Version dated: October 30, 2023

Scalable gradients enable Hamiltonian Monte Carlo sampling for phylodynamic inference under episodic birth-death-sampling models

Yucai Shao[1], Andrew F. Magee[2], Tetyana I. Vasylyeva[3,5] and Marc A. Suchard[1,2,4]

[1]*Department of Biostatistics, Jonathan and Karin Fielding School of Public Health, University of California, Los Angeles, United States*

[2]*Department of Biomathematics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, United States*

[3]*Department of Medicine, University of California San Diego, La Jolla, United States*

[4]*Department of Human Genetics, David Geffen School of Medicine at UCLA, Universtiy of California, Los Angeles, United States*

[5] *Department of Population Health and Disease Prevention, University of California Irvine, Irvine, United States*

Corresponding author: Marc A. Suchard, Departments of Biostatistics, Biomathematics, and Human Genetics, University of California, Los Angeles, 695 Charles E. Young Dr., South, Los Angeles, CA 90095-7088, USA; E-mail: msuchard@ucla.edu

Abstract Birth-death models play a key role in phylodynamic analysis for their interpretation in terms of key epidemiological parameters. In particular, models with piecewise-constant rates varying at different epochs in time, to which we refer as episodic birth-death-sampling (EBDS) models, are valuable for their reflection of changing transmission dynamics over time. A challenge, however, that persists with current time-varying model inference procedures is their lack of computational efficiency. This limitation hinders the full utilization of these models in large-scale phylodynamic analyses, especially when dealing with high-dimensional parameter vectors that exhibit strong correlations. We present here a linear-time algorithm to compute the gradient of the birth-death model sampling density with respect to all time-varying parameters, and we implement this algorithm within a gradient-based Hamiltonian Monte Carlo (HMC) sampler to alleviate the computational burden of conducting inference under a wide variety of structures of, as well as priors for, EBDS processes. We assess this approach using three different real world data examples, including the HIV epidemic in Odesa, Ukraine, seasonal influenza A/H3N2 virus dynamics in New York state, America, and Ebola outbreak in West Africa. HMC sampling exhibits a substantial efficiency boost, delivering a 10- to 200-fold increase in minimum effective sample size per unit-time, in comparison to a Metropolis-Hastings-based approach. Additionally, we show the robustness of our implementation in both allowing for flexible prior choices and in modeling the transmission dynamics of various pathogens by accurately capturing the changing trend of viral effective reproductive number.

1 Introduction

Phylodynamic models constitute a sophisticated toolset employed to decipher the complex interplay between epidemiological and evolutionary processes, providing valuable insights into population dynamics (Lau et al. 2019). In this paper, our primary emphasis is directed toward the inference of epidemiological dynamics, rather than estimation of the underlying phylogeny through sequence analysis. Specifically, we start with a sample of molecular sequences, which can be used to reconstruct the evolutionary relationships between organisms, often viral pathogens, and yield inference on dynamics of the larger pathogen population over time while relegating the phylogeny the status of a nuisance parameter. To provide this link, a vital component of phylodynamic analysis is the use of birth-death models, which belong to an important subclass of continuous-time Markov chains (CTMCs). We use birth-death models to define the probability distribution on time-calibrated phylogenies for reflecting the fluctuations of the population size (MacPherson et al. 2022). In this context, birth-death models posit three major types of events: birth, which refers to the creation of new lineages through pathogen transmission between hosts; death, which represents host death/recovery or other removal from the studied population, and sampling, which means the collection of a sequence derived from the pathogen in a single infected host and included in the data set under analysis (Crawford 2012).

The past few decades have delivered a wide range of birth-death models. These span from a simple, constant-over-time formulation (Yang & Rannala 1997) to models that allow both birth and death rates to vary over time (Stadler et al. 2013, Höhna 2014). Further extensions incorporate additional processes, both statistical and biological, such as the collection of samples in continuous time (Stadler 2010), migration (Barido-Sottani et al. 2020), or the dependency of rates of birth and death on key biological traits (Maddison et al. 2007, FitzJohn 2010, 2012). One powerful variant, the episodic birth-death-sampling (EBDS) model (Lambert & Stadler 2013, Stadler et al. 2013, Gavryushkina et al. 2014, Du Plessis 2016) permits birth, death, and sampling rates to change in discrete epochs throughout time

to capture more complicated population dynamics. Recent inference based on EBDS models has found its way already into many applications, especially on the understanding of the spread of infectious disease (Novitsky et al. 2015, Vasylyeva et al. 2020, Minosse et al. 2021).

With increasingly rich and complex molecular sequence datasets across fields, improving the scalability of inference under EBDS models remains challenging both in terms of the number of sequences and the number of epochs. The most commonly employed inference methods based on Markov chain Monte Carlo (MCMC) (Hastings 1970, Morlon et al. 2011) use random-walk transition kernels generally to propose new parameter values in a blind fashion. Consequently, they lead to many birth-death model likelihood evaluations and slow exploration across the state space, especially for high-dimensional problems. The potentially complex correlation structure between epoch parameters can further exacerbate inference. This is where gradient-based sampling methods, such as Hamiltonian Monte Carlo (HMC) (Duane et al. 1987, Neal et al. 2011), are expected to shine. HMC has recently become very popular as a MCMC algorithm that overcomes many of the limitations of random-walk Metropolis-Hasting (MH) methods. Instead of making random proposals, HMC exploits the gradient of the log posterior with respect to (wrt) its model parameters to propose new states that are likely to be accepted and are far from the current state. Since HMC can make large moves in the state space while still maintaining a high acceptance rate, it can lead to faster convergence and better mixing than MH approaches, if one can efficiently evaluate not only the log posterior (up to a constant) but also its gradient. Successful implementation of HMC transition kernels has proved fruitful in terms of boosting sampling performance in other phylogenetic inference frameworks, including for different clock models (which describe how rates of molecular evolution vary among different organisms over time, Ji et al. 2020, Fisher et al. 2021), divergence times (the internal-node heights of phylogenies, Ji et al. 2021) and non-parametric coalescent models (which fall into another category of phylodynamic models assuming effective population size as a piecewise-constant form of time, Baele et al. 2020).

In this paper, we incorporate gradient-based sampling methods into phylodynamic analysis based on EBDS models, thereby enabling scalable inference within this framework. First, we refactor the EBDS (log) likelihood to show explicitly that the computational complexity scales linearly both in terms of the number of sequences and the number of epochs. With this refactoring in hand, we deliver a novel linear-time algorithm to evaluate the gradient of this (log-)likelihood wrt all epoch parameters simultaneously. Then we design and deploy an efficient HMC sampler that enables us to fit a large class of EBDS models in a Bayesian framework and provide an open-source implementation in the popular Bayesian Evolutionary Analysis by Sampling Trees (BEAST) software (Suchard et al. 2018).

Current approaches to Bayesian inference for EBDS epoch parameters employ a variety of prior assumptions to model the dependence structure between parameters across epochs. Some priors assume that birth, death and sampling rates across epochs are independent and identically distributed (iid) (Stadler et al. 2013, Gavryushkina et al. 2014, Vasylyeva et al. 2020). To smooth rate variation over time, temporally-auto-correlated priors such as Ornstein-Uhlenbeck smoothing prior (Du Plessis 2016), Gaussian Markov random fields (GMRF) priors (Condamine et al. 2018, Silvestro et al. 2019) and the horseshoe Markov random field for EBDS models (Magee et al. 2020) have been considered. Conveniently, both our linear-time gradients and our HMC approach generalize across all of these choices of prior without the need to construct model-specific sampling techniques and allow us to introduce the Bayesian bridge shrinkage prior to yield parsimonious time-varying rate patterns.

Across three real-world infectious disease examples that vary in the number of sequences, model dimension, and prior specification, we demonstrate the performance gain achieved by our implementation of an HMC transition kernel compared to random walk transition kernels. Moreover, for each of these datasets we infer key epidemiological parameters and demonstrate the utility of our scalable approach for providing reasonable estimates of pathogen transmission dynamics over time.

2 Methods

2.1 Setup

In an infectious disease setting, suppose an infected individual initiates an epidemic at time (measured backwards from the present day) $t_{or} > 0$, called the time of the origin. Then, each currently and newly infected individual disseminates the pathogen to others at a time-varying birth rate $\lambda(t)$ and transitions into a noninfectious state at a time-varying death rate $\mu(t)$. At any given time, we may sample an infected individual with time-varying sampling rate $\psi(t)$, at which point we add the time of sampling and a molecular sequence of their infectious agent into our time-stamped molecular sequence alignment $\mathbf{Y}$. Further, we may posit K fixed time-points at which we randomly sample all infected individuals with associated vector of probabilities $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_K)$, adding the time and molecular sequence to $\mathbf{Y}$. Note that this means that several individuals can be sampled at the same time point. The choice of the time-points is dependent on the dataset at hand and will be discussed later in this section. Every sampled infected individual may be treated and then become noninfectious with time-varying probability $r(t)$ which we assume equal to one everywhere for complete sampling.

The model defined above provides a forward in time portrayal of the epidemiological process. Considering the N sampled and time-stamped sequences in $\mathbf{Y}$ as tree tips, there exists a (possibly unknown) phylogeny $\mathcal{T}$ that depicts the evolutionary relationships among these sequences. Specifically, $\mathcal{T}$ is a rooted, bifurcating tree with N tip nodes that correspond to the sampled sequences or their hosts from the population and $N-1$ internal nodes that represent transmission events between hosts. We define the height of the nodes as the length of time between the time of the corresponding transmission/sampling events and the time of the most recent sampled sequence, which we refer to the present time, 0. Each node of $\mathcal{T}$ is then associated with a node-height ≥ 0 relative to the present, such that the difference between the parent node-height and its child node-height is a branch length measured in

the units of real time (e.g., years). We call the earliest internal node in $\mathcal{T}$ the root and its node-height corresponds to the time of the most recent common ancestor (TMRCA). Therefore, we can further define the node heights of internal nodes to be bifurcation times and that of leave nodes to be sampling times. Accordingly, for a vector of bifurcation times, we have $\boldsymbol{v} = (v_1, v_2, \ldots, v_{N-1})$ where $v_1 < \cdots < v_{N-1}$. And we let $\boldsymbol{u} = (u_1, u_2, \ldots, u_N)$ be a vector of sampling times where $u_1 < \cdots < u_N$.

For an episodic model, we make the assumption that all the rate parameters are piecewise constant across K different epochs with cut points $\boldsymbol{t} = (t_0, \ldots, t_K)$, with $t_0 = 0 < t_1 < \cdots < t_{K-1} < t_K$. We also require $t_{or} \leq t_K$. Under this assumption, we can rewrite the time dependent birth rate $\lambda(t)$ in terms of some unknown epoch-specific birth rate $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$, where $\lambda(t) = \lambda_k$ for $t_{k-1} < t \leq t_k$. Similar parametrization applies to other parameters, so that we can express $\mu(t)$ in terms of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, $\psi(t)$ in terms of $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_K)$ and $r(t)$ in terms of $\boldsymbol{r} = (r_1, \ldots, r_K)$. Without loss of generality, we let intensive sampling events happen at every time points in $\boldsymbol{t}$. Then we define $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_K)$, where $\rho(t) = \rho_k$ for $t = t_{k-1}$. We can remove these intensive sampling events at the epoch switching times from our model simply by setting $\boldsymbol{\rho} = \mathbf{0}$.

After reparametrizing the rates of the EBDS model, we can arrive at some key epidemiological quantities. For example, if we assume there are no intensive sampling events, we can specify the effective reproductive number as $R_e(t) = \frac{\lambda(t)}{\mu(t)+\psi(t)r(t)}$. Other parameters that are important include the total rate of becoming noninfectious, which is defined as $\delta(t) = \mu(t) + \psi(t)r(t)$, and the sampling proportion, defined as $\zeta(t) = \frac{\psi(t)r(t)}{\mu(t)+\psi(t)r(t)}$. If we also assume removal of lineages upon sampling, these formulas can be further simplified by letting $r(t)$ be constant and always equal to 1.

2.2 Probability Density of a Sampled Phylogeny

Recall we break time into intervals with cut points $\boldsymbol{t} = (t_0, \ldots, t_K)$ defined by epochs. Within each epoch, we define a series of subintervals such that a new subinterval start at

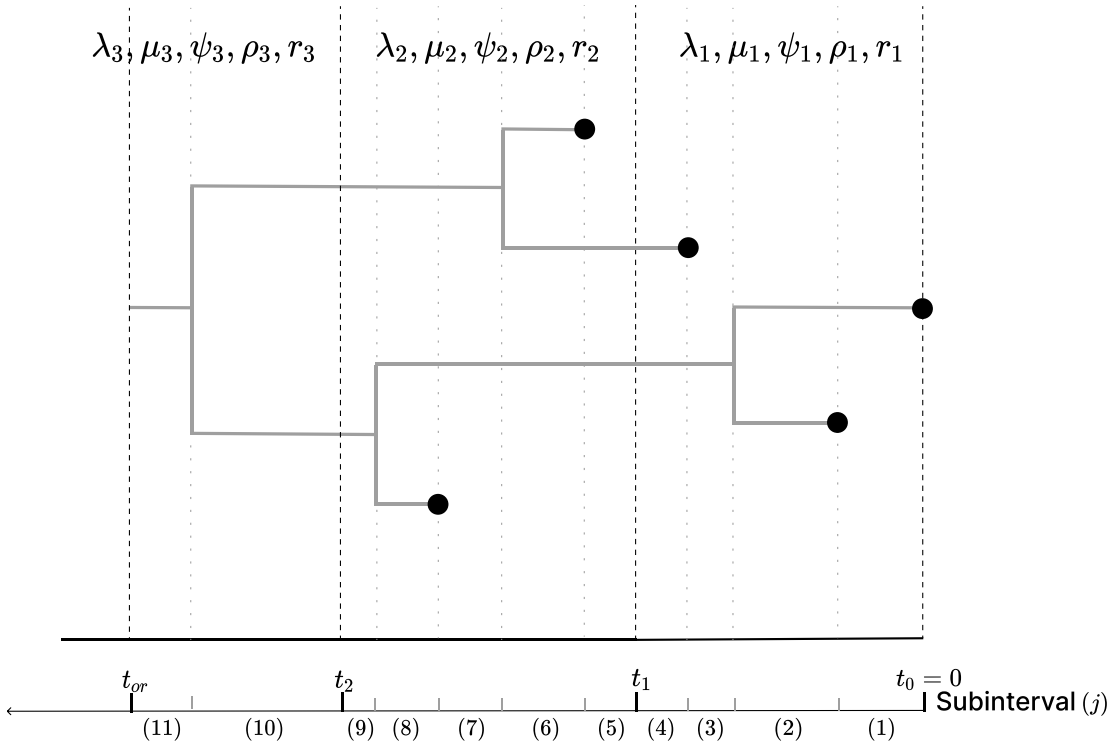

Figure 1: A phylogeny arising from an EBDS model. This sampled phylogeny has three epochs (with epoch switching time t_1, t_2) and thus three sets of model parameters including rates and probabilities. For every epoch, each branch is further divided into subinterval that starts at s_j and ends at time s_{j+1} so that no epoch switching, birth or sampling event occurs within it. Each subinterval within each epoch k is represented by a phylogeny segment index, j.

every bifurcation time $\boldsymbol{v}$, sampling time $\boldsymbol{u}$ and epoch switching time $\boldsymbol{t}$. We delineate the subinterval by indices j, which begins at s_j and terminates at s_{j+1} (where $s_j < s_{j+1}$). If $t_{or} = t_K$, then the grids $\boldsymbol{s} = (s_1, \ldots, s_{2N-2+K})$ can be obtained by joining the time points in $\boldsymbol{v}$, $\boldsymbol{u}$ and $\boldsymbol{t}$ according to their ascending order when none of these times coincide with each other. If $t_{or} < t_K$, we have $s_{2N-2+K} = t_{or}$ instead of t_K.

