# Deep learning and likelihood approaches for viral phylogeography converge on the same answers whether the inference model is right or wrong

Ammon Thompson[1,*,†], Benjamin Liebeskind[2,†], Erik J. Scully[2,†], Michael Landis[3,*]

[1]*Participant in an education program sponsored by U.S. Department of Defense (DOD) at the National Geospatial-Intelligence Agency, Springfield, VA, 22150, USA*

[2]*National Geospatial-Intelligence Agency, Springfield, VA, 22150, USA*

[3]*Department of Biology, Washington University in St. Louis, Rebstock Hall, St. Louis, Missouri, 63130, USA*

† The views presented here are those of the authors and do not necessarily represent the views of DoD or its Components.

[*]**Corresponding authors:** E-mail: Ammon.M.Thompson.ctr@nga.mil and michael.landis@wustl.edu.

*Abstract.*— Analysis of phylogenetic trees has become an essential tool in epidemiology. Likelihood-based methods fit models to phylogenies to draw inferences about the phylodynamics and history of viral transmission. However, these methods are computationally expensive, which limits the complexity and realism of phylodynamic models and makes them ill-suited for informing policy decisions in real-time during rapidly developing outbreaks. Likelihood-free methods using deep learning are pushing the boundaries of inference beyond these constraints. In this paper, we extend, compare and contrast a recently developed deep learning method for likelihood-free inference from trees. We trained multiple deep neural networks using phylogenies from simulated outbreaks that spread among five locations and found they achieve close to the same levels of accuracy as Bayesian inference under the true simulation model. We compared robustness to model misspecification of a trained neural network to that of a Bayesian method. We found that both models had comparable performance, converging on similar biases. We also implemented a method of uncertainty quantification called conformalized quantile regression which we demonstrate has similar patterns of sensitivity to model misspecification as Bayesian highest posterior intervals (HPI) and greatly overlap with HPIs, but have lower precision (more conservative). Finally, we trained and tested a neural network against phylogeographic data from a recent study of the SARS-Cov-2 pandemic in Europe and obtained similar estimates of region-specific epidemiological parameters and the location of the common ancestor in Europe. Along with being as accurate and robust as likelihood-based methods, our trained neural networks are on average over 3 orders of magnitude faster. Our results support the notion that neural networks can be trained with simulated data to accurately mimic the good and bad statistical properties of the likelihood functions of generative phylogenetic models.

(Keywords: phylogeography, SSE, phylodynamics, machine learning, deep learning, epidemiology)

# Introduction

Viral phylodynamic models use genomes sampled from infected individuals to infer the evolutionary history of a pathogen and its spread through a population (Holmes and Garnett 1994; Volz et al. 2013). By linking genetic information to epidemiological data, such as the location and time of sampling, these generative models can provide valuable insights into the transmission dynamics of infectious diseases, especially in the early stages of cryptic disease spread when it is more difficult to detect and track (Holmes et al. 1995; Rambaut et al. 2008; Lemey et al. 2009; Pybus et al. 2012; Worobey et al. 2016, 2020; Lemey et al. 2021; Washington et al. 2021; Pekar et al. 2022). This information can be used to inform public health interventions and improve our understanding of the evolution and spread of pathogens. Many phylodynamic models are adapted from state-dependent birth-death (SDBD) processes or, equivalently, state-dependent speciation-extinction (SSE) models (Maddison et al. 2007; FitzJohn 2012; Kühnert et al. 2014; Beaulieu and O'Meara 2016). These birth-death models correspond to the well-known Susceptible-Infectious-Recovered (SIR) model during an exponential growth phase, when nearly all individuals in the population are susceptible to infection (Anderson and May 1979). The simplest SIR models only track the number of susceptible, infected, and recovered individuals across populations over time, with more advanced models also allowing the movement of individuals among localized populations. The phylodynamic models we are interested in track the incomplete transmission tree (phylogeny) of sampled, infected individuals that emerges from host-to-host pathogen spread among populations over space and time. Within this broader context, we will refer to the state as location and the models as location-dependent birth-death (LDBDS) models that include serial sampling of taxa (Kühnert et al. 2016).

Analysts typically fit these birth-death models to data using likelihood-based inference methods, such as maximum likelihood (Maddison et al. 2007; Richter et al. 2020) or Bayesian inference (Kühnert et al. 2016; Scire et al. 2020). Likelihood-based inference

relies upon a likelihood function to evaluate the relative probability (likelihood) that a given phylogenetic pattern (i.e., topology, branch lengths, and tip locations) was generated by a phylodynamic process with particular model parameter values. In this sense the likelihood of any possible phylodynamic data set is mathematically encoded into the likelihood as a function of (unknown) data-generating model parameters.

Computing the likelihood requires high-dimensional integration over a large and complex space of evolutionary histories. Analytically integrated likelihood functions, however, are not known for LDBDS models. Methods developers instead use ordinary differential equation (ODE) solvers (Maddison et al. 2007; Kühnert et al. 2016) to numerically approximate the integrated likelihood. These clever approximations perform well statistically, but are too computationally expensive to use with large epidemic-scale data sets. Thus, while Nextstrain (Hadfield et al. 2018) and similar efforts have provided useful visualizations to policy makers during the COVID response, most phylogeographical methods are used forensically, providing insight on the past, and are not used to provide parameter estimates in response to emerging events to inform policy decisions in real-time due to the complexity and long run-times of these models.

As phylodynamic models become more biologically realistic, they will necessarily grow more mathematically complex, and therefore less able to yield likelihood functions that can be approximated using ODE methods. Because of this, phylodynamic model developers tend to explore only models for which a likelihood-based inference strategy is readily available. As a consequence, the lack of scalable inference methods impedes the design, study, and application of richer phylodynamic models of disease transmission, in particular, and richer phylogenetic models of lineage diversification, in general.

To avoid the computational limitations associated with likelihood-based methods, deep learning inference methods that are likelihood-free have emerged as a complementary framework for fitting a wide variety of evolutionary models (Bokma 2006). Deep learning methods rely on training many-layered neural networks to extract information from data

patterns. These neural networks can be trained with simulated data as another way to approximate the latent likelihood function (Cranmer et al. 2020). Once trained, neural networks have the benefit of being fast, easy to use, and scalable. Recently, likelihood-free deep learning neural network methods have successfully been applied to phylogenetics (da Fonseca et al. 2020; Suvorov et al. 2020; Nesterenko et al. 2022; Solis-Lemus et al. 2022; Suvorov and Schrider 2022) and phylodynamic inference (Lambert et al. 2022; Voznica et al. 2022).

Here we extend new methods of deep learning from phylogenetic trees (Lambert et al. 2022; Voznica et al. 2022) to explore their potential when applied to phylogeographic problems in geospatial epidemiology. Phylodynamics of birth-death-sampling processes that include migration among locations have been under development for more than a decade (Stadler 2010; Stadler et al. 2012; Kühnert et al. 2014, 2016; Scire et al. 2020; Gao et al. 2022, 2023). Given the added complexity of location-specific dynamics (e.g. location-specific infection rates) and recent successes in deep learning with phylogenetic time trees (Voznica et al. 2022) under state-dependent diversification models (Lambert et al. 2022), we sought to evaluate this approach when applied to viral phylodynamics and phylogeography by including location data when training deep neural networks with phylogenetic trees.

A current limitation of likelihood-free approaches is that it remains unknown how brittle the inference machinery is when the assumptions used for simulation and training are violated (Schmitt et al. 2022). For example, a brittle deep learning method would be more easily misled by model misspecification when compared to a likelihood-based method. Likelihood approaches may have some advantages because the simplifying assumptions are explicit in the likelihood function while for trained neural networks it is difficult to know how those same assumptions implemented in the simulation are encoded in data patterns in the training data and learned network weights. However, with complex likelihood models, there may be unexpected interactions among simplifying assumptions that can result in

large biases when applied to real-world data (Gao et al. 2023). Characterizing the relative robustness and brittleness of these two inference paradigms is essential for those who wish to confidently develop and deploy likelihood-free methods of inference from real world data.

To explore relative robustness to model misspecification, we trained multiple deep convolutional neural networks (CNNs) with transmission trees generated from epidemic simulations. We were able to achieve accuracy very close to that of a likelihood-based approach and through several model misspecification experiments show that our CNNs are no more sensitive to model violations than the likelihood approach. Significantly, both methods consistently show similar biases induced by model violations in test data sets. We find that for the models tested here, the migration rate estimates are highly sensitive to misspecification of infection rate and sampling rates, but that estimates of the infection and sampling rates are fairly robust to misspecification of the migration models. We also show that the rate parameter estimates are fairly robust to misspecification of both the number of locations in the model and phylogenetic error. We also estimated prediction intervals for the rate parameters and compared and contrasted their performance to the Bayesian highest posterior density intervals (HPI). We show that they produce intervals that greatly overlap with HPIs in all experiments, but have, on average, wider intervals making them relatively conservative. Finally, we compared a simulation-trained neural network to a recent phylodynamic study of the first wave of the COVID pandemic in Europe (Nadeau et al. 2021) and obtain similar inferences about the dynamics and history of SARS-CoV-2 in the European clade.

# Methods

First, we define the SIR model we assume here that is approximately equivalent to the LDBDS model (Kühnert et al. 2016). Following that, is a description of the simulation method to generate the training, validation, and test data sets of phylogenies under the model. The simulation and data processing pipeline is shown in Figure 1. We next describe

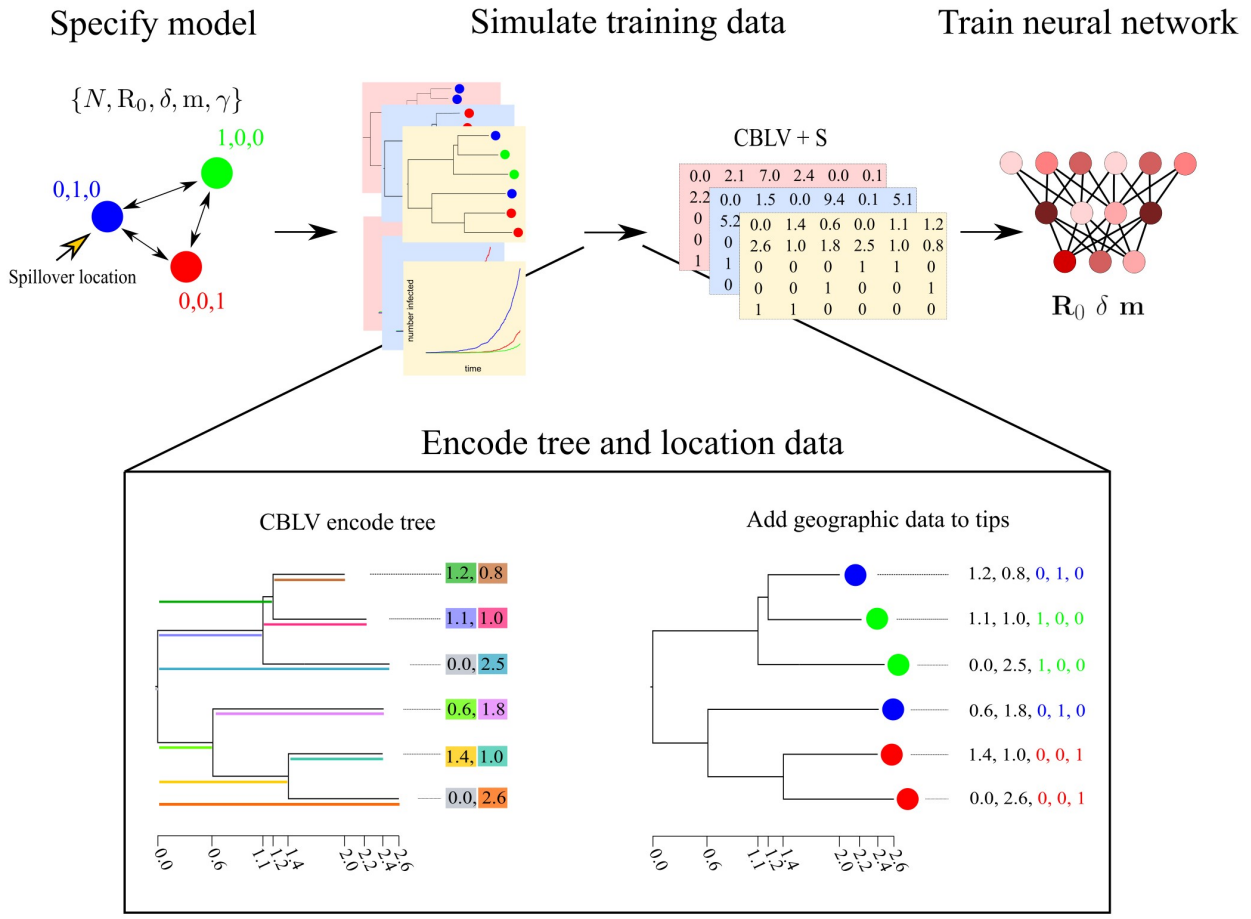


Figure 1: Simulation and tree encoding pipeline for generating training data. 1) Specify a model, for example an SIR model with serial sampling and migration among three locations (colored circles). 2) Run simulations of outbreaks under the model to generate population trajectories and phylogenetic trees. 3) Encode trees and location data into the Compact Bijective Ladderized Vector + States (CBLV+S) format. 4) Train the neural network with CBLV+S training data.

our implementation of simulation-trained deep learning inference with convolutional neural networks (CNN) as well as a likelihood-based method using Bayesian inference. We then describe our methods for measuring and comparing their performance when tested against data sets generated by simulations under the inference model as well as several data sets simulated under models that violate assumptions of the inference model. Finally, we describe how we tested our simulation-trained CNN against a real-world data set.

## *Model definition*

We first define a general location-dependent SIR stochastic process used for simulations and likelihood function derivation in the format of reaction equations we specified in MASTER (Vaughan and Drummond 2013). Reaction equations 1 through 4 specify the SIR compartment model with migration and serial sampling where $S$, $I$, and $R$ denote the number of individuals in each compartment. The $S$ and $I$ compartments are indexed by geographic location using $i$ and $j$. $N_i$ is the total population size in location $i$ and $N_i = S_i + I_i + R_i$. To simplify notation, we consider all local recoveries to lead to the same global compartment and absorbing state, $R$. The symbols for each rate parameter is placed above each reaction arrow.

$$S_i + I_i \xrightarrow{\beta_i/N_i} 2I_i \qquad \text{infection} \tag{1}$$

$$I_i \xrightarrow{m_{ij}} I_j \qquad \text{migration} \tag{2}$$

$$I_i \xrightarrow{\gamma} R \qquad \text{recovery} \tag{3}$$

$$I_i \xrightarrow{\delta_i} R \qquad \text{sample and recovery.} \tag{4}$$

We parameterize the model with the basic reproduction number in location $i$, $R_{0_i}$, which is related to $\beta_i$ and $\delta_i$ by equation 5,

$$R_{0_i} = \frac{\beta_i}{\gamma + \delta_i}. \tag{5}$$

In particular, our study considers a location-independent SIR (LISIR) model with sampling that assumes $R_{0_i}$ was equal among all locations, and a location-dependent (LDSIR) model with sampling that assumes $R_{0_i}$ varied among locations. During the exponential growth phase of an outbreak, the LISIR and LDSIR models are equivalent to the location-independent birth-death-sampling (LIBDS) and location-dependent birth-death-sampling (LDBDS) models, respectively, that are often used in viral phylogeography (Kühnert et al. 2014, 2016; Douglas et al. 2021).

Each infectious individual transitions to recovered at rate $\gamma$. We assumed that sampling a virus in an individual occurs at rate $\delta_i$ in location $i$ and immediately removes that individual from the infectious compartment and places them in the recovered compartment. Thus the effective recovery rate in location $i$ is $\gamma + \delta_i$. The above reactions correspond to the following coupled ordinary differential equations.

$$\begin{aligned}
\frac{dS_i}{dt} &= -\frac{\beta_i}{N_i} S_i I_i \\
\frac{dI_i}{dt} &= \frac{\beta_i}{N_i} S_i I_i + \sum_{j \neq i}^{n} m_{ij} I_j - \sum_{j \neq i}^{n} m_{ji} I_i - (\gamma + \delta_i) I_i \\
\frac{dR}{dt} &= \sum_{i=1}^{n} (\gamma + \delta_i) I_i
\end{aligned} \tag{6}$$

When the migration rate is constant among locations and the model is a location-independent SIR model, or equivalently, LIBDS, and we set $S_i(t = 0) \approx N_i$ at the beginning of the outbreak, the equation set 6 reduces to

$$\frac{dS_i}{dt} = -\beta I_i$$
$$\frac{dI_i}{dt} = \beta I_i + m\left(\sum_{j\neq i}^{n} I_j - (n-1)I_i\right) - (\gamma+\delta)I_i$$
$$\frac{dR}{dt} = (\gamma+\delta)\sum_{i=1}^{n} I_i$$

The number of infections and the migration of susceptible individuals is at negligible levels on the timescales investigated here. The infection rate is, therefore, approximately constant and the migration of susceptible individuals can be safely ignored requiring only migration of infectious individuals to be simulated.

At the beginning of an outbreak, it is often easier to know the recovery period from clinical data than the sampling rate which requires knowing the prevalence of the disease. Therefore, we treat the average recovery period as a known quantity and use it to make the other two parameters (the sampling rate and the basic reproduction number $R_0$) identifiable. This was done by fixing the corresponding rate parameter in the likelihood function to the true simulated value for each tree, and by adding the true simulated value to the training data for training the neural network.

## *Simulated training and validation data sets*

Epidemic simulations of the SIR+migration model that approximates the LIBDS process were performed using the MASTER package v. 6.1.2 (Vaughan et al. 2014) in BEAST 2 v. 2.6.6 (Bouckaert et al. 2019). MASTER allows users to simulate phylodynamic data sets under user-specified epidemiological scenarios, for which MASTER simultaneously simulates the evolution of compartment (population type) sizes and tracks the branching lineages (transmission trees in the case of viruses) from which it samples over time. We

trained neural networks with these simulated data to learn about latent populations from the shape of sampled and subsampled phylogenies. In addition to the serial sampling process, at the end of the simulation 1% of infected lineages were sampled. In MASTER this was approximated by setting a very high sampling rate and very short sampling time such that the expected number sampled was approximately 1%. This final sampling event was required to make a 1-to-1 comparison of the likelihood function used for this study (see Likelihood method description below) which assumes at least one extant individual was sampled to end the process. Coverage statistics from our MCMC samples closely match expectations (see Likelihood method description below; SI Figure 2 C). Simulation parameters under LIBDS and LDBDS models for training the neural network under the phylogeography model were drawn from the following distributions:

$$\begin{aligned} R_0 &\sim \text{Uniform}(2,\ 8) \\ \delta &\sim \text{Uniform}(0.0001,\ 0.005) \\ m &\sim \text{Uniform}(0.0001,\ 0.005) \\ \gamma &\sim \text{Uniform}(0.01,\ 0.05) \\ \text{spillover location} &\sim \text{Multinomial}(\text{k} = 1,\ \text{p}_i = 1/5),\ \text{for 5 locations} \end{aligned} \tag{7}$$

All five locations had initial population sizes of 1,000,000 susceptible individuals and one infected individual in a randomly sampled spillover location. Simulations were run for 100 time units or until 50,000 individuals had been infected to restrict simulations to the approximate exponential phase of the outbreak. For the experiments comparing the CNN to the likelihood-based method under the LIBDS model, if this population threshold was reached the simulation was rejected. This criterion was not enforced for simulations under the LDBDS model. This ensured the LIBDS model used in the likelihood-based analyses are equivalent to more complex density-dependent SIR models. After simulation, trees with

500 or more tips were uniformly and randomly downsampled to 499 tips and the sampling proportion was recorded for training the neural networks and to adjust estimates of $\delta$.

We simulated 410,000 outbreaks under these LIBDS settings to generate the training, validation, and test sets for deep learning. Any simulation that generated a tree with less than 20 tips was discarded, leaving a total of 111,157 simulated epidemiological data sets. Of these, 104,157 data sets were used to train and 7,000 were used to validate and test each CNN. A total of 193,110 LDBDS data sets were simulated, with 186,110 used to train and 7,000 used to validate and test the LDBDS CNNs.

To make phylodynamic inferences about the first wave of the SARS-CoV-2 epidemic in Europe we used the LDBDS model on the data set from Nadeau et al. (2021). Training simulation parameters for the LDBDS process were drawn from the same distributions as LIBDS except $R_0$ which was unique for each location. We assume that the variability of $R_0$ among different pathogens (simulated outbreaks) is greater than the variability of the same pathogen's $R_0$ among different locations within the same simulation. To implement this assumption, all $R_0$ was drawn from a joint distribution to narrow the magnitude of differences among locations within simulations to be within 6 of each other but expand the magnitude of differences between simulations to range from 0.9 to 15:

$$\alpha \sim \text{Uniform}(3.9,\ 12)$$

$$R_{0_i} \mid \alpha \sim \text{Uniform}(\alpha - 3,\ \alpha + 3)$$

For the empirical analysis, population sizes at each location were also set to 500,000 and instead of running the simulations for 100 time units, time was scaled by the recovery period, $1/\gamma$, and was drawn from a uniform distribution:

$$\text{time} \sim \text{Uniform}(1,\ 20)$$

## *Simulated test data sets with and without model misspecification*

All simulation models used for training and testing are listed in Table 1. We first simulated a test set of 138 trees under the training model to compare the accuracy of the CNN and the likelihood-based estimates when the true model is specified. These data sets were simulated by random draws of parameter values from the same distributions described above for generating the training data set.

Sensitivity to model misspecfication for each of the three rate parameters, $R_0$, $\delta$, and $m$, was tested. All sensitivity experiments used the same LIBDS model for inference for both the CNN and the Likelihood-based methods. Sensitivity experiments were conducted by simulating a test data set of trees that were generated by an epidemic process that was more complex than or different from the LIBDS model.

The tree data set for the misspecified $R_0$ experiment consisted of simulating outbreaks where each location had a unique $R_0$ drawn from the same distribution as above. Likewise, the misspecified sampling model test set was generated by simulating outbreaks where each location had a unique sampling rate, $\delta$, drawn from the same distribution used for the global sampling rate described above. For the misspecified migration model, a random pair of coordinates, each drawn from a uniform(0,5) distribution in a plane, were generated for the five locations, and a pairwise migration rate was computed such that pairwise migration rates were symmetric and proportional to the inverse of their euclidean distances and the average pairwise migration rate was equal to a random scalar which was also drawn from a uniform distribution (see equations 7 above).

The tree set for the misspecified number of locations experiment was generated by

| Description | Simulation model parameters and data |
|---|---|
| Generate training data | $\{N, R_0, \delta, m, \gamma, \Psi\}$ |
| Misspecify $R_0$ | $\{N, R_{0_1}, R_{0_2}, R_{0_3}, R_{0_4}, R_{0_5}, \delta, m, \gamma, \Psi\}$ |
| Misspecify $\delta$ | $\{N, R_0, \delta_1, \delta_2, \delta_3, \delta_4, \delta_5, m, \gamma, \Psi\}$ |
| Misspecify $m$ | $\{N, R_0, \delta, m_{ij} \forall i \neq j \in \{1, \ldots, N\}, \gamma, \Psi\}$ |
| Misspecify number of locations | $\{2N, R_0, \delta, m, \gamma, \Psi\}$ |
| Tree error | $\{N, R_0, \delta, m, \gamma, \Psi^{\text{error}}\}$ |
| Analyze Nadeau et al. (2021) dataset | $\{N, R_{0_1}, R_{0_2}, R_{0_3}, R_{0_4}, R_{0_5}, \delta, m, \gamma, \Psi\}$ |

Table 1: Models used in this study. All simulations assume an SIR compartmental epidemic model. N = 5 is the number of locations, $R_0$ is the basic reproduction number, $\delta$ is the sampling rate, m is the migration rate, $\gamma$ is the recovery rate (treated as data), and $\Psi$ is the phylogenetic tree + locations (also treated as data).

simulating outbreaks among ten locations instead of five. After simulations, six locations were chosen at random and re-coded as being sampled from the same location.

To generate a test set where the time tree used for inference has incorrect topology and branch lengths, we implemented a basic pipeline of tree inference from simulated genetic data to mimic a worst case real world scenario. We simulated trees under the same settings as before. Phylogenetic error was introduced in two ways: the amount of site data (short sequences) and misspecification of the DNA sequence evolution inference model using seq-gen V. 1.3.2 (Rambaut and Grassly 1997). We simulated the evolution of a 200 base-pair sequence under an HKY model with $\kappa = 2$, equal base frequencies and 4 discretized-gamma(2, 2) rate categories for among site rate variation. The simulated alignment as well as the true tip dates (sampling times) was then used to infer test trees. Test tree inference was done using IQ-Tree v. 2.0.6 (Minh et al. 2020) assuming a Jukes-Cantor model of evolution where all transition rates are equal. The inference model also assumed no among-site rate variation. The number of shared branches between the true transmission tree and the test tree inferred by IQ-Tree was measured using gotree v. 0.4.2 (Lemoine and Gascuel 2021). Polytomies were resolved using phytools (Revell 2012) and a small, random number was added to each resolved branch. These trees were then used for likelihood inference and CNN prediction.

## *Deep learning inference method*

The resulting trees and location metadata generated by our pipeline were converted to a modified CBLV format (Voznica et al. 2022), which we refer to as the CBLV+S (+State of character, *e.g.* location) format (Figure 1). The CBLV format uses an in-order tree traversal to translate the topology and branch lengths of the tree into an 2 x n matrix where n is the maximum number of tips allowed for trees. The matrix is initialized with zeroes. We then fill the matrix starting with the root then proceed to the tip with largest root-to-tip distance rather than starting with that tip as in Voznica et al. (2022). We chose this to separate the the zero value of the root age from the zeroes used to pad matrices where the tree has less than the maximum number of tips, though we expect this to make marginal to no difference in performance. The CBLV representation gives each sampled tip a pair of coordinates in 'tree-traversal space'. Our CBLV+S format associates geographic information corresponding with each sampled taxon by appending each vector column with a one-hot encoding vector of length $g$ states to yield a $(2 + g) \times n$ CBLV+S matrix. The CBLV+S format allows for multiple characters and/or states to be encoded, extending the single binary character encoding format introduced by Lambert et al. (2022). Our study uses CBLV+S to encode a single character with $g = 5$ location-states. In addition to the the CBLV+S data, we also include a few tree summary statistics and known simulating parameters; the number of tips, mean branch length, the tree height and the recovery rate and the subsampling proportion. Trees were rescaled such that their mean branch length was the default for phylodeep (Voznica et al. 2022) before training and testing of the CNN. The mean pre-scaling branch length and tree heights were also fed into the neural networks. Trees were not rescaled for the likelihood-based analysis. Recall that tree height did not vary for the LIBDS CNN training set but did for the LDBDS training set (see simulation time settings above). Varying the time-scale for the LDBDS model was necessary for analyzing real world data where time-scales of outbreaks can vary considerably.

Our CNNs were implemented in Python 3.8.10 using keras v. 2.6.0 and

tensorflow-gpu v. 2.6.0. (Chollet; Abadi et al. 2016). CNNs consist of one or more layers specifically intended for structural feature extraction. CNNs utilize a filter, akin to a sliding window, that executes a mathematical operation (convolution) on the input data. When dealing with structured data like the CBLV+S matrix, multiple 1D filters slide across the matrix's columns, embedding each scanned window into an N-dimensional vector representation. This architectural design imparts CNNs with translation invariance, enabling them to recognize and learn repeating patterns throughout the input space, regardless of their specific location. Stacking multiple convolutional layers enables CNNs to decipher hierarchical structures within the data. See Alzubaidi et al. (2021) and Khan et al. (2020) for reviews of the subject.

For each model, LIBDS and LDBDS, we designed and trained two CNN architectures, one to predict epidemiological rate parameters and the other to predict the outbreak location resulting in four total CNNs trained by two training data sets (LIBDS and LDBDS). We used the mean-squared-error for the regression neural loss function in the network trained to estimate epidemiological rates, and the categorical cross-entropy loss function for the categorical network trained to estimate outbreak location. We assessed the performance of the network by randomly selecting 5,000 samples for validation before each round of training. We measured the mean absolute error and accuracy using the validation sets. We used these measures to compare architectures and determine early stopping times to avoid overfitting the model to the training data. We also added more simulations to the training set until we could no longer detect an improvement in error statistics. After comparing the performance of several networks, we found that the CNN described in SI Figure S1 performed the best. In brief, the networks have three parallel sets of sequential convolutional layers for the CBLV+S tensor and a parallel dense layer for the priors and tree statistics. The three sets of convolution layers differed by dilation rate and stride lengths. These three segments and the dense layer were concatenated and then fed into a segment consisting of a sequential set of dense layers, each layer gradually narrowing to the

output size to either three or five for the rates and origin location networks, respectively, for the LIBDS model, and seven and five for the seven rates and five locations, respectively, for the LDBDS model.

All layers of the CNN used rectified linear unit (ReLU) activation functions. We used the Adam optimizer algorithm for batch stochastic gradient descent (Kingma and Ba 2017) with batch size of 128. We selected the number of epochs by monitoring the mean absolute error and accuracy of the validation data set. This set was not used in training or testing. These metrics suggested stopping after 15 epochs for the regression network and ten epochs for the root location network would maximize accuracy/minimize error for out-of-sample test data. The output layer activation for the network that predicted the $R_0, \delta$ and m parameters was linear with three nodes. For the output layer predicting the outbreak location the activation function was softmax with five nodes for the five locations. The input layer and all intermediate (latent) layers were the same for all four networks, namely the CBLV+S tensor and the recovery rate, mean branch lengths, tree height and number of tips in the tree. The LDBDS neural network was trained with simulated trees where $R_{0_i}$ varied among locations and had an output layer with seven nodes; five for the each location's $R_{0_i}$ and a node each for the sampling rate and the migration rate. We tested networks with max-pooling layers between convolution layers as well as dropout at several rates and found no improvement or a decrease in performance.

## *Likelihood-based method of inference*

We compared the performance of our trained phylodynamic CNN to likelihood-based Bayesian phylodynamic inferences. We specified LIBDS and LDBDS Bayesian models that were identical to the LIBDS and LDBDS simulation models that we used to train our CNNs. The most general phylodynamic model in the birth-death family applied to epidemiolgoical data is the state-dependent birth-death-sampling process (SDBDS; (Kühnert et al. 2016; Scire et al. 2020)), where the state or type on which birth, death, and

sampling parameters are dependent is the location in this context. The basic model used for experiments here is a phylogeographic model that is similar to the serially sampled birth-death process (Stadler 2010) where rates do not depend on location, which we refer to as the LIBDS model. The death rate, $\mu$, is equivalent to the recovery rate, $\gamma$, in SIR models. Standard phylogenetic birth-death models assume the birth and death rates, $\lambda$ and $\mu$, are constant or time-homogeneous, while the SIR model's infection rate is proportional to $\beta$ and $S$ and varies with time as $S$ changes. However, when the number of infected is small relative to susceptible people, as in the initial stages of an outbreak, the infection rate, $\beta$, is approximately constant and approximately equal to the birth rate $\lambda$;

$$\lambda = \frac{\beta S}{N} \approx \beta \tag{8}$$

The joint prior distribution was set to the same model parameter distributions that were used to simulate the training and test sets of phylogenetic trees in the first section with $\gamma$ treated as known and the proportion of extant lineages sampled, $\rho$, set to 0.01 as in the simulations. The likelihood was conditioned on the tree having extant samples (*i.e.* the simulation ran for the allotted time without being rejected). All simulated trees in this study had a stem branch and the outbreak origins were inferred for the parent node of the stem branch.

We used Markov chain Monte Carlo (MCMC) to simulate random sampling from the posterior distribution implemented in the TensorPhylo plugin (https://bitbucket.org/mrmay/tensorphylo/src/master/) in RevBayes (Höhna et al. 2016). After a burnin phase, a single chain was run for 7,500 cycles with 4 proposals per cycle and at least 100 effective sample size (ESS) for all parameters. If the effective sample size (ESS) was less than 100, the MCMC was rerun with a higher number of cycles. We also analyzed the coverage of the 5, 10, 25, 50, 75, 90, and 95% HPI to verify that our

simulation model and inference model are the same and that the MCMC simulated draws from the true posterior distribution. Bayesian phylogeographic analysis recovered the true simulating parameters at the expected frequencies (Figure 2 C), thus validating the simulations were working as expected and confirming that the MCMC was accurately simulating draws from the true posterior distribution.

## *Quantifying errors and error differences*

We measure the absolute percent error (APE) of the predictions from the CNN and the mean posterior estimate (MPE) of the likelihood-based method. The formula for APE of a prediction/estimate, $y^{\text{estimate}}$, of $y^{\text{truth}}$ is

$$\text{APE} = \left| \frac{y^{\text{estimate}} - y^{\text{truth}}}{y^{\text{truth}}} \right| \times 100$$

The Bayesian alternative to significance testing is to analyze the posterior distribution of parameter value differences between groups. In this framework, the probability that a difference is greater than zero can be easily interpreted. We therefore used Bayesian statistics to infer the median difference in error between the CNN and likelihood-based methods and the increase in median error of each method when analyzing misspecified data compared to when analyzing data simulated under the true inference model.

We used Bayesian inference to quantify population error by performing three sets of analyses: (1) inferred the population median APE under the true model (this will be the reference group for analysis 3), (2) the effect of inference method — CNN or likelihood-based (Bayesian) — on error by inferring the median difference between the CNN estimate and the likelihood-based estimate, (3) the effect of misspecification on error for each parameter by comparing the median error of estimates under misspecified

experiments and the reference group defined by analysis 1. See SI Figures S3 - S13 and SI Table S1 for summaries and figures for all analyses for this section.

To infer these differences between groups we used the R package `BEST` (Meredith and Kruschke). BEST assumes the data follow a t-distribution parameterized by a location parameter, $\mu$, a scale parameter, $\sigma$, and a shape parameter, $\nu$, which they call the "normality parameter" (*i.e.* if $\nu$ is large the distribution is more Normal). Because the posterior distribution does not have a closed form, BEST uses Gibbs sampling to simulate draws from the posterior distribution. 20,000 samples were drawn from the posterior distribution for each BEST analysis. BEST uses automatic posterior predictive checks to indicate that a model adequately describes the data distributions. Posterior predictive checks indicate the BEST model adequately fits each data set analyzed below.

*Inferring the median APE.*— Before inferring differences between groups, we inferred the population median APE for predictions of $R_0$, $\delta$, and $m$ from test data simulated under the inference model using the CNN and likelihood-based methods. Histograms of the sampled log-transformed APE appears to be symmetric with heavy tails so we fit the log APE to the BEST model. This implies that the sampled APE scores are drawn from a log-t distribution. The log-t distribution has a mean of $\infty$ and median of $e^{\mu}$, we therefore focus our inference on estimating posterior intervals for the population median APE from the sampled APE values for each parameter estimated by the CNN method and likelihood-based method which we denote $\text{APE}^{\text{CNN}}$, and $\text{APE}^{\text{Like}}$ respectively. The data analyzed here and likelihood assumed by BEST is

$$y = \text{APE}^{\text{CNN}} \text{ or } \text{APE}^{\text{Like}}$$

$$\log y \mid \mu, \sigma, \nu \sim t_{\nu}(\mu, \sigma).$$

The priors were set to the vague priors that BEST provides by default,

$$\mu \sim \text{Normal}(\text{mean(y)}, \text{sd(y)} \times 1000)$$

$$\sigma \sim \text{Uniform}(\text{sd(y)}/1000, \text{ sd(y)} \times 1000)$$

$$\nu \sim \text{Exponential}(1/29) + 1.$$

95% HPI for the median APE, $\tilde{\mu}$, was estimated by the following transformation of simulated draws from the posterior distribution

$$\tilde{\mu} = e^{\mu}.$$

In summary, the results we present are 95% HPI from the posterior distributions of the median error, $\tilde{\mu}$.

*Inferring the relative accuracy of the CNN and likelihood-based method.*— To quantify the difference in error between the CNN and the likelihood-based method, we fit the difference in sampled APE scores, $\Delta$APE, between the CNN method and the likelihood-based method to the BEST model. Histograms of $\Delta$APE appear symmetric with weak to strong outliers making the BEST model a good candidate for inference from this data. The data and likelihood are

$$\Delta y = \text{APE}^{\text{CNN}} - \text{APE}^{\text{Like}}$$

$$\Delta y \mid \mu, \sigma, \nu \sim t_{\nu}(\mu, \sigma)$$

We used the same default priors as above.

Because, $\Delta y$ is not log-transformed, it is drawn from a t-distribution and the

marginal posterior of the parameter $\mu$ is an estimate of the population mean, $\mu^d$. Because the mean and the median are equivalent for a t-distribution, we again report the posterior distribution of the median difference, $\tilde{\mu}^d$ to simplify the results.

In summary, the results we present are 95% HPI from the posterior distribution of the median difference between the two methods, $\tilde{\mu}^d$.

When comparing CNN to the likelihood-based approach, positive values for $\tilde{\mu}^d$ indicate the CNN is less accurate, and negative indicate the likelihood-based estimates less accurate. We emphasise that this quantity is the median difference in contrast to the difference in medians, $\Delta\tilde{\mu}$, reported in the next section.

*Inferring sensitivity to model misspecification.*— Finally, to quantify the overall sensitivity of each rate parameter to model misspecification under each inference method, we infer the difference in median APE, $\tilde{\mu}$ of predictions under a misspecified model relative to predictions under the true model. In other words we are inferring differences in medians between experiments. For example, to infer the sensitivity of the CNN's inference of the sampling rate, $\delta$, to phylogenetic error, we inferred the difference between the median APE of the CNN's predictions for misspecified trees and the median APE of CNN predictions for true trees. The data is concatenated as below.

$$(y_1, y_2) = (\text{APE}^{\text{CNN}}, \text{APE}^{\text{CNN Ref}}) \text{ or}$$
$$(y_1, y_2) = (\text{APE}^{\text{Like}}, \text{APE}^{\text{Like Ref}})$$

We inferred the difference between group median APE scores, denoted $\Delta\tilde{\mu}$, by assuming that the model parameters conditioned on the observed APE from the two groups, $y_1$ and $y_2$, follow a posterior distribution that is proportional to

$$P(y_1 \mid \mu_1, \sigma_1, \nu)P(y_2 \mid \mu_2, \sigma_2, \nu)P(\mu_1, \mu_2, \sigma_1, \sigma_2, \nu),$$

where $\log y_1$ and $\log y_2$ follow t distributions with means $\mu_1$ and $\mu_2$ and standard deviations $\sigma_1$ and $\sigma_2$, respectively while sharing a common normality parameter, $\nu$.

The posterior sample of $\Delta\tilde{\mu}$ is obtained by transforming samples from the joint marginal posterior distribution of $\mu_1$ and $\mu_2$ with the following equation,

$$\Delta\tilde{\mu} = e^{\mu_1} - e^{\mu_2}.$$

The two components of the likelihood are each t-distributed and share the $\nu$ parameter which means we assume both samples are drawn from a similarly shaped distribution (similarly heavy tails).

$$\log y_1 \mid \mu_1, \sigma_1, \nu \sim t_\nu(\mu_1, \sigma_1)$$

$$\log y_2 \mid \mu_2, \sigma_2, \nu \sim t_\nu(\mu_2, \sigma_2)$$

The prior distribution for the parameters of the model were set to the defaults for BEST,

$$\mu_1 \sim \text{Normal}(\text{mean}(\log\, y_1), \text{sd}(\log\, y_1) \times 1000)$$

$$\mu_2 \sim \text{Normal}(\text{mean}(\log\, y_2), \text{sd}(\log\, y_2) \times 1000)$$

$$\sigma_1 \sim \text{Uniform}(\text{sd}(\log\, y_1)/1000, \text{sd}(\log\, y_1) \times 1000)$$

$$\sigma_2 \sim \text{Uniform}(\text{sd}(\log\, y_2)/1000, \text{sd}(\log\, y_2) \times 1000)$$

$$\nu \sim \text{Exponential}(1/29) + 1$$

As before, interpretation of the posterior distribution of the difference in medians is straightforward: the more positive the difference in median APE from the misspecified model test set and the median APE from the true model test set, the more sensitive the parameter is to model misspecification in the experiment.

## *CNN uncertainty quantification*

We used conformalized quantile regression (CQR) to construct calibrated probability intervals (CPI), ensuring accurate predictive coverage (Lei et al. 2018; Romano et al. 2019; Sousa et al. 2022; Vovk et al. 2022; Angelopoulos et al. 2023). CQR is implemented in two stages: first a network is trained to predict conditional quantiles, then a hold-out simulated dataset is used to estimate bias adjustment terms to ensure correct coverage on future data *i.e.* 95% intervals contain the true value 95% of the time for test data.

To implement quantile regression with a neural network and predict lower and upper quantiles, we adjusted the general network architecture used for point estimates above to have two outputs each with a mean pinball loss function instead of the mean squared error,

$$L_\tau(y, \hat{q}) = \frac{1}{N} \sum_i^N \left[ (y_i - \hat{q}_i)\tau \mathbb{1}\{y_i \geq \hat{q}_i\} + (\hat{q}_i - y_i)(1-\tau)\mathbb{1}\{y_i \leq \hat{q}_i\} \right].$$

Here, $y$ is the label or true parameter value (not a quantile) and $\hat{q}$ is the trained neural network's prediction of a given quantile. $\tau$ is the quantile level and is equal to $1-\alpha$, where $\alpha$ is the mis-coverage rate, or the probability the true value is not below the quantile. To estimate inner quantiles with miscoverage rate $\alpha$, the lower quantile output was set to predict the $\alpha/2$ quantile for each rate parameter and the other layer to predict the $1-\alpha/2$ upper quantile (Steinwart and Christmann 2011) (SI figure S2). We refer to CNNs of this type as qCNN. Though often close, these inner quantiles are not guaranteed to have the correct coverage on test data sets (Figure 3) necessitating the calibration (conformalization) step (Romano et al. 2019).

To calibrate the predictions of quantile regression neural networks, CQR finds an adjustment term for each quantile through computing a non-comformity score, such as the distance of the predicted value from the predicted quantile. If the estimated quantile is well calibrated, then the same quantile of the scores in a calibration set will be zero. If the estimated quantile is, for example, too high then too high a proportion of the labels will fall below the estimated quantile and the empirical quantile, $Q$, of the nonconformity score $y-\hat{q}$ at $1-\alpha/2$ will be negative. In other words it will over cover the calibration set. $Q$ thus becomes the adjustment term for calibrating the qCNN's quantile estimate (equations 9, and 10) by simply adding the term to the corresponding estimated quantile as shown in equation 11.

$$Q_{\text{lower}} \text{ s.t. } P(y-\hat{q}_{\text{lower}} < Q_{\text{lower}}) = \frac{\alpha}{2}\left(1+\frac{1}{n}\right) \tag{9}$$

$$Q_{\text{upper}} \text{ s.t. } P(y-\hat{q}_{\text{upper}} < Q_{\text{upper}}) = \left(1-\frac{\alpha}{2}\right)\left(1+\frac{1}{n}\right) \tag{10}$$

$$\text{CPI} = [\hat{q}_{\text{lower}} + Q_{\text{lower}}, \hat{q}_{\text{upper}} + Q_{\text{upper}}] \tag{11}$$

Note that the quantiles of the score for finite sample sizes require adjustment by $(1+\frac{1}{n})$ where $n$ is the number of samples in the calibration set (Romano et al. 2019).

We simulated 108,559 more datasets (trees) to estimate the calibration amounts for the upper and lower qCNN-estimated quantiles. After calibration through conformalization, we clipped intervals to the prior boundary for intervals that extended beyond the prior distribution's range. To examine the consistency of quantile regression for neural networks trained on different quantiles we trained seven different quantile networks to predict the same quantiles used for validating our Bayesian analysis and simulation model: {0.05, 0.25, 0.5, 0.75, 0.9, 0.95}. We checked the coverage of these adjusted CPIs on another simulated test dataset of 5,000 trees.

## *Real data*

We compared the inferences of a LDBDS simulation trained neural network to that of a phylodynamic study of the first COVID wave in Europe (Nadeau et al. 2021). These authors analyzed a phylogenetic tree of viruses sampled in Europe and Hubei, China using a location-dependent birth-death-sampling model in a Bayesian framework using priors informed by myriad other sources of information. We simulated a new training set of trees under an LDBDS model where $R_{0_i}$ depends on the geographic location, and the sampling process only consists of serial sampling and no sampling of extant infected individuals. We estimated 95% CPIs for model parameters with a simulated calibration dataset of 101,219 trees using CQR as above and confirmed accurate coverages with another dataset of 5,000 trees.

We then analyzed the whole tree from Fig. 1 in (Nadeau et al. 2021) as well as the European clade which Nadeau et al. (2021) labeled as A2 in the same figure. We note that our simulating model is not identical to the inference model used in (Nadeau et al. 2021). We model migration with a single parameter with symmetrical migration rates among locations and all locations having the same sampling rate. Nadeau and colleagues parameterize the migration process with asymmetric pairwise migration rates and assume location-specific sampling rates. We also do not include the information the authors used

to inform their priors as that requires an extra level of simulation and training on top of simulations done here, and is thus beyond the scope of this study.

The time tree from (Nadeau et al. 2021) was downloaded from GitHub (https://github.com/SarahNadeau/cov-europe-bdmm). The recovery rate assumed in (Nadeau et al. 2021) was 0.1 days$^{-1}$ which was set to 0.05 to bring the recovery rate to within the range of simulating values used to train the CNN. Consequently, the branch lengths of the tree were then scaled by 2. The number of tips, tree height, and average branch lengths were measured from the rescaled trees and fed into the network. The full tree and A2 clade were then analyzed using the LDBD CNN and compared to the posterior distributions from (Nadeau et al. 2021).

### *Hardware used*

Simulations were run on a 16 core Intel(R) Xeon(R) Platinum 8175M CPU @ 2.50GHz. For each simulation, an XML file with random parameter settings was generated using custom scripts. These XML files were the inputs for MASTER which was run in the BEAST2 platform. Neural network training and testing and predictions were conducted on an 8 core Intel(R) Core(TM) i7-7820HQ CPU @ 2.90GHz laptop with a NVIDIA Quadro M1200 GPU for training.

## Results

### *Comparing deep learning to likelihood*

Our first goal in this study was to train a CNN that produced phylodynamic parameter point estimates that were as accurate as likelihood-based Bayesian posterior mean estimates under the true model. This will serve as a reference for quantifying level of sensitivity in our misspecification experiments. Using viