMM-Prime, trajectory inference).

[23] Remco Bouckaert, Timothy G. Vaughan, Joèlle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, Graham Jones, Denise Kühnert, Nicola De Maio, Michael Matschiner, Fábio K. Mendes, Nicola F. Müller, Huw A. Ogilvie, Louis Du Plessis, Alex Popinga, Andrew Rambaut, David Rasmussen, Igor Siveroni, Marc A. Suchard, Chieh Hsi Wu, Dong Xie, Chi Zhang, Tanja Stadler, and Alexei J. Drummond. Beast 2.5: An advanced software platform for bayesian evolutionary analysis. *PLoS Computational Biology*, 15(4), 2019.

[24] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[25] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232, 2001.

[26] Luke Merrick and Ankur Taly. The explanation game: Explaining machine learning models using shapley values. In Andreas Holzinger, Peter Kieseberg, A Min Tjoa, and Edgar Weippl, editors, *Machine Learning and Knowledge Extraction*, pages 17–38, Cham, 2020. Springer International Publishing.