Consequently, each subinterval, inclusive on the left, is partitioned in such a way that it precludes the occurrence of an epoch switching, birth or sampling event within its boundaries. Within the kth epoch, the first subinterval starts at $s_j = t_{k-1}$ and the last subinterval ends at $s_{m_k+1} = t_k$. (Note for the last epoch K, the last subinterval ends at t_{or}.) We assign $L(j)$ to account for the number of lineages in $\mathcal{T}$ that are extant in subinterval time $(s_j, s_{j+1}]$.

Our likelihood derivation falls into the common framework with Stadler et al. (2013), Gavryushkina et al. (2014) and Magee & Höhna (2021). However, instead of writing the likelihood in terms of the times of node and epochs, we write it in terms of the subintervals j. This representation highlights the fact that the likelihood can be computed in one pass, starting at the present and ending at the origin. The interval-based representation of the likelihood is as follows:

$$\mathbb{P}[\mathcal{T} \mid \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\psi}, \boldsymbol{\rho}, \boldsymbol{r}, \boldsymbol{t}] = N_1 \log \rho_1 + \underbrace{\sum_{k=1}^{K} \sum_{j=1}^{m_k}}_{\sum_{k=1}^{K} m_k \leq 2N+K-2^*} \left(\log I_k(E_j) + L(j) \log \left(\frac{q_k\left(s_{j+1}\right)}{q_k\left(s_j\right)} \right) \right), \tag{1}$$

where m_k is the total number of subintervals in epoch k. (*: equality holds when no events happens at the exact same time except for the current).

The indicator function $I_k(E_j)$ is labelled by the index k. This implies that the function is concerned with events occurring within the time frame $(t_{k-1}, t_k]$. We have E_j represent the event that takes place at the termination of subinterval j within epoch k. In most phylodynamic studies, ancestral sampling scenarios are not taken into account; therefore, our model is based on the assumption of a strictly bifurcating phylogenetic tree and does not involve considerations of ancestral sampling cases, which is distinctive from the work of Gavryushkina et al. (2014). Nonetheless, incorporating ancestral sampling into our framework is relatively straightforward. This can be achieved by setting the treatment probability to be less than 1 and adding the term $\psi_k(1 - r_k)$ to our indicator function to account for events involving ancestral samples. Consequently, this indicator function takes the following

form:

$$I_k(E_j) = \begin{cases} 1, & E_j = \text{a epoch switching event happens on } s_{j+1} \\ \lambda_k, & E_j = \text{a birth event happens at } s_{j+1} \\ \psi_k((1-r_k)p_k(s_{j+1}) + r_k), & E_j = \text{a tip sampling event happens at } s_{j+1} \\ \rho_k^{N_k}((1-r_k)p_{k-1}(s_{j+1}) + r_k)^{N_k} \cdot (1-\rho_k)^{L(j)-N_k}, & E_j = \text{an intensive sampling event happens at } s_{j+1} = t_{k-1}. \end{cases} \tag{2}$$

Note that $p_k(t)$ is the probability that an infected individual at time t has no sampled descendants when the process is stopped (i.e., at time t_0), and $q_k(t)$ is the probability density of an individual at time t giving rise to an edge between t and t_{k-1} (not t_k since we define time to flow backwards which is the reverse of the generative process) for $t_{k-1} < t < t_k$ in epoch k. We have $p_0(t_0 = 0) = 1$.

The intensive sampling probability at time t_{k-1} is ρ_k and the corresponding number of leaves sampled at that time is N_k. The index here is intentionally misaligned to reconcile the fact that we model the epoch as left inclusive in time.

The definitions of the underlying functions, $q_k(t)$ and $p_k(t)$, follow the work from Stadler et al. (2013) and the detailed formulas are included in Supplementary Material S1. Note that our equation 1 does not condition the tree likelihood upon any particular properties, such as the presence of at least one sampled individual. Without loss of generality, additional conditioning schemes can be integrated by adding a factor to the log-likelihood; relevant discussions on this subject are available in Table S3 from the study by MacPherson et al. (2022).

As stated previously, our representation of the likelihood differs from the more standard nodewise representation (see for example Stadler et al. 2013, Gavryushkina et al. 2014, Wu 2014, Magee & Höhna 2021). Our representation makes it explicit that the likelihood computation can be accomplished in $\mathcal{O}(N+K)$ time (see Algorithm 1 for computational details). We demonstrate this behavior empirically in Supplementary Material S6. On the other hand, as we show in Supplementary Material S5, the conventional nodewise representation leads to ambiguities in the cost and a wide potentially range of computational complexities de-

pending on implementation decisions. In Supplementary Material S6 we show empirically that formulations based on the nodewise representation include both implementations which are of the same computational order as ours (namely BEAST2 (Bouckaert et al. 2019) and RevBayes (Höhna et al. 2016)) and which scale worse in the number of epochs (TreePar (Stadler et al. 2013)).

2.3 Inference

In a Bayesan inference procedure, as introduced in Section 2.1, we use a multiple sequence alignment with the sampling times, the time-stamped sequences, $\mathbf{Y}$, as the input data. Based on $\mathbf{Y}$, we can form the posterior distribution over the product space of trees and EBDS model parameters as follows. First, a phylogeny $\mathcal{T}$ is generated from the EBDS process defined in Section 2. Then we specify a molecular clock model that controls the rate at which evolution occurs on each branch of $\mathcal{T}$. Under a molecular character-based CTMC substitution model, the columns in the sequence alignment evolve independently along the branches of the tree. Adoption of different substitution models is contingent upon the distinct attributes of the dataset under investigation (see Section 2.6.1). For the sake of notational convenience, we refer to the vector encompassing both substitution and clock model parameters as $\boldsymbol{\omega}$. We denote by $\mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{\omega}, \mathcal{T})$ the probability of the time-stamped sequences under the CTMC substitution model, known as the phylogenetic likelihood. Subsequently, we can factorize the posterior in the following manner:

$$\begin{aligned}\mathbb{P}[\mathcal{T}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\psi}, \boldsymbol{\rho}, \boldsymbol{r}, \boldsymbol{t}, \boldsymbol{\omega} \mid \boldsymbol{Y}] &\propto \mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{\omega}, \mathcal{T})\mathbb{P}[\mathcal{T} \mid \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\psi}, \boldsymbol{\rho}, \boldsymbol{r}, \boldsymbol{t}] \\ &\quad \times \mathbb{P}[\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\psi}, \boldsymbol{\rho}, \boldsymbol{r}, \boldsymbol{t}, \boldsymbol{\omega}] \\ &\propto \mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{\omega}, \mathcal{T})\mathbb{P}(\boldsymbol{\omega})\mathbb{P}[\mathcal{T} \mid \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\psi}, \boldsymbol{\rho}, \boldsymbol{r}, \boldsymbol{t}] \\ &\quad \times \mathbb{P}(\boldsymbol{\lambda})\mathbb{P}(\boldsymbol{\mu})\mathbb{P}(\boldsymbol{\psi})\mathbb{P}(\boldsymbol{\rho})\mathbb{P}(\boldsymbol{r})\mathbb{P}(\boldsymbol{t}).\end{aligned} \tag{3}$$

In phylodynamic analyses, it is sometimes advantageous to streamline the model by

maintaining the death rate as constant. We can also presume the intensive sampling probability to be 0 and treatment probability to uniformly be 1 across all epochs. In handling time-varying parameters, we choose either iid priors or Markov random field models based on dataset-dependent assumptions pertaining to the patterns of change expected in rate parameters. In this paper, we specifically consider the GMRF and the Bayesian bridge Markov random field model, the latter of which we describe below.

With increasing complexity of the existing EBDS models, we seek to integrate Bayesian regularization methods to help manage the potentially vast quantity of model parameters. Specifically, we consider Markov random field priors which specify distributions on the incremental difference between the log-transformed rate parameters. By assigning a normal distribution to the incremental changes, we arrive at the GMRF priors that induce a smoothing effect on the change of rate parameters across contiguous epochs. This approach naturally leads to adjacent epochs exhibiting similar rate values. However, a strong data signal indicative of a rate change can still manifest in the resulting trajectory. By placing a heavy-tailed Bayesian bridge prior (Piironen & Vehtari 2017) on these, we achieve a more generalized extension of the GMRF model. The key distinction resides in the specification of the standard deviation arising from the normal priors on the increments. In this resulting Bayesian bridge Markov random field framework, each epoch's increment is assigned an additional variable to account for variation, thereby affording greater flexibility to the model.

Supposing we have varied birth rates, we define the birth rate on the log scale $\lambda_k^* = \log(\lambda_k)$. Then we have the prior on increments, $\mathbb{P}\left(\lambda_k^* - \lambda_{k-1}^* \mid \tau\right) \propto \exp\left\{-\left|\frac{\lambda_k^* - \lambda_{k-1}^*}{\tau}\right|^\alpha\right\}$ for $k > 1$, where τ is the global scale parameter that controls the overall degree of parameter variation. As α diminishes, the function $\mathbb{P}\left(\lambda_k^* - \lambda_{k-1}^*\right)$ accrues an increased density close to zero. For the purpose of our study, we establish $\alpha = 0.25$ to address a potent prior assumption that $\lambda_k^* - \lambda_{k-1}^*$ is proximate to 0 without inducing any problems related to mixing issues. In other words, we do not anticipate substantial fluctuations in the birth rates across consecutive epochs (but allow for rapid rate shift, for example during the exponential growth

phase.) Another important parameter is the local scale, denoted as ν_k, which is specific to an individual increment $\lambda_k^* - \lambda_{k-1}^*$. Its density regulates the magnitude of the spike and the tail behavior of the above marginal $\lambda_k^* - \lambda_{k-1}^* \mid \tau$.

Note that the GMRF model can be perceived as a specific instance of the Bayesian bridge MRF, where all the local scale parameters are equalized to 1 and α is fixed at 2. In this case, the increment differences adhere to a normal distribution whose variance is solely governed by a single global scale parameter.

To complete our model, a normal prior is assigned to λ_1^* in adherence with the method outlined in Magee et al. (2020). We obtain the mean parameter of the prior using an empirical Bayes method. This provides a crude estimate of the log rate parameter, coupled with a standard deviation that is sufficiently large to encompass all possible values (See S3). We apply a Gamma(1,1) prior to $\phi = \tau^{-\alpha}$. This selection is grounded on a combination of theoretical considerations and empirical validation and allows for an efficient Gibbs sampler for τ.

To regularize the tail behavior, we leverage the shrunken-shoulder version of the Bayesian bridge prior and limit the bridge to have light tails past the slab width, ξ (Piironen & Vehtari 2017, Nishimura & Suchard 2023). An efficient update of Markov random field models global and local scale parameters (for Bayesian bridge priors) follows Nishimura & Suchard (2023). In this framework, the prior on the increment space represented as a scale mixture of normal distributions:

$$\mathbb{P}\left(\lambda_k^* - \lambda_{k-1}^* \mid \nu_k, \tau, \xi\right) = \mathrm{N}\left(0, \left(\frac{1}{\xi^2} + \frac{1}{\nu_k^2 \tau^2}\right)^{-1}\right), \tag{4}$$

where ν_k is called the local scale parameter and τ is the global scale parameter. (Note that ν_k has an exponentially tilted stable distribution with characteristic exponent $\alpha/2$.) This mixture representation aids in clarifying the local adaptivity of the Bayesian bridge prior as considerable changes in rates can be accommodated by an increase in ν_k without

necessitating a rise in τ. The inclusion of the slab width helps to bound the variance of increments to ξ^2. We set $\xi = 2$, which creates a reasonable upper limit on the variations in birth rate between consecutive epochs.

In our study, we primarily focus on sampling $\mathbb{P}[\mathcal{T} \mid \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\psi}, \boldsymbol{t}]$. With increasing numbers of epochs, the parameter space of the EBDS model expands quickly, exhibiting substantial correlation between adjacent epochs. To improve the sampling efficiency, we utilize HMC method to concurrently sample the time varying model parameters and ensure a high acceptance rate.

2.4 Hamiltonian Monte Carlo Sampling

Hamiltonian Monte Carlo is a widely-used Markov chain Monte Carlo method to sample from a target distribution effectively. Given a target parameter $\boldsymbol{\theta}$ with a posterior probability density $\pi(\boldsymbol{\theta})$, HMC iteratively generates samples from the target distribution by simulating the dynamics of a physical system whose equilibrium distribution is equal to $\pi(\boldsymbol{\theta})$. In particular, HMC introduces an auxiliary momentum parameter $\boldsymbol{d}$, which is typically chosen to follow a multivariate normal distribution with zero mean and covariance matrix $\boldsymbol{M}$, i.e., $\boldsymbol{d} \sim \mathcal{N}(0, \boldsymbol{M})$. $\boldsymbol{M}$ is also known as the mass matrix, which serves as a hyperparameter. The Hamiltonian function of the system is defined as:

$$H = U(\boldsymbol{\theta}) + K(\boldsymbol{d}), \tag{5}$$

where $U(\boldsymbol{\theta}) = -\log(\pi(\boldsymbol{\theta}))$ is the potential energy, and $K(\boldsymbol{d}) = \boldsymbol{d}^\top \boldsymbol{M} \boldsymbol{d}$ is the kinetic energy of the system.

Starting from the current state $(\boldsymbol{\theta}_0, \boldsymbol{d}_0)$, HMC updates the state according to the fol-

lowing differential equations:

$$\begin{aligned} \frac{\mathrm{d}\boldsymbol{d}}{\mathrm{d}t} &= -\nabla U(\boldsymbol{\theta}) = \nabla \log \pi(\boldsymbol{\theta}) \\ \frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} &= +\nabla K(\boldsymbol{d}) = \boldsymbol{M}^{-1}\boldsymbol{d}. \end{aligned} \tag{6}$$

The simple and effective "leapfrog" method (Neal et al. 2011) approximates the solution to (6) numerically:

$$\begin{aligned} \boldsymbol{d}_{t+\epsilon/2} &= \boldsymbol{d}_t + \frac{\epsilon}{2}\nabla \log \pi\left(\boldsymbol{\theta}_t\right) \\ \boldsymbol{\theta}_{t+\epsilon} &= \boldsymbol{\theta}_t + \epsilon \boldsymbol{M}^{-1}\boldsymbol{d}_{t+\epsilon/2} \\ \boldsymbol{d}_{t+\epsilon} &= \boldsymbol{d}_{t+\epsilon/2} + \frac{\epsilon}{2}\nabla \log \pi\left(\boldsymbol{\theta}_{t+\epsilon}\right), \end{aligned} \tag{7}$$

where ϵ is the size of each leapfrog step, and n steps are required to simulate the Hamiltonian dynamics from time $t = 0$ to $t = n\epsilon$. In practice, the "leapfrog" method has been shown to be stable and accurate for a wide range of step sizes (Neal et al. 2011).

The default choice of the mass matrix is the identity matrix. However, using a different $\boldsymbol{M}$, such as a log-posterior Hessian approximation can largely enhance the efficiency of HMC sampling. In this work, $\boldsymbol{M}$ is adaptively tuned to estimate the expected (diagongal) Hessian averaged over the prior distribution. This design choice alleviates some computational burden, following the work of Ji et al. (2020).

2.5 Gradient

HMC sampling of the model parameters requires the gradient of the log-likelihood derived from (1) wrt the EBDS model rate parameters. The gradient is the collection of derivatives wrt model parameters:

$$\nabla_{\boldsymbol{\theta}} \mathbb{P}[\mathcal{T} \mid \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\psi}, \boldsymbol{\rho}, \boldsymbol{r}, \boldsymbol{t}] = \left(\frac{\partial \mathbb{P}}{\partial \theta_1}, \cdots, \frac{\partial \mathbb{P}}{\partial \theta_k}, \cdots, \frac{\partial \mathbb{P}}{\partial \theta_K} \right)^{\top}, \tag{8}$$

where $\theta_k \in \{\lambda_k, \psi_k, \mu_k, \rho_k\}$ is a unified parameter to reduce notation clutter.

Given the piece-wise constant nature of the model, the likelihood assumes a consistent form across all epochs. Therefore, we can examine the gradient of the log-likelihood at each epoch separately. We denote the log-likelihood at epoch k and phylogeny segment j as:

$$\mathbb{P}_k(j) = \log I_k(E_j) + L(j) \log \left(\frac{q_k\left(s_{j+1}\right)}{q_k\left(s_j\right)} \right). \tag{9}$$

We can further get individual terms in (8) by accumulating contributions from each epoch and the corresponding phylogeny segments:

$$\frac{\partial \mathbb{P}}{\partial \theta_k} = \sum_{a=k}^{K} \sum_{j=1}^{m_k} \frac{\partial \mathbb{P}_a(j)}{\partial \theta_k}, \quad \theta_k \in \{\lambda_k, \psi_k, \mu_k, \rho_k\}. \tag{10}$$

By examining the interdependency between epochs, we discern that a given epoch k exerts influence on the gradient of parameters pertaining to that and all preceding epochs. Consequently, it becomes necessary to consider $\frac{\partial \mathbb{P}_k(j)}{\partial \theta_k}$ and $\frac{\partial \mathbb{P}_k(j)}{\partial \theta_{k-i}}$ respectively, where i is a positive integer ranging between 1 and $(k-1)$.

First, we consider the gradient contribution at epoch k wrt the current model parameters $\frac{\partial \mathbb{P}_k(j)}{\partial \theta_k}$, where $\theta_k \in \{\lambda_k, \psi_k, \mu_k, \rho_k\}$.

Then we have the following cases:

$$\frac{\partial \mathbb{P}_k(j)}{\partial \theta_k} = \begin{cases} \text{If } E_j \text{ is a birth event happens at subinterval end } s_{j+1}\text{:} \\ \quad \mathbb{1}_{\theta_k=\lambda_k} \dfrac{1}{\theta_k} + L(j) \cdot \dfrac{\partial Q_k(s_{j+1}, s_j)}{\partial \theta_k}, & (11) \\ \text{If } E_j \text{ is a serial sampling event happens at subinterval end } s_{j+1}\text{:} \\ \quad \mathbb{1}_{\theta_k=\psi_k} \dfrac{1}{\theta_k} + \dfrac{1-r_k}{(1-r_k)p_k\left(s_j\right)+r_k} \cdot \dfrac{\partial p_k(s_j)}{\partial \theta_k} + L(j) \cdot \dfrac{\partial Q_k(s_{j+1}, s_j)}{\partial \theta_k}, & (12) \\ \text{If } E_j \text{ is an intensive sampling event happens at subinterval end } s_{j+1} = t_{k-1}\text{:} \\ \quad \mathbb{1}_{\theta_k=\rho_k} \left(\dfrac{N_k}{\theta_k} + \dfrac{L(j)-N_k}{(1-\theta_k)} \right) + \dfrac{1-r_k}{(1-r_k)p_{k-1}\left(s_j\right)+r_k} \cdot \dfrac{\partial p_{k-1}(s_j)}{\partial \theta_k}, & (13) \\ \text{If } E_j \text{ is a epoch switching event happens at subinterval end } s_{j+1}\text{:} \\ \quad L(j) \cdot \dfrac{\partial Q_k(s_{j+1}, s_j)}{\partial \theta_k}. & (14) \end{cases}$$

Note that $\mathbb{1}$ is the indicator function. We leave the explicit expression of the shared terms in (11)-(14) to Supplementary Material S2.

Second, we consider the gradient at epoch k wrt the previous model parameters $\frac{\partial \mathbb{P}_k(j)}{\partial \theta_{k-i}}$, where $\theta_{k-i} \in \{\lambda_{k-i}, \psi_{k-i}, \mu_{k-i}, \rho_{k-i}\}$:

$$\frac{\partial \mathbb{P}_k(j)}{\partial \theta_{k-i}} = \begin{cases} \text{If } E_j \text{ is a birth event or epoch switching event happens at subinterval end } s_{j+1}: \\ \quad L(j) \cdot \dfrac{\partial Q_k(s_{j+1}, s_j)}{\partial \theta_{k-i}}, & (15) \\ \text{If } E_j \text{ is a serial sampling event happens at subinterval end } s_{j+1}: \\ \quad \dfrac{1-r_k}{(1-r_k)p_k(s_j)+r_k} \cdot \dfrac{\partial p_k(s_j)}{\partial \theta_{k-i}} + L(j) \cdot \dfrac{\partial Q_k(s_{j+1}, s_j)}{\partial \theta_{k-i}}. & (16) \end{cases}$$

We also leave the explicit expression of the shared terms in (15)-(16) in Section S2.

Third, we discuss the gradient at epoch k wrt the treatment probability $\mathbf{r}$. In (1), the treatment probabilities at different epochs only affect the current epoch. Therefore, we only need to consider $\frac{l_{k,j}}{\partial r_k}$ as follows:

$$\frac{\partial \mathbb{P}_k(j)}{\partial r_k} = \begin{cases} \text{If } E_j \text{ is a serial sampling event happens at subinterval end } s_{j+1}: \\ \quad \dfrac{1-p_k(s_j)}{(1-r_k)p_k(s_j)+r_k}, & (17) \\ \text{If } E_j \text{ is a intensive sampling event happens at subinterval end } s_{j+1} = t_{k-1}: \\ \quad \dfrac{1-p_{k-1}(s_j)}{(1-r_k)p_{k-1}(s_j)+r_k}. & (18) \end{cases}$$

The total gradient wrt $\mathbf{r}$ can be obtained similar to (10).

To determine the computation complexity of gradient evaluation, we can assume the gradient calculation for $\frac{\partial \mathbb{P}_k(j)}{\partial \theta_k}$ takes constant time. The model has K epochs, where each epoch has $\frac{(2N-1+K)}{K}$ phylogeny segments in average. According to (10), the total computation complexity is $\mathcal{O}(K \cdot \frac{(2N-1+K)}{2}) \sim \mathcal{O}(NK)$, since $K \ll N$. We demonstrate this result through a series of timing experiments presented in Supplementary Material S6 where we also compare the efficiency of gradients calculations with the automatic differentiation algorithm implemented in the VBSKY (Ki & Terhorst 2022) package based on JAX library (Bradbury

et al. 2018). Figure S5 shows our analytical gradients implemented in BEAST significantly outpace the VBSKY method.

2.6 Analysis

2.6.1 Examples

We evaluate the relative effectiveness of MH-MCMC and HMC transition kernels under the EBDS model using three phylodynamic examples. The first example comprises 274 sequences of the Pol locus of HIV-1 subtype A sampled in Odesa, Ukraine from 2000 to 2020 that Vasylyeva et al. (2020) previously analyzed to assess the population-level impact of the transmission reduction intervention project (TRIP) on HIV transmission (Nikolopoulos et al. 2016). Following this previous analysis, we establish a cutoff point of 50 years for the EBDS model. Within this period of time, we let the birth, death and sampling rates vary across 10 epochs mirroring the grid points specified by Vasylyeva et al. (2020). Note that for better comparability to the original work (Vasylyeva et al. 2020), we place iid lognormal priors on the rate parameters. Both the previous and our analysis assume an HKY nucleotide substitution (Hasegawa et al. 1985) model with discrete-gamma-distributed rate variation among sites (HKY+G) (Yang 1994), and an uncorrelated lognormal relaxed molecular clock model (Drummond et al. 2006) (UCLD), with a CTMC rate-reference prior (Ferreira & Suchard 2008) on the clock-model mean, truncated between 1×10^{-3} - 3×10^{-3}, and a normal prior (with mean $= 5 \times 10^{-4}$ and standard deviation $= 5 \times 10^{-4}$) on the standard deviation. We use a normal distribution prior (with mean $= 35$, standard deviation $= 5$) on the time to the most recent common ancestor, in accordance with the previous study.

Second, we examine the transmission dynamics of 637 human influenza A/H3N2 HA genes across 12 epidemic seasons sampled from New York state Rambaut et al. (2008) following the study of Parag et al. (2020). We set an EBDS model cutoff value of 13 years and infer time-varying birth and sampling rates across 78 epochs, each representing 2 months in time, and a constant-over-time death rate. Preceding studies focused on the evolution-

ary dynamics of influenza A/H1N1 virus mostly utilize the coalescent models. These studies predominantly rely on Gaussian process smoothing (Karcher et al. 2020, Bhattacharjee et al. 2023). Following the same path, we seek to use GMRF prior distributions for the birth and sampling rates. Our approach accommodates the considerable variability in the effective reproductive number across different flu seasons from 1993 to 2005. We adopt the same substitution and clock models from Rambaut et al. (2008). Specifically, to account for potential differences in the rate of substitution between the first and second codon positions compared to the third, we employ the SRD06 substitution model (Shapiro et al. 2006) and apply an HKY nucleotide substitution model with discrete-gamma distributed rate heterogeneity for both codon-position partitions (1st + 2nd, and 3rd). We further assume a UCLD clock model and employ the default priors from BEAST on the substitution and clock model parameters.

Lastly, to demonstrate the potential our linear-time algorithms afford phylodynamic analyses on larger data sets, we examine 1610 full Ebola virus (EBOV) genomes sampled between 17 March 2014 and 24 October 2015 from West Africa (Dudas et al. 2017) to explored the factors contributing to the spread of Ebola during the 2014-2016 epidemic. We set a EBDS model cutoff value of 2 years and infer time-varying birth and sampling rates for 24 epochs, each corresponding to a month in time, and a constant death rate. For choosing the priors on the rate parameters, we incorporate information from previous studies on the transmission dynamics of Ebola virus disease in West Africa (Fang et al. 2016, Nyenswah et al. 2016). The number of confirmed cases first persisted at a relatively low level and started to soar in the mid-Summer of 2014, followed by a consistent peak and a dramatic decrease after the initiation of some key intervention events. Considering the potential fast shifts projected to the effective reproductive number, we apply the Bayesian bridge MRF model as the prior for the incremental differences in the birth and sampling rates. Based on Dudas et al. (2017), we assume a HKY+G substitution model independently across four partitions (codon positions 1, 2, 3 and non-coding intergenic regions) and a log-normally-

distributed relaxed molecular clock model with a CTMC reference prior on the clock model mean, and leave all other priors on substitution and clock model parameters at their BEAST defaults.

2.6.2 Implementation

We conduct all analyses using extensions to BEAST 1.10 (Suchard et al. 2018) and the high-performance BEAGLE 4.0 library (Ayres et al. 2019) for efficient computation on central processing units (CPUs). We take the timing measurements using a Macbook Pro equipped with an M1 Pro chip that features 8 CPU cores and 32GB of RAM. For all experiments involving BEAST, we utilized the Azul Zulu Builds of OpenJDK version 18 on the ARM architecture.

To compare the performance of the two transition kernels in estimating the EBDS model parameters, we conduct efficiency comparison analyses that focused solely on the estimation of the birth-death model's rate parameters. Specifically, we fix the phylogeny to the maximum clade credibility (MCC) tree, a tree with the maximum product of the posterior clade probabilities summarized from the Bayesian joint phylogeny inference. We analyze all data sets using BEAST with logging performed every 1000 iterations. We run our algorithm on the HIV example for 300 million iterations when using MH-MCMC transition kernel and 30 million iterations for HMC transition kernel. Also, to obtain convergent results for the influenza example, we run analyses using MH-MCMC and HMC transition kernels for 300 million and 50 million states, respectively. For the Ebola example, we run analyses using MH-MCMC and HMC transition kernels for 100 million and 30 million states, respectively. For all analyses, we discard 10% of the MCMC chain samples as burn-in.

We calculate the effective sample size (ESS) for each rate parameter of interest using the coda package (Plummer et al. 2006) in CRAN (R Core Team 2021). ESS quantifies the degree to which auto-correlation within MCMC iterations contributes to uncertainty in estimates (Ripley 2009). We average ESS per compute-hour for each parameter across 10

independent runs to reduce Monte Carlo error in each estimate, aiming for a maximal Monte Carlo error of 10%. We report the relative increase in ESS per hour of the HMC sampler compared with the MH-MCMC sampler over all rate parameters.

We also conduct phylodynamic analysis for each of the three examples under a joint phylogeny inference scheme to mitigate potential bias from the fixed phylogeny, following the model specifications discussed in Section 2.6.1. Under these settings, we simulate MCMC chains for all examples of‘ 500 million iterations using HMC transition kernel with logging performed every 1000 iterations.

3 Results

3.1 Performance Improvements

Figure 2 shows the binned ESS per hour estimates of the EBDS model rates (λ, μ, ψ) that the MH-MCMC and HMC samples generate for all three viral examples. Table 1 summarizes the performance improvements by reporting the relative increase in the minimum ESS per hour comparing both samplers across all rate parameters.

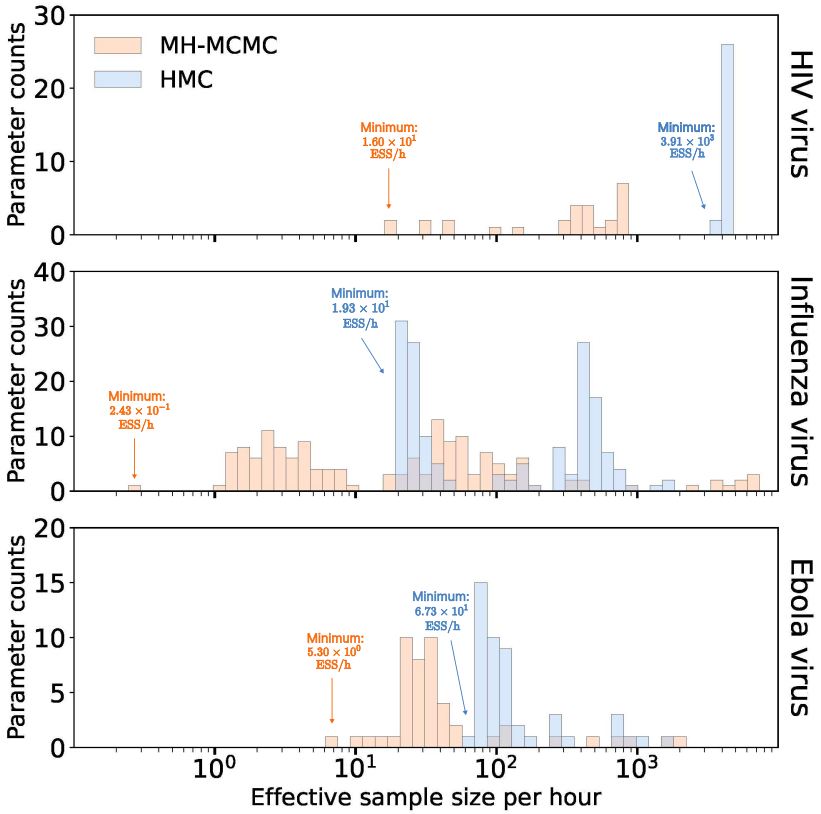

Figure 2: Efficiency Comparison between random walk Metropolis-Hastings (MH-MCMC) and Hamiltonian Monte Carlo (HMC) samplers. Bars correspond to the estimated effective sample size per hour averaged across 10 independent runs for all rate parameters. The height of each bar indicates the number of parameters that achieve the given ESS per hour value.

The HIV example assumes that time-varying rates are *a priori* independent across epochs and HMC demonstrates an approximate 245-fold acceleration relative to MH-MCMC. Likewise, the influenza example imposes a GMRF across epochs and returns an approximate 79.4-fold speed-up. On the other hand, the EBOV example enforces heavier shrinkage, and hence higher *a priori* correlation between epochs, and yields a smaller yet computationally impactful (approximately 12.7-fold) performance increase.

Example	Minimum ESS/h		HMC Speedup
	MH-MCMC	HMC	
HIV (10 epochs)	1.60×10^{1}	3.91×10^{3}	2.45×10^{2} times
Influenza (78 epochs)	2.43×10^{-1}	1.93×10^{1}	7.94×10^{1} times
Ebola (24 epochs)	5.30×10^{0}	6.73×10^{1}	1.27×10^{1} times

Table 1: Relative speedup in terms of effective sample size per hour (ESS/h) of HMC Over MH-MCMC for all three data Sets from fixed phylogeny analyses.

3.2 HIV dynamics in Odesa, Ukraine

In the context of conducting phylodynamic analyses using EBDS models, we are primarily interested in the value and trend of effective reproductive number over time $R_e(t)$ that is the average number of secondary cases per infectious case in a population made up of both susceptible and non-susceptible hosts. If $R_e > 1$, the number of cases is growing, such as at the start of an epidemic; if $R_e = 1$, the disease is endemic; and if $R_e < 1$, there is an expected decrease in transmission (Nishiura & Chowell 2009). Under the EBDS model, given the absence of intensive sampling events, if an individual becomes infected at time t, we can use the rate parameters at time t to obtain an estimated $R_e(t) = \frac{\lambda(t)}{\mu(t)+r(t)\psi(t)}$. Furthermore, in all our analyses for infectious disease phylodynamics, we maintain $r(t) = 1$ as constant. This assertion carries the assumption that upon diagnosis and sequencing, an individual ceases to be a source of infection. This could be due to treatment, death, or geographical relocation, rendering them incapable of onward transmission.

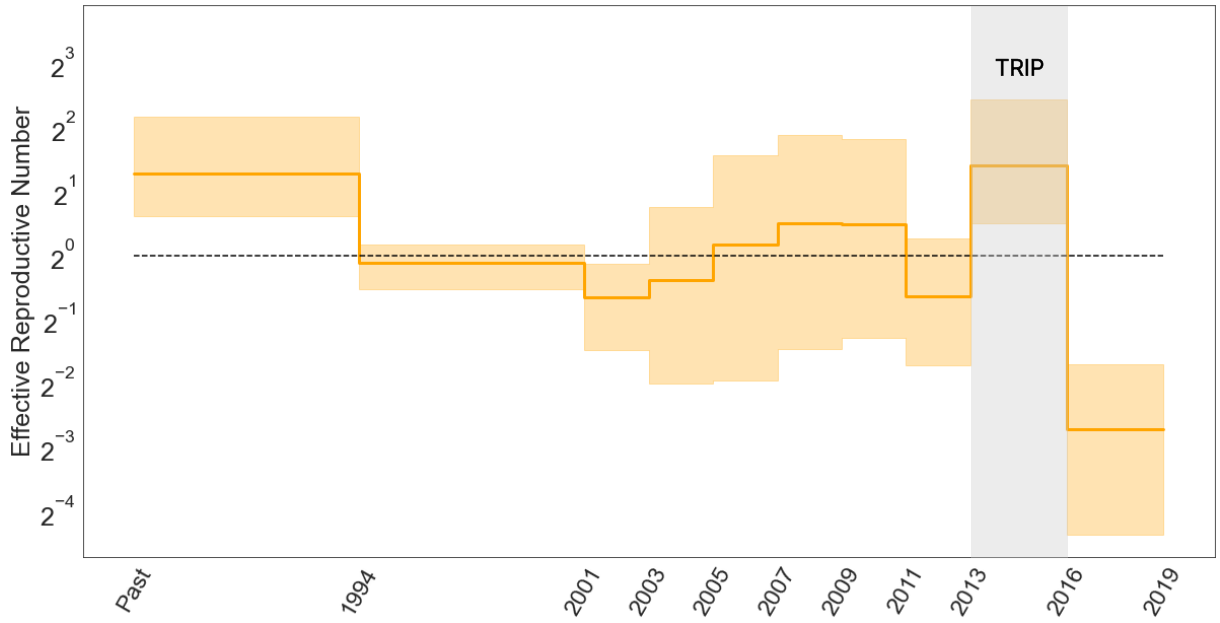

Figure 3: Posterior median (solid line) and 95% credible intervals (CI) indicated by the shaded areas of the effective reproductive number estimates (R_e) through time for HIV epidemic in Odesa, where the black dotted line represents the epidemiological threshold of $R_e(t) = 1$.

To assess the effects of TRIP for reducing the transmission of HIV in Odesa, we fit the EBDS model with varying birth, death and sampling rates and plot the resulting $R_e(t)$ trend estimate in Figure 3. We apply iid lognormal priors on the rate parameters to stay consistent with the methods in previous study (Vasylyeva et al. 2020).

Estimates of $R_e(t)$ appear mostly to accord with previous findings that identify a drop in infection rate subsequent to the implementation of the TRIP intervention. Focusing on the period from 2013 to early 2016, when TRIP was enacted, our posterior mean estimate of R_e is 2.64 (95 % CI: 1.18 - 5.43); while post-intervention, the posterior mean reduces to 0.152 (95 % CI: 0.03 - 0.32). This latter value, falling below the critical threshold of 1, signifies the potential deceleration of HIV transmission.

3.3 Seasonal Influenza in New York State

While influenza viruses circulate throughout the year, peak influenza outbreaks in the United States typically occurs between December and February. Rambaut et al. (2008) employed a

non-parametric coalescent model to elucidate the cyclical patterns of variation in the population size, uncovering a notable increase in genetic diversity at the beginning of each winter flu season. Subsequently, Parag et al. (2020) demonstrated that incorporating sampling intensity into the otherwise sampling-naive non-parametric coalescent process improves the precision of these inferred cycles. With a GMRF smoothing prior on increments, our model also offers the potential for accurately inferring seasonal behaviour and achieving the precision of parameter estimations.

Figure 4 presents posterior estimates of the effective reproductive number $R_e(t)$ for the alignment of 637 A/H3N2 HA sequences from New York state. As expected, the trajectory is highly cyclic, and all peaks lie near the midpoint of the influenza seasons with estimated R_e larger than 1. For the 2000/2001 and 2002/2003 seasons, where almost all infections were attributed to other sub-types of influenza viruses as indicated by the surveillance data and previous work (Centers for Disease Control and Prevention n.d., Parag et al. 2020), we observe the 95% CI of the estimated peak cover values from 0.68 to 1.3 and from 0.48 to 1.4, respectively. This suggests that their true R_e values might have fallen below 1. Similar to the results given by the non-parametric coalescent with sampling analysis (Parag et al. 2020), we capture a minor peak in the 1995/1996 season, where the inferred R_e is slightly above one. This again echoes with the fact that the influenza case composition during the 1995/1996 season was characterized by a mix of A/H1N1 and A/H3N2 infections (Ferguson et al. 2003). This diversity in infection types led to a less significant elevation in the effective reproductive number for that specific year.

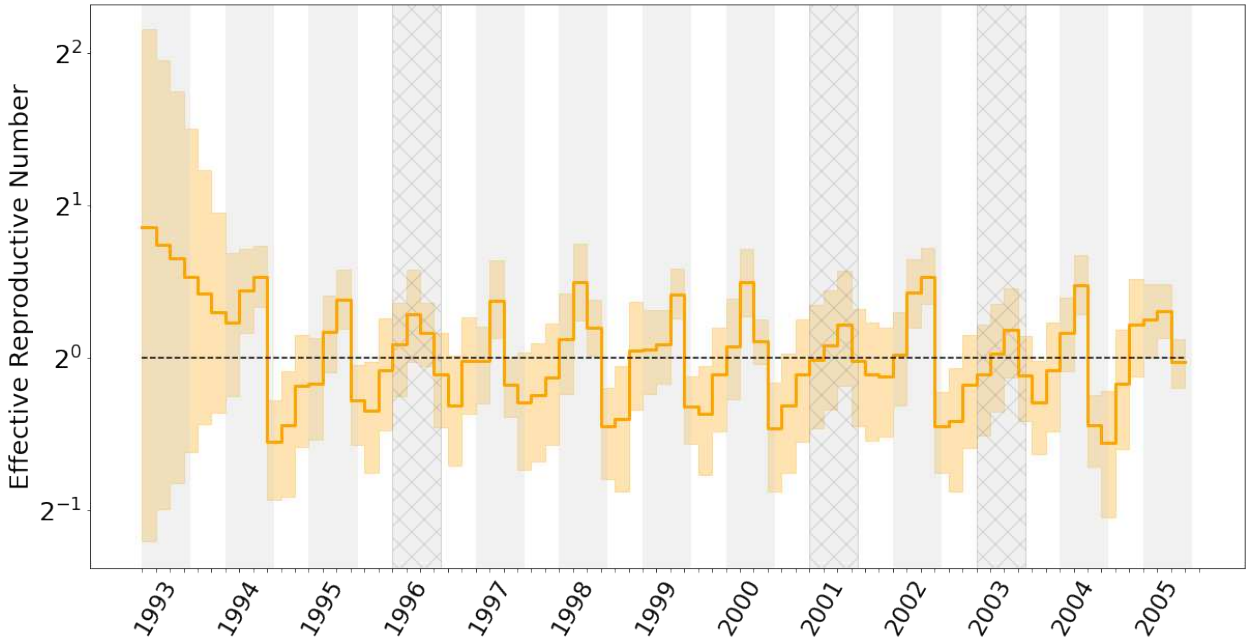

Figure 4: Median (solid orange line) and 95% credible intervals indicated by the shaded orange areas for the effective reproductive number estimates (Re) through time. Gray shading in the graph represents the rough duration of influenza monitored in New York state for each season, spanning from epidemiological week 40 to week 20 of the following year. Seasons where A/H3N2 was not the dominant influenza virus subtype are cross-hatched.

3.4 Ebola epidemic in West Africa

Using EBDS model assisted by the HMC sampler, we are able to analyze the 2014 Ebola epidemic in West Africa using the full 1610-sequence alignment and metadata of sampling times taken from the work by Dudas et al. (2017). Previously, researchers have applied birth-death models extensively for the phylodynamic analysis of the Ebola outbreak. Stadler et al. (2014) adopted a series of birth death models to capture the early trend of the infection of Ebola virus in Sierra-Leone. They used 72 Ebola samples from late May to mid June 2014 with three epochs, and estimated the corresponding effective reproductive number in each period. Zhukova et al. (2022) applied the multi-type birth death models to the 1610 sequence data. However, their analysis was based on the maximum likelihood estimation. To demonstrate the scalability of our method, we also take the 1610 sequence data and fit the EBDS model with 24 epochs for a finer time resolution to provide more precise estimation

of the effective reproductive number. Here, we employ a Bayesian bridge MRF prior on rate increments to avoid spurious rate variations while capturing significant rate shifts.

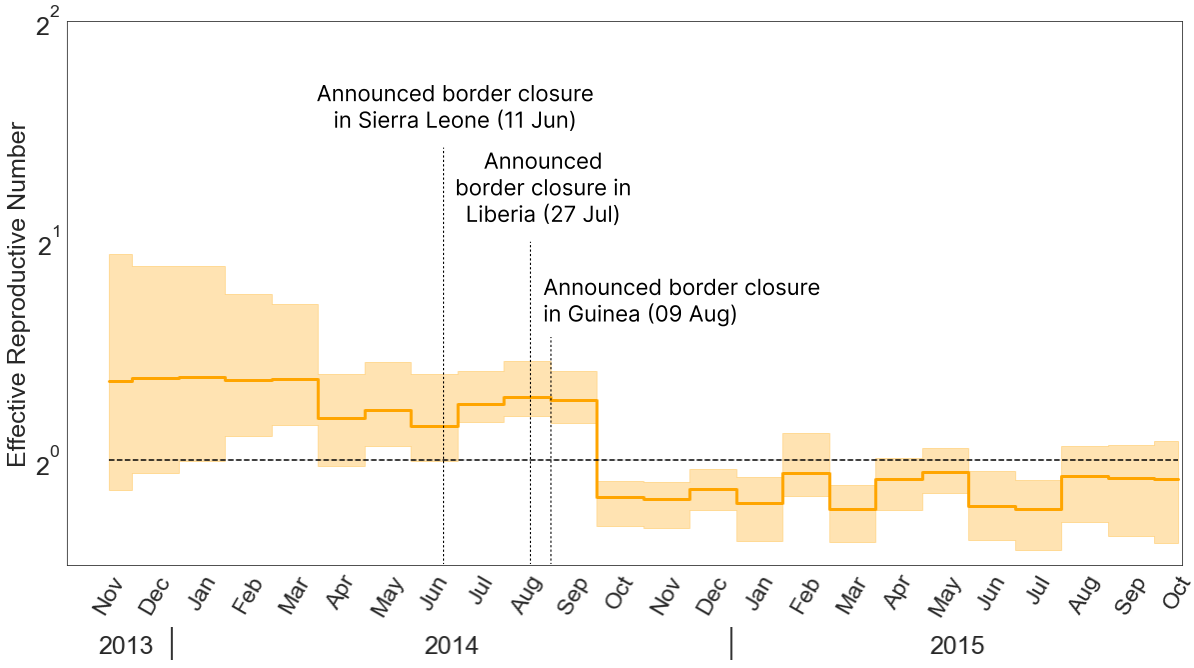

Figure 5: Median (solid line) and 95% credible intervals indicated by the shaded areas of the effective reproductive number estimates (R_e) through time for Ebola outbreak in west Africa. The black dotted line represents the epidemiological threshold of $R_e = 1$.

Our inference results give an estimated posterior mean effective reproductive number at the beginning of the epidemic before December 2013 as 1.65 (95 % CI: 0.41 - 3.05). Dudas et al. (2017) show that after the international border closure of Sierra Leone on 11 June 2014, followed by Liberia on 27 July 2014, and Guinea on 9 August 2014, the relative contribution of international border to overall viral migration is significantly lower. The change-point probability is the highest from August to September. This finding stands clearly compatible with our EBDS inference that demonstrates a drop of R_e from 1.3 (posterior mean, 95 % CI: 1.01 - 1.59) to 0.79 (95 % CI: 0.62 - 0.91) after September 2014 when the international travel restrictions are in place across the three countries.

4 Discussion

Birth-death models serve as fundamental tools for modeling the temporal progression of epidemics. In extending the work of Stadler et al. (2013), Gavryushkina et al. (2014), we have provided a systematic representation of the EBDS model for phylodynamics that promotes scalability. Our general re-formalization of the EBDS likelihood identifies that its computation is simply $\mathcal{O}(N + K)$, foreshadowing an $\mathcal{O}(NK)$ algorithm to deliver its gradient wrt time-varying birth, death or sampling rates across K epochs. This optimal scaling enables HMC sampling to more efficiently explore the high-dimensional joint distribution of rates as we increase the number of sequences and the number of model epochs to learn these processes at a finer time-resolution. HMC also emits an agnostic approach to incorporate a variety of prior assumptions about these time-varying trends, without the need to hand-craft specialized transitions kernels for specific priors. Moreover, as suggested by Ji et al. (2020), we take measures to enhance the efficiency of our HMC sampler by preconditioning the mass matrix based on the Hessian of the log-prior.

Through three viral epidemic examples, we show that our HMC-assisted approach considerably accelerates Bayesian inference across three very different choices of prior models. Our preconditioned HMC sampler achieves roughly 10- to 200-fold increase over the widely used MH-MCMC sampler in terms of the minimum ESS per unit-time. The enhanced efficiency gains are particularly beneficial given the increasing use of phylodynamic inference techniques in conducting real-time evaluations of outbreak patterns.

For applying our model in phylodynamic analyses of disease epidemics, we first examine our EBDS model on the effects of TRIP for reducing the transmission of HIV in Ukraine, and our inference results support a decreased rate of transmission following the TRIP intervention. Applied to seasonal Influenza in New York city, our model is able to accurately capture the complex pattern of variation in R_e during each influenza season. Applied to the Ebola outbreak in West Africa, our model supports the effect of international travel restrictions characterized as a noticeable decrease in R_e following the border closure of the

three countries in West Africa.

In the EBDS model, Stadler and colleagues (Stadler et al. 2013) have indicated that the three rate parameters, λ, μ, and ψ, cannot be simultaneously identified. This issue of unidentifiability in complex birth-death processes has also been recently discussed by Louca & Pennell (2020). In our own empirical analysis, problems related to unidentifiability seldom manifest when we restrict ourselves to estimating no more than two time-varying rate parameters. Instead, the primary challenge appears to be the multimodal nature of the posterior distribution. Legried & Terhorst (2022) have demonstrated that, under certain conditions, piecewise constant birth-death models can be reliably inferred and differentiated. Furthermore, Kopperud et al. (2023) showed that rapidly changing speciation or extinction rates can be accurately estimated. This lends credence to the identifiability of patterns we observed in our phylodynamic analysis of pandemics such as the seasonal influenza and the Ebola outbreaks.

Current methods to estimate the expected Hessian averaged over the posterior distribution improves upon the previous work (Girolami & Calderhead 2011) by avoiding excessive computational burden. However, it relies on numerical approximations to compute the Hessian, leaving room for potential performance enhancements. To further optimize the methodology, we can advance beyond analytical solutions solely for gradients and extend them to encompass the analytical Hessian. This would smooth the path of updating the adaptive mass matrix, offering opportunities for better outcomes in terms of both efficiency and accuracy.

In many scenarios, the examination of EBDS models is contingent upon having some preliminary understanding of how to identify the epoch switching time and the length of duration of each epoch. However, it is possible that information available through epidemiological surveillance is insufficient. Moreover, the choice of epoch duration can be related to the uncertainty in the timing of the rate shifts (Magee et al. 2020). In this study, our strategy aims to increase the number of epochs and leverage regularizing priors, striving to achieve a

refined grid of timelines. Nevertheless, constraints persist on the maximum epochs feasible with our HMC algorithm, particularly when confronted with computational limitations or models exhibiting multimodality challenges. One possible solution entails simultaneously inferring epoch duration, epoch switching times, and rate parameters via the reversible-jump MCMC method (Wu 2014). However, this method requires one to integrate across models with differing dimension, which demands substantial effort and might be impractical for large datasets.

Considering these cases, if the piece-wise constant model assumptions can be lifted so that we can obtain a smoothly differentiable likelihood function, it would inherently aid in deriving gradients concerning node ages and epoch switching times. This advancement would, in turn, improve our current implementation, empowering us to infer, rather than presuppose, epoch switching times, with enhanced scalability prospects. It would also enhance the sampling efficiency from joint phylogeny posterior distributions, by enabling us to take advantage of recent work by Ji et al. (2021), yielding a pronounced improvement in the analytical capacity of our models.

In anticipation of future advancements that will improve upon standard HMC methods and broaden the applicability of the current EBDS model, we present a comprehensive framework in this manuscript. This framework facilitates phylodynamic analysis on large-scale sequence data and employs regularization techniques to yield a finely-resolved, regular grid that effectively aids in our understanding of the impact of the pandemics.

5 Acknowledgements

This work was supported through National Institutes of Health grants U19 AI135995, R01 AI153044 and R01 AI162611. We gratefully acknowledge support from Advanced Micro Devices, Inc. with the donation of parallel computing resources used for this research.

Bibliography

Ayres, D. L., Cummings, M. P., Baele, G., Darling, A. E., Lewis, P. O., Swofford, D. L., Huelsenbeck, J. P., Lemey, P., Rambaut, A. & Suchard, M. A. (2019), 'BEAGLE 3: improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics', Systematic Biology **68**, 1052–1061.

Baele, G., Gill, M. S., Lemey, P. & Suchard, M. A. (2020), 'Hamiltonian Monte Carlo sampling to estimate past population dynamics using the skygrid coalescent model in a Bayesian phylogenetics framework', Wellcome Open Research **5**, 53.

Barido-Sottani, J., Vaughan, T. G. & Stadler, T. (2020), 'A multitype birth–death model for Bayesian inference of lineage-specific birth and death rates', Systematic Biology **69**, 973–986.

Bhattacharjee, U., Chakrabarti, A. K., Kanungo, S. & Dutta, S. (2023), 'Evolutionary dynamics of influenza A/H1N1 virus circulating in India from 2011 to 2021', Infection, Genetics and Evolution **110**, 105424.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N. et al. (2019), 'BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis', PLoS computational biology **15**, e1006650.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S. et al. (2018), 'JAX: Composable transformations of python+ numPy programs (v0. 2.5)', Software available from https://github.com/google/jax .

Centers for Disease Control and Prevention (n.d.), 'Key facts about influenza (flu)', https://www.cdc.gov/flu/about/keyfacts.html. Accessed: 2023-05-31.

621 Condamine, F. L., Rolland, J., Höhna, S., Sperling, F. A. & Sanmartín, I. (2018), 'Testing
622 the role of the Red Queen and Court Jester as drivers of the macroevolution of Apollo
623 butterflies', Systematic Biology **67**, 940–964.

624 Crawford, F. W. (2012), General birth-death processes: probabilities, inference, and appli-
625 cations, PhD thesis, UCLA.

626 Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. (2006), 'Relaxed phyloge-
627 netics and dating with confidence', PLoS Biology **4**, e88.

628 Du Plessis, L. (2016), Understanding the spread and adaptation of infectious diseases using
629 genomic sequencing data, PhD thesis, ETH Zurich.

630 Duane, S., Kennedy, A. D., Pendleton, B. J. & Roweth, D. (1987), 'Hybrid Monte Carlo',
631 Physics Letters B **195**, 216–222.

632 Dudas, G., Carvalho, L. M., Bedford, T., Tatem, A. J., Baele, G., Faria, N. R., Park, D. J.,
633 Ladner, J. T., Arias, A., Asogun, D. et al. (2017), 'Virus genomes reveal factors that
634 spread and sustained the Ebola epidemic', Nature **544**, 309–315.

635 Fang, L.-Q., Yang, Y., Jiang, J.-F., Yao, H.-W., Kargbo, D., Li, X.-L., Jiang, B.-G., Kargbo,
636 B., Tong, Y.-G., Wang, Y.-W. et al. (2016), 'Transmission dynamics of Ebola virus disease
637 and intervention effectiveness in Sierra Leone', Proceedings of the National Academy of
638 Sciences **113**, 4488–4493.

639 Ferguson, N. M., Galvani, A. P. & Bush, R. M. (2003), 'Ecological and immunological
640 determinants of influenza evolution', Nature **422**, 428–433.

641 Ferreira, M. A. & Suchard, M. A. (2008), 'Bayesian analysis of elapsed times in continuous-
642 time Markov chains', Canadian Journal of Statistics **36**, 355–368.

643 Fisher, A. A., Ji, X., Nishimura, A., Lemey, P. & Suchard, M. A. (2021), 'Shrinkage-based
644 random local clocks with scalable inference', arXiv preprint arXiv:2105.07119 .

645 FitzJohn, R. G. (2010), 'Quantitative traits and diversification', Systematic biology **59**, 619–
646 633.

647 FitzJohn, R. G. (2012), 'Diversitree: comparative phylogenetic analyses of diversification in
648 R', Methods in Ecology and Evolution **3**, 1084–1092.

649 Gavryushkina, A., Welch, D., Stadler, T. & Drummond, A. J. (2014), 'Bayesian inference
650 of sampled ancestor trees for epidemiology and fossil calibration', PLoS Computational
651 Biology **10**, e1003919.

652 Girolami, M. & Calderhead, B. (2011), 'Riemann manifold Langevin and Hamiltonian
653 Monte Carlo methods', Journal of the Royal Statistical Society: Series B (Statistical
654 Methodology) **73**, 123–214.

655 Hasegawa, M., Kishino, H. & Yano, T.-a. (1985), 'Dating of the human-ape splitting by a
656 molecular clock of mitochondrial DNA', Journal of Molecular Evolution **22**, 160–174.

657 Hastings, W. K. (1970), 'Monte Carlo sampling methods using Markov chains and their
658 applications', Biometrika **57**, 97–109.

659 Höhna, S. (2014), 'Likelihood inference of non-constant diversification rates with incomplete
660 taxon sampling', PLoS one **9**, e84184.

661 Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck,
662 J. P. & Ronquist, F. (2016), 'RevBayes: Bayesian phylogenetic inference using graphical
663 models and an interactive model-specification language', Systematic biology **65**, 726–736.

664 Ji, X., Fisher, A. A., Su, S., Thorne, J. L., Potter, B., Lemey, P., Baele, G. & Suchard,
665 M. A. (2021), 'Scalable Bayesian divergence time estimation with ratio transformations',
666 arXiv preprint arXiv:2110.13298 .

667 Ji, X., Zhang, Z., Holbrook, A., Nishimura, A., Baele, G., Rambaut, A., Lemey, P. &

668 Suchard, M. A. (2020), 'Gradients do grow on trees: a linear-time O(N)-dimensional
669 gradient for statistical phylogenetics', Molecular Biology and Evolution **37**, 3047–3060.

670 Karcher, M. D., Carvalho, L. M., Suchard, M. A., Dudas, G. & Minin, V. N. (2020), 'Es-
671 timating effective population size changes from preferentially sampled genetic sequences',
672 PLoS Computational Biology **16**, e1007774.

673 Ki, C. & Terhorst, J. (2022), 'Variational phylodynamic inference using pandemic-scale data',
674 Molecular Biology and Evolution **39**, msac154.

675 Kopperud, B. T., Magee, A. F. & Höhna, S. (2023), 'Rapidly changing speciation and
676 extinction rates can be inferred in spite of nonidentifiability', Proceedings of the National
677 Academy of Sciences **120**, e2208851120.

678 Lambert, A. & Stadler, T. (2013), 'Birth–death models and coalescent point processes:
679 The shape and probability of reconstructed phylogenies', Theoretical Population Biology
680 **90**, 113–128.

681 Lau, M. S., Grenfell, B. T., Worby, C. J. & Gibson, G. J. (2019), 'Model diagnostics and
682 refinement for phylodynamic models', PLoS Computational Biology **15**, e1006955.

683 Legried, B. & Terhorst, J. (2022), 'A class of identifiable phylogenetic birth–death models',
684 Proceedings of the National Academy of Sciences **119**, e2119513119.

685 Louca, S. & Pennell, M. W. (2020), 'Extant timetrees are consistent with a myriad of diver-
686 sification histories', Nature **580**, 502–505.

687 MacPherson, A., Louca, S., McLaughlin, A., Joy, J. B. & Pennell, M. W. (2022), 'Unifying
688 phylogenetic birth–death models in epidemiology and macroevolution', Systematic Biology
689 **71**, 172–189.

690 Maddison, W. P., Midford, P. E. & Otto, S. P. (2007), 'Estimating a binary character's effect
691 on speciation and extinction', Systematic biology **56**, 701–710.

692 Magee, A. F. & Höhna, S. (2021), 'Impact of K-Pg mass extinction event on crocodylomorpha
693 inferred from phylogeny of extinct and extant taxa', bioRxiv pp. 2021–01.

694 Magee, A. F., Höhna, S., Vasylyeva, T. I., Leaché, A. D. & Minin, V. N. (2020), 'Locally
695 adaptive Bayesian birth-death model successfully detects slow and rapid rate shifts', PLoS
696 Computational Biology **16**, e1007999.

697 Minosse, C., Salichos, L., Taibi, C., Luzzitelli, I., Nardozi, D., Capobianchi, M. R., D'Offizi,
698 G., McPhee, F. & Garbuglia, A. R. (2021), 'Phylogenetic and phylodynamic analy-
699 ses of HCV strains circulating among patients using injectable drugs in central Italy',
700 Microorganisms **9**, 1432.

701 Morlon, H., Parsons, T. L. & Plotkin, J. B. (2011), 'Reconciling molecular phylogenies with
702 the fossil record', Proceedings of the National Academy of Sciences **108**, 16327–16332.

703 Neal, R. M. et al. (2011), 'MCMC using Hamiltonian dynamics', Handbook of Markov Chain
704 Monte Carlo **2**, 2.

705 Nikolopoulos, G. K., Pavlitina, E., Muth, S. Q., Schneider, J., Psichogiou, M., Williams,
706 L. D., Paraskevis, D., Sypsa, V., Magiorkinis, G., Smyrnov, P. et al. (2016), 'A network
707 intervention that locates and intervenes with recently hiv-infected persons: The transmis-
708 sion reduction intervention project (TRIP)', Scientific reports **6**, 38100.

709 Nishimura, A. & Suchard, M. A. (2023), 'Shrinkage with shrunken shoulders: Gibbs sampling
710 shrinkage model posteriors with guaranteed convergence rates', Bayesian Analysis **18**, 367–
711 390.

712 Nishiura, H. & Chowell, G. (2009), 'The effective reproduction number as a prelude to
713 statistical estimation of time-dependent epidemic trends', Mathematical and Statistical
714 Estimation Approaches in Epidemiology pp. 103–121.

Novitsky, V., Kühnert, D., Moyo, S., Widenfelt, E., Okui, L. & Essex, M. (2015), 'Phylodynamic analysis of HIV sub-epidemics in Mochudi, Botswana', Epidemics **13**, 44–55.

Nyenswah, T. G., Kateh, F., Bawo, L., Massaquoi, M., Gbanyan, M., Fallah, M., Nagbe, T. K., Karsor, K. K., Wesseh, C. S., Sieh, S. et al. (2016), 'Ebola and its control in Liberia, 2014–2015', Emerging Infectious Diseases **22**, 169.

Parag, K. V., du Plessis, L. & Pybus, O. G. (2020), 'Jointly inferring the dynamics of population size and sampling intensity from molecular sequences', Molecular Biology and Evolution **37**, 2414–2429.

Piironen, J. & Vehtari, A. (2017), 'Sparsity information and regularization in the horseshoe and other shrinkage priors'.

Plummer, M., Best, N., Cowles, K. & Vines, K. (2006), 'CODA: Convergence diagnosis and output analysis for MCMC', R News **6**, 7–11.
URL: *https://journal.r-project.org/archive/*

R Core Team (2021), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Rambaut, A., Pybus, O. G., Nelson, M. I., Viboud, C., Taubenberger, J. K. & Holmes, E. C. (2008), 'The genomic and epidemiological dynamics of human influenza A virus', Nature **453**, 615–619.

Ripley, B. D. (2009), Stochastic simulation, John Wiley & Sons.

Shapiro, B., Rambaut, A. & Drummond, A. J. (2006), 'Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences', Molecular Biology and Evolution **23**, 7–9.

738 Silvestro, D., Tejedor, M. F., Serrano-Serrano, M. L., Loiseau, O., Rossier, V., Rol-
739 land, J., Zizka, A., Höhna, S., Antonelli, A. & Salamin, N. (2019), 'Early arrival and
740 climatically-linked geographic expansion of New World monkeys from tiny African ances-
741 tors', Systematic Biology **68**, 78–92.

742 Stadler, T. (2010), 'Sampling-through-time in birth–death trees', Journal of Theoretical
743 Biology **267**, 396–404.

744 Stadler, T., Kühnert, D., Bonhoeffer, S. & Drummond, A. J. (2013), 'Birth–death skyline
745 plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV)',
746 Proceedings of the National Academy of Sciences **110**, 228–233.

747 Stadler, T., Kühnert, D., Rasmussen, D. A. & du Plessis, L. (2014), 'Insights into the early
748 epidemic spread of Ebola in Sierra Leone provided by viral sequence data', PLoS Currents
749 **6**.

750 Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J. & Rambaut, A.
751 (2018), 'Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10',
752 Virus Evolution **4**, vey016.

753 Vasylyeva, T. I., Zarebski, A., Smyrnov, P., Williams, L. D., Korobchuk, A., Liulchuk, M.,
754 Zadorozhna, V., Nikolopoulos, G., Paraskevis, D., Schneider, J. et al. (2020), 'Phylody-
755 namics helps to evaluate the impact of an HIV prevention intervention', Viruses **12**, 469.

756 Velásquez, G. E., Aibana, O., Ling, E. J., Diakite, I., Mooring, E. Q. & Murray, M. B. (2015),
757 'Time from infection to disease and infectiousness for Ebola virus disease, a systematic
758 review', Clinical Infectious Diseases **61**, 1135–1140.

759 Wu, C.-H. (2014), Bayesian approaches to model uncertainty in phylogenetics, Ph.d. thesis,
760 University of Auckland.

761 Yang, Z. (1994), 'Maximum likelihood phylogenetic estimation from DNA sequences with
762 variable rates over sites: approximate methods', Journal of Molecular Evolution **39**, 306–
763 314.

764 Yang, Z. & Rannala, B. (1997), 'Bayesian phylogenetic inference using DNA sequences: a
765 Markov Chain Monte Carlo method.', Molecular biology and evolution **14**, 717–724.

766 Zhukova, A., Hecht, F., Maday, Y. & Gascuel, O. (2022), 'Fast and accurate maximum-
767 likelihood estimation of multi-type birth-death epidemiological models from phylogenetic
768 trees', medRxiv pp. 2022–08.

Supplementary Material

S1 Likelihood Derivation

S1.1 Formulas for Likelihood Related Functions

$$A_k = \sqrt{(\lambda_k - \mu_k - \psi_k)^2 + 4\lambda_k\psi_k}, \tag{19}$$

$$B_k = \frac{(1 - 2(1-\rho_k)\, p_{k-1}(t_{k-1}))\,\lambda_k + \mu_k + \psi_k}{A_k} \tag{20}$$

$$p_k(t) = \frac{\lambda_k + \mu_k + \psi_k - A_k \frac{e^{A_k(t-t_{k-1})}(1+B_k)-(1-B_k)}{e^{A_k(t-t_{i-1})}(1+B_k)+(1-B_k)}}{2\lambda_k} \tag{21}$$

$$q_k(t) = \frac{4e^{A_k(t-t_{k-1})}}{(e^{A_k(t-t_{k-1})}(1+B_k) + (1-B_k))^2} \tag{22}$$

$$g_1 = e^{A_k(t-t_{k-1})} \cdot (1+B_k) + (1-B_k) \tag{23}$$

$$g_2 = A_k(1 - \frac{2(1-B_k)}{g_1}) \tag{24}$$

$$g_3 = 1 - 2(1-\rho_k))P_{k-1}(t_{k-1}) \tag{25}$$

S1.2 Implementation Algorithm:

Detailed algorithm for likelihood calculation is shown below based on the equations listed in Section 2.2 of the main text and from the section above.

Algorithm 1: Likelihood Calculation

Initialize: $p_0(t_0) = 1$
for $k = 0, \ldots, K-1$ **do**
  /* Intermediate quantities */
  Load the value of $p_k(t_k)$
  Calculate A_{k+1}, B_{k+1} via Equation (19), (20)
  for $j = 0, \ldots, m_{k+1} - 1$ **do**
    Calculate $q_{k+1}(s_{j+1})$ via Equation (22)
    if s_{j+1} is a serial sampling event **then**
      Calculate $p_{k+1}(s_{j+1})$ via Equation (21)
    end
    if $j \geq 1$ **then**
      Calculate $I_k(E_j)$ via Equation (2)
    end
  end
  Calculate and store $p_{k+1}(t_{k+1})$ via Equation (21)
end
/* Likelihood */
Calculate $\mathbb{P}[\mathcal{T} \mid \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\psi}, \boldsymbol{\rho}, \boldsymbol{r}, \boldsymbol{t}]$ via Equation (1)

S2 Gradient Derivation

S2.1 For $\frac{\partial \mathbb{P}_k(j)}{\partial \theta_k}$:

$$\frac{\partial q_k(t)}{\partial \theta_k} = \frac{8e^{A_k(t-t_{k-1})}((t-t_{k-1})\frac{\partial A_k}{\partial \theta_k}(\frac{1}{2}\cdot g_1 - e^{A_k(t-t_{k-1})}\cdot(1+B_k))}{g_1^3} - \frac{\frac{\partial B_k}{\partial \theta_k}(e^{A_k(t-t_{k-1})}-1))}{g_1^3} \tag{26}$$

$$\frac{\partial A_k}{\partial \theta_k} = \begin{cases} \frac{\lambda_k-\mu_k+\psi_k}{A_k}, & \text{If } \theta = \lambda \\ \frac{-\lambda_k+\mu_k+\psi_k}{A_k}, & \text{If } \theta = \mu \\ \frac{\lambda_k+\mu_k+\psi_k}{A_k}, & \text{If } \theta = \psi \\ 0, & \text{If } \theta = \rho \end{cases} \tag{27}$$

$$\frac{\partial B_k}{\partial \theta_k} = \begin{cases} \frac{2\lambda_k p_{k-1}(t_{k-1})}{A_k}, & \text{If } \theta = \rho \\ \frac{\partial B_k}{\partial \theta_k} = \frac{A_k\cdot \text{temp} - \frac{\partial A_k}{\partial \theta_k}\cdot(g_3\cdot\lambda_k+\mu_k+\psi_k)}{A_k^2}, & \text{Otherwise} \end{cases} \tag{28}$$

$$\frac{\partial p_k(t)}{\partial \theta_k} = \begin{cases} \frac{1}{2\lambda_k^2}(-\mu_k-\psi_k-\lambda_k\frac{\partial g_2}{\partial \lambda_k}+g_2), & \text{If } \theta = \lambda \\ -\frac{A_k}{\lambda_k}\frac{\left((1-B_k)(e^{A_k(t-t_{k-1})}-1)+g_1\right)\frac{\partial B_k}{\partial \rho_k}}{g_1^2}, & \text{If } \theta = \rho \\ \frac{1}{2\theta_k}(1-\frac{\partial g_2}{\partial \theta_k}), & \text{Otherwise} \end{cases} \tag{29}$$

$$\frac{\partial Q_k(s_{j+1},s_j)}{\partial \theta_k} = \frac{1}{q_k(s_{j+1})}\frac{\partial q_k(s_{j+1})}{\partial \theta_k} - \frac{1}{q_k(s_j)}\frac{\partial q_k(s_j)}{\partial \theta_k} \tag{30}$$

$$\frac{\partial g_2}{\partial \theta_k} = \frac{dA_k}{d\theta_k} - \frac{2}{g_1^2}\cdot\left(g_1\left\{\frac{dA_k}{d\theta_k}(1-B_k) - \frac{dB_k}{d\theta_k}\cdot A_k\right\} - \left(e^{A_k(t-t_{k-1})}\frac{\partial A_k}{\partial \theta_k}(1+B_k)\cdot(t-t_{k-1}) + (e^{A_k(t-t_{k-1})}-1)\frac{\partial B_k}{\partial \theta_k}\right)\cdot A_k(1-B_k)\right) \tag{31}$$

S2.2 For $\frac{\partial \mathbb{P}_k(j)}{\partial \theta_{k-i}}$ (i is an integer smaller than k):

$$\frac{\partial q_k(t)}{\partial \theta_{k-i}} = -\frac{8e^{A_k(t-t_{k-1})}\frac{\partial B_k}{\partial \theta_{k-i}}(e^{A_k(t-t_{k-1})}-1)}{g_1^3} \tag{32}$$

$$\frac{\partial B_k}{\partial \theta_{k-i}} = \frac{\partial B_k}{\partial p_{k-1}(t_{k-1})}\cdot\frac{\partial p_{k-1}(t_{k-1})}{\partial \theta_{k-i}} = \frac{-2(1-\rho_k)\lambda_k}{A_k}\frac{\partial p_{k-1}(t_{k-1})}{\partial \theta_{k-i}} \tag{33}$$

$$\frac{\partial p_k(t)}{\partial \theta_{k-i}} = -\frac{A_k}{\lambda_k}\frac{\left((1-B_k)(e^{A_k(t-t_{k-1})}-1)+g_1\right)\frac{\partial B_k}{\partial \theta_{k-i}}}{g_1^2} \tag{34}$$

$$\frac{\partial Q_k(s_{j+1},s_j)}{\partial \theta_{k-i}} = \frac{1}{q_k(s_{j+1})}\frac{\partial q_k(s_{j+1})}{\partial \theta_{k-i}} - \frac{1}{q_k(s_j)}\frac{\partial q_k(s_j)}{\partial \theta_{k-i}} \tag{35}$$

S2.3 Implementation Algorithm:

We implement a recursive algorithm to compute the necessary gradient of the log-likelihood within our rate parameter space. Intermediate quantities are stored in between epochs to alleviate computational burden. Detailed algorithm is shown below based on the equations listed in 2.5 and previous sections in the supplement.

Algorithm 2: Gradient Calculation

Initialize: $p_0(t_0) = 1$
for $k = 0, \ldots, K-1$ **do**
  /* Intermediate quantities */
  if $k == 0$ **then**
    Calculate $\frac{\partial A_1}{\partial \theta_1}$, $\frac{\partial B_1}{\partial \theta_1}$ using $p_0(t_0)$ via Equation (27), (28)
  end
  else if $k \geq 1$ **then**
    Load the values of $\{\frac{\partial p_k(t_k)}{\partial \theta_i}\}_{i=1}^{k}$
    Calculate $\frac{\partial A_{k+1}}{\partial \theta_{k+1}}$, $\{\frac{\partial B_{k+1}}{\partial \theta_i}\}_{i=1}^{k+1}$ using $\{\frac{\partial p_k(t_k)}{\partial \theta_i}\}_{i=1}^{k}$ via Equation (27), (28), (33)
  end
  Calculate and store $\{\frac{\partial p_{k+1}(t_{k+1})}{\partial \theta_i}\}_{i=1}^{k+1}$ using $\{\frac{\partial B_{k+1}}{\partial \theta_i}\}_{i=1}^{k}$ via Equation (29), (34)
  /* Gradient */
  Calculate $\{\frac{\partial \mathbb{P}_k(j)}{\partial \theta_i}\}_{i=1}^{k}$ via Equations (11)-(18) in Section 2.5
end

S3 Prior distributions for EBDS models

S3.1 HIV dynamics in Odesa, Ukraine

We refer to the prior settings on the compound parameters from previous work (Vasylyeva et al. 2020), and try to roughly match their priors by adopting the following prior distributions on each of the rate parameters. Note that the sampling proportion was fixed to 0 before the first sampling date in their study, so we also set the sampling rate to 0 for the last two epochs for consistency.

Parameter	Prior	Role
λ	Lognormal (Mean = 0.85, SD = 1.0)	Birth rate
μ	Lognormal (Mean = -0.25, SD = 1.0)	Death rate
ψ	Lognormal (Mean = -9.0, SD = 0.50)	Serial sampling rate
t_{or}	Uniform (Lower = 19, Upper = 60)	Age of phylogeny

Table S1: Prior specifications for the EBDS model in HIV virus analysis

S3.2 Seasonal Influenza in New York State

We follow the same framework for setting the priors for the GMRF-based model as in Section S3.3. Similarly, the prior distribution for the constant death rate is acquired by estimating the credible range for the duration of the infectious period according to reports by Centers for Disease Control and Prevention (n.d.), with 95% confidence intervals encompassing 6 to 11 days. Comprehensive information regarding the specific prior distributions is shown in the following table:

Parameter	Prior	Role
λ_1^*	Normal (Mean = 3.08, SD = 1.17)	Log-scale birth rate at present
μ_k^*	Normal (Mean = 3.82, SD = 0.16)	Log-scale death rate for all epochs
ψ_1^*	Normal (Mean = -0.77, SD = 1.17)	Log-scale sampling rate at present
t_{or}	Normal (Mean = 12.5, SD = 15.0)	Age of phylogeny
α	Fixed to 2.0	Exponent of the MRF
ϕ	Gamma (Shape = 1.0, Scale = 1.0)	Transformed global scale of the MRF
ν_k	Fixed to 1.0	Local scale of MRF

Table S2: Prior specifications for the EBDS model in Influenza virus analysis

S3.3 Ebola epidemic in West Africa

We assume a constant death rate, μ for this data set, and we employ an empirical Bayes approach proposed by Magee et al. (2020) to set the prior on the first log-birth-rate and log-sampling-rate in our Bayesian bridge MRF models. The prior for the constant death rate is obtained from an estimation of the plausible duration of infectious period with 95% confidence intervals covering 8 to 40 days (Velásquez et al. 2015). The detailed prior distributions

803 can be found in the table below:

Parameter	Prior	Role
λ_1^*	Normal (Mean = 1.26, SD = 0.58)	Log-scale birth rate at present
μ_k^*	Normal (Mean = 3.02, SD = 0.41)	Log-scale death rate for all epochs
ψ_1^*	Normal (Mean = 1.27, SD = 0.58)	Log-scale sampling rate at present
t_{or}	Normal (Mean = 1.89, SD = 15.0)	Age of phylogeny
α	Fixed to 0.25	Exponent of the MRF
ϕ	Gamma (Shape = 1.0, Scale = 1.0)	Transformed global scale of the MRF
ν_k	Exponentially tilted stable distributions	Local scale of Bayesian bridge MRF
ξ	Fixed to 2.0	Slab width of Bayesian bridge MRF

Table S3: Prior specifications for the EBDS model in Ebola virus analysis

S4 Inferred trajectories for birth/death/sampling rates

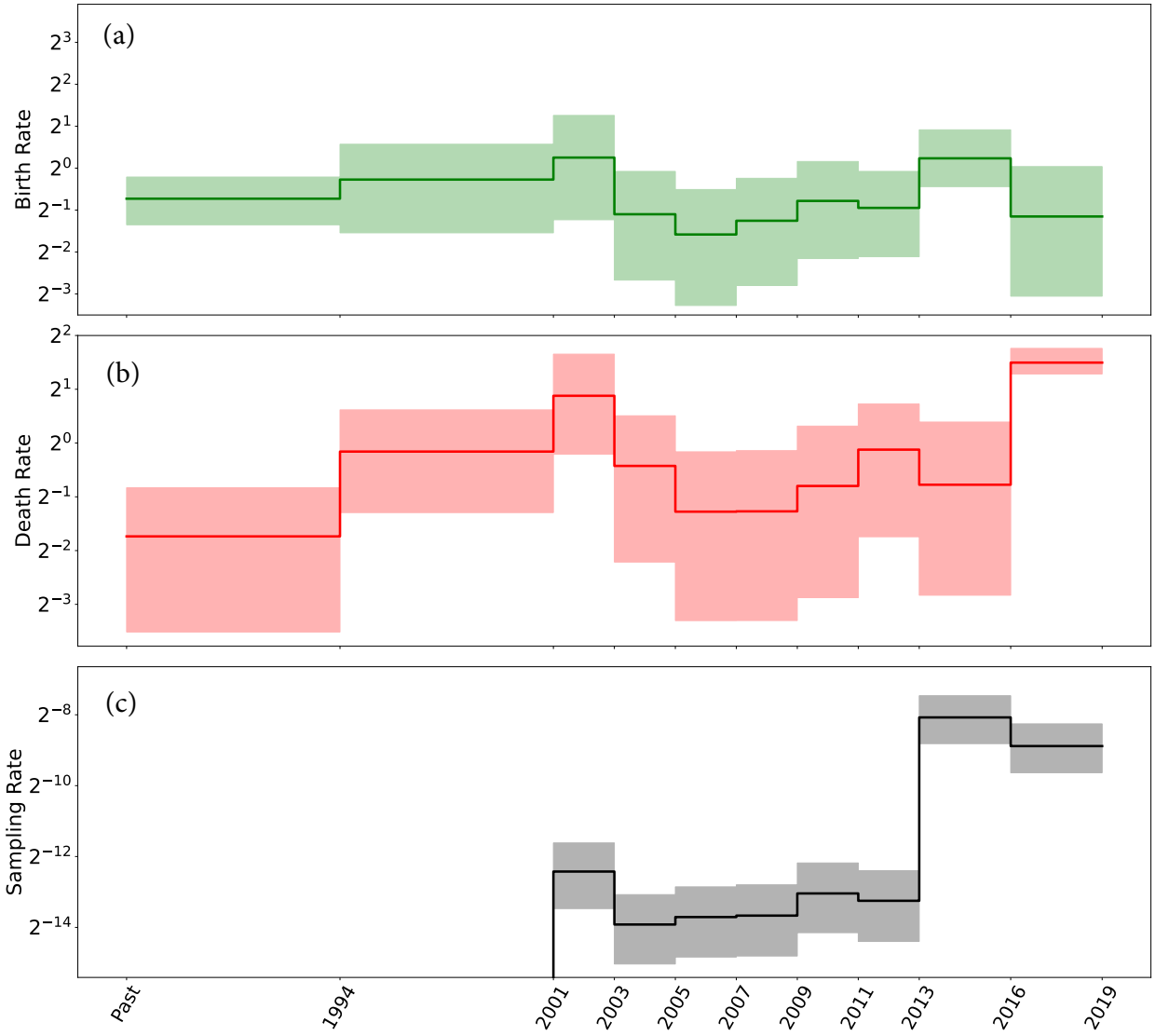

Figure S1: HIV virus: Median (solid line) and 95% credible intervals indicated by the shaded areas of the (a) birth rate, (b) death rate, and (c) sampling rate estimates through time.

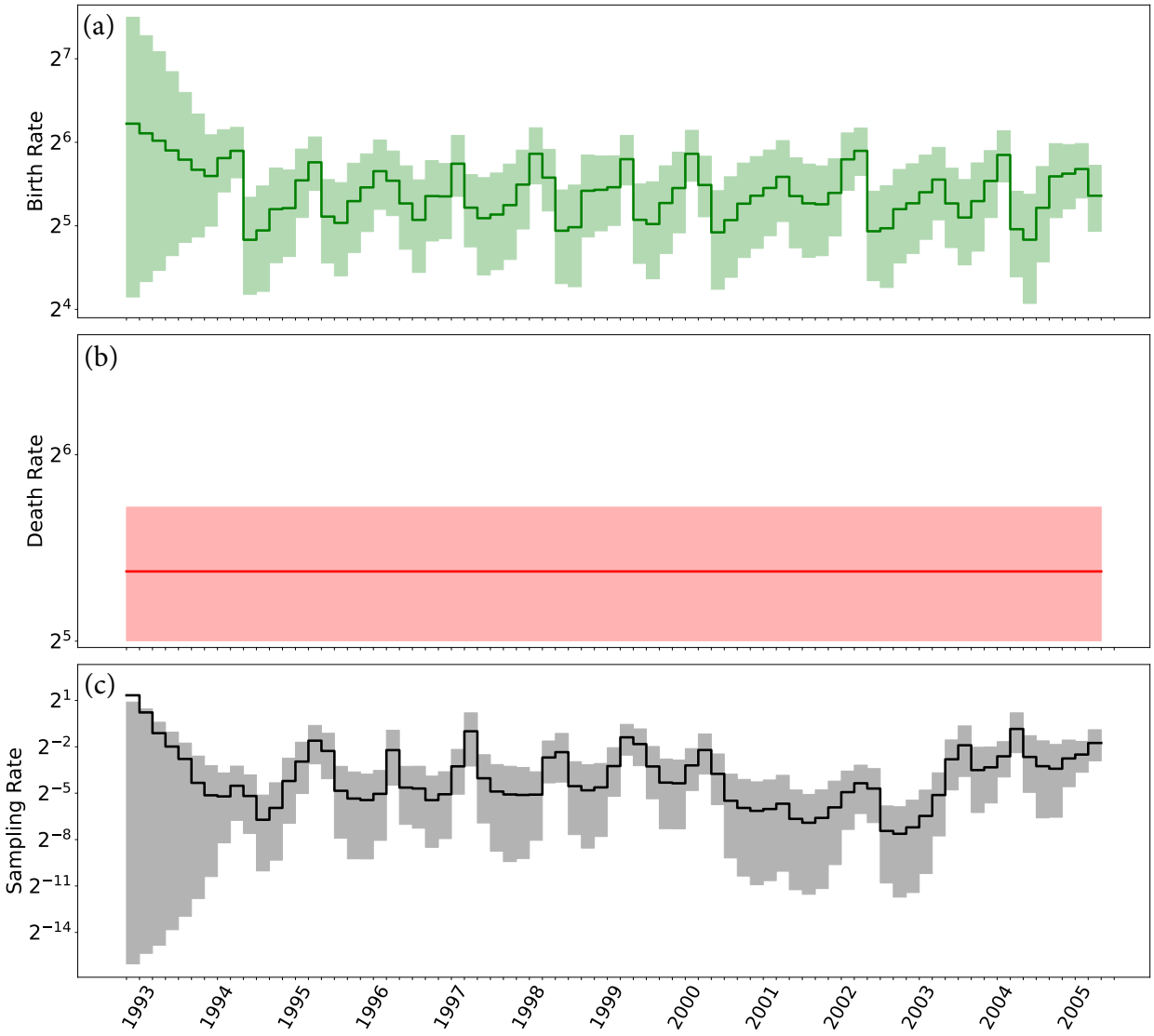

Figure S2: Influenza virus: Median (solid line) and 95% credible intervals indicated by the shaded areas of the (a) birth rate, (b) death rate, and (c) sampling rate estimates through time.

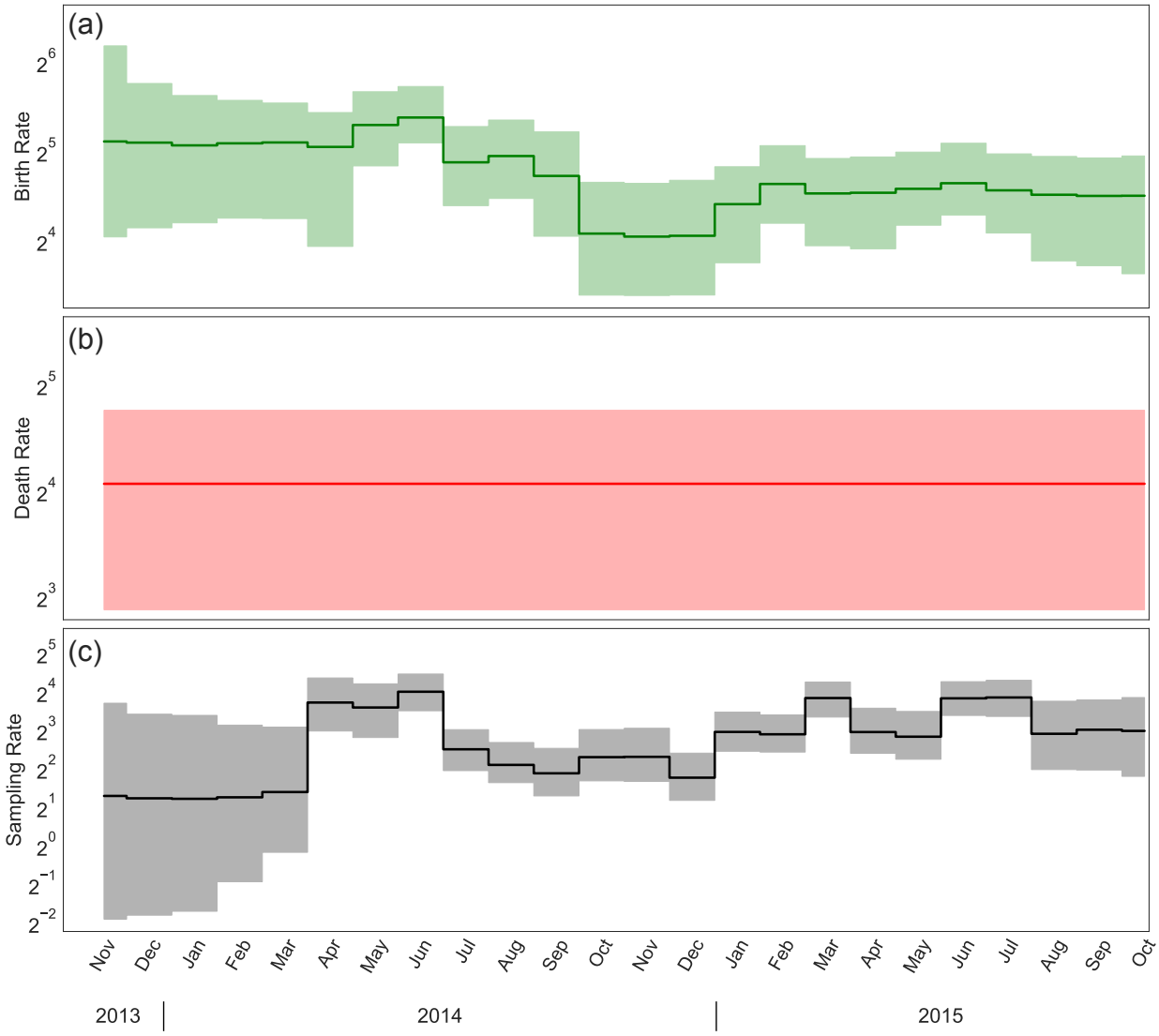

Figure S3: Ebola virus: Median (solid line) and 95% credible intervals indicated by the shaded areas of the (a) birth rate, (b) death rate, and (c) sampling rate estimates through time.

S5 Computational complexity of the nodewise likelihood

The computational complexity of evaluating node-based representations of the likelihood is much less explicit. First, we need to write out an equivalent expression for the likelihood of Equation 1 node-wise. It will be helpful to distinguish different types of samples. In particular, let us denote serially-sampled tips $\bar{\boldsymbol{u}}_\psi$ with a particular serially-sampled tip being $\bar{u}_{\psi i}$. With a slight abuse of notation, let us denote intensively-sampled tips $\bar{\boldsymbol{u}}_\rho$, with $\bar{\boldsymbol{u}}_{\rho i}$ denoting the *vector* of intensively-sampled tips at the ith intensive-sampling event. Then we can write

$$\begin{aligned}\mathbb{P}[\mathcal{T} \mid \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\psi}, \boldsymbol{\rho}, \boldsymbol{r}, \boldsymbol{t}] = \log(q_{k(t_{or})}(t_{or})) + \left(\sum_{i=1}^{||\boldsymbol{v}||} \log(\lambda_{k(v_i)}) + \log(q_{k(v_i)}(v_i))\right) + \\ \left(\sum_{i=1}^{||\bar{\boldsymbol{u}}_\psi||} \log(\psi_{k(\bar{u}_{\psi i})}) + \log(r_{k(\bar{u}_{\psi i})} + (1 - r_{k(\bar{u}_{\psi i})})p_{k(\bar{u}_{\psi i})})) - \log(q_{k(\bar{u}_{\psi i})}(\bar{u}_{\psi i}))\right) + \\ \left(\sum_{i=1}^{K} ||\bar{\boldsymbol{u}}_{\rho i}|| + \log(\rho_i) + (L(t_{i-1}) - ||\bar{\boldsymbol{u}}_{\rho i}||)\log((1-\rho_i)q_{i-1}(t_{i-1})) + \right. \\ \left. + ||\tilde{\boldsymbol{u}}_{\rho i}||\log(1-r_i)q_{i-1}(t_{i-1})||\bar{\boldsymbol{u}}_{\rho i}||\log(r_i + (1-r_i)p_{i-1}(t_{i-1}))\right)\end{aligned} \tag{36}$$

The complexity here is not immediately apparent for a number of reasons. For one, the complexity appears to depend on the relative proportion of samples of different types, which affects the number of values of $p_k(t)$ and $q_k(t)$ which must be computed. Importantly, the complexity of computing those $p_k(t)$ and $q_k(t)$ is not immediately apparent either, and that these costs are somewhat hard to disentangle, as $p_k(t_i)$ builds recursively on $p_{k-1}(t_i)$ and $q_k(t)$ depends on $p_k(t)$.

S5.1 Node lookups

Regardless of such ambiguities, all nodes in the tree require an interval lookup. For births, the lookup is required to find the correct λ_k term to use. For samples, the lookup is either to find the appropriate sampling rate, for serial samples, or to determine to which intensive-sampling event a sample belongs, for intensive samples. The time requirement here depends on the algorithm, for a binary search it is $\mathcal{O}(\log(K))$, making the total lookup cost $\mathcal{O}(N \log(K))$.

S5.2 How many computations of $q_k(t)$ are required?

In the worst, but most common, case, there are no intensive-sampling events and $q_k(t)$ must be computed for the times of all samples, all births, and all epoch times (note that even when ρ_i is 0, there is a term $L(t_i) \log(q_{i-1}(t_i))$ which must be computed in the final summation). In the best case, all samples are at intensive-sampling events, and $q_k(t)$ only needs to be computed for the times of all births and all epoch times. These are both $\mathcal{O}(N + K)$, though there is a factor of two's worth of variation in front of the N depending on which side of this spectrum a tree falls in. Calling the cost of computing $q_k(t)$ Q, this makes the contribution to the complexity here $\mathcal{O}(Q(N + K))$.

S5.3 How many computations of $p_k(t)$ are required?

The likelihood contains a number of explicit computations of $p_k(t)$ in the terms pertaining to (both serially- and intensively-)sampled tips. When all samples are serial samples, there are $\mathcal{O}(N)$ direct computations of $p_k(t)$, while when all samples are intensive samples, there are $\mathcal{O}(K)$. Taking the cost of computing $p_k(t)$ to be P, the addition to the cost here is between $\mathcal{O}(PN)$ and $\mathcal{O}(PK)$.

S5.4 What is the cost of computing $p_k(t)$ and $q_k(t)$?

We have thus far shown that the cost of computing the nodewise likelihood appears to be between $\mathcal{O}(N\log(K)+Q(N+K)+PN)$ and $\mathcal{O}(N\log(K)+Q(N+K)+PK)$. But this is not particularly revealing without considering P and Q.

While $q_k(t)$ depends on $p_{l:l<k}(t)$ through $\mathbf{A}$ and $\mathbf{B}$, once A_k and B_k have been computed, let us assume (as we did when evaluating the cost of the interval-wise likelihood) that the cost of $q_k(t)$ is $\mathcal{O}(1)$. In other words, let us assume that $\mathcal{O}(Q(N+K))=\mathcal{O}(P(N+K))$. This makes the implied cost of the nodewise likelihood between $\mathcal{O}(N\log(K)+P(N+K)+PN)$ and $\mathcal{O}(N\log(K)+P(N+K)+PK)$, which both simplify to $\mathcal{O}(N\log(K)+P(N+K))$. Naïvely, we might choose to compute $p_k(t)$ recursively every time we need it, which is $\mathcal{O}(K^2)$. In this case, the implied cost of the nodewise likelihood is $\mathcal{O}(N\log(K)+NK+K^2))$.

S5.5 Precomputing $\mathbf{A}$ and $\mathbf{B}$

One can instead choose to pre-compute A_k, B_k, as once these are computed the cost to compute $p_k(t)$ and $q_k(t)$ becomes $\mathcal{O}(1)$. Working backwards from the present allows re-computation to be avoided. As we did when we approximated the cost of the interval-wise likelihood, we will take the cost of the update (computing (A_k,B_k) from (A_{k-1},B_{k-1})) to be $\mathcal{O}(1)$. Thus, the cost of the precomputation is $\mathcal{O}(K)$. This puts the implied cost of computing the nodewise likelihood between $\mathcal{O}(N\log(K)+N+K)$.

S5.6 Counting lineages at epoch times

Regardless of whether the model includes intensive-sampling (that is, regardless of whether $\boldsymbol{\rho}=0$), one must compute $L(t_i)$ for all epoch times. This can be solved essentially the same way as the subintervals are obtained, at a cost of $\mathcal{O}(N+N\log(N))$. Alternately, it can be obtained by counting the number of births and sampled tips older (or younger) than each epoch time, at a cost of $\mathcal{O}(KN)$. This makes the lower end of the computational cost once

again a range, from $\mathcal{O}(NK + N\log(K) + N + K)$ to $\mathcal{O}(N\log(K) + N\log(N) + N + K)$.

In practice, the constants in front of all the sorting and node-lookup terms appear to be so small as to be unnoticeable in real-world computation. We demonstrate this in our timing experiments in the next section. Thus, for all practical purposes, the likelihood appears to be $\mathcal{O}(N + K)$ regardless of representation, as long as one avoids recursive computation of $p_k(t)$.

S6 Timing Experiments

With the reformulation of the likelihood and derivation of the analytical gradients, our method notably gains in speed, as we highlight in this section. For a comprehensive assessment, we compare our approach with four other specialized packages for EBDS model inference concerning likelihood calculations. These include the BDSKY (Stadler et al. 2013) package within BEAST2 (Bouckaert et al. 2019), TreePar (Stadler et al. 2013) package in R (R Core Team 2021) and RevBayes (Höhna et al. 2016). Furthermore, we present a benchmark comparing the gradient calculation efficiency of automatic differentiation implemented in VBSKY (Ki & Terhorst 2022)package using JAX library (Bradbury et al. 2018) isolated from the variational inference procedure against our algorithm based analytical gradients implemented in BEAST.

To assess the scalability of the aforementioned methods in terms of likelihood/gradient calculation, we simulated a set of trees under the EBDS model with increasing number of tips. To investigate the scalability of different methods wrt the number of sequences, we fix the number of epochs to 5 for both likelihood and gradient calculation.

Regarding scalability with respect to the number of epochs, we adjust the model by progressively increasing the number of epochs. To keep other variables constant, we maintain the tree topology and set the number of tips at 12 (in scenarios where $K >> N$, this allows us to negate the effect of N in $\mathcal{O}(N+K)$) for likelihood computation. For gradient calculations, we set the number of tips to 8198 (to minimize the impact of K^2 in $\mathcal{O}(NK + K^2)$).

For methods that employ just-in-time (JIT) compilation, including BEAST, BEAST2 and VBSKY, we run a short MCMC chain or variational inference algorithm to compute likelihood or gradient across 100,000 iterations and take the average run time.

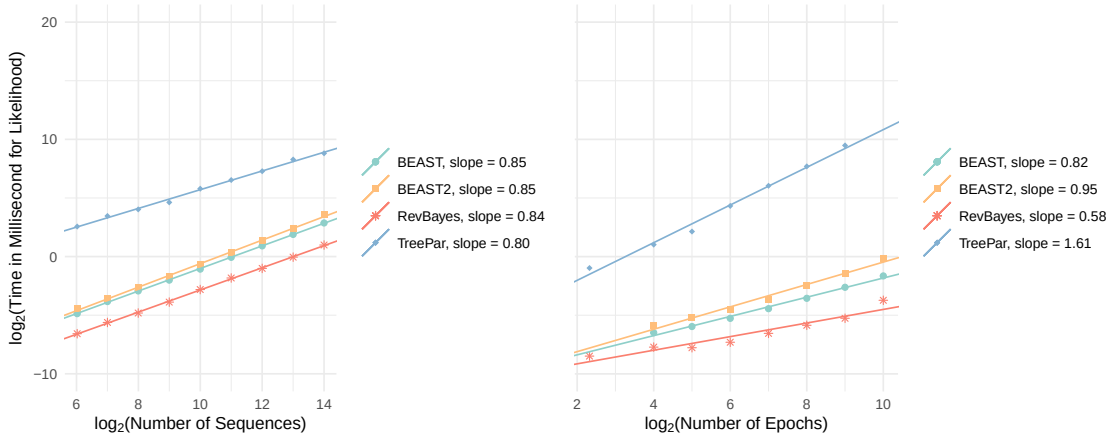

Figure S4: Speed of implementations for the likelihood calculations of increasing number of sequences (left plot) or number of epochs (right plot) for EBDS model. Note the time and number of sequences/epochs are laid out according to a logarithmic scale with base 2.

In our analysis, we observe that for likelihood computations, the implementations in BEAST, BEAST2, and RevBayes offer similar speed performance when adjusting both the number of sequences and epochs. In contrast, the TreePar package consistently lags, being several hundred times slower than its counterparts across all tested scenarios. It is also the sole implementation that exhibits a quadratic scaling with the number of epochs. The algorithms of BEAST, BEAST2, and RevBayes seem to demonstrate approximately linear scaling relative to both tree size and model epochs. It's worth noting that RevBayes delivers the quickest calculation speed, which might be attributed to the inherent speed advantages of precompiled codes, particularly for quick likelihood calculations in our context. Result for TreePar with epochs exceeding than 512 is not not included as TreePar fail to process such large models.

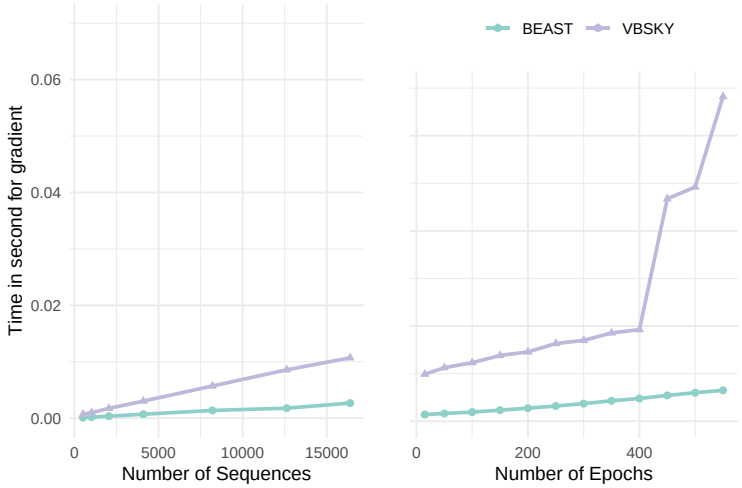

Figure S5: Speed of implementations for and gradient calculations of increasing number of sequences (left plot) or number of epochs (right plot) for EBDS model.

In terms of gradient calculations, our analytical gradients deployed within BEAST is remarkably faster than VBSKY approach using automatic differentiation. The gradient computation scales approximately linearly with the number of sequences for both BEAST and VBSKY. However, wrt the number of epochs, the scaling remains linear for BEAST but seems quadratic for VBSKY. We further confirm that the runtime slowness exhibited in VBSKY is not due to memory issues or JIT compilation difficulty. Therefore, our analysis demonstrates that analytically calculating the gradients of the EBDS likelihood is critical for improving the running time of gradient based methods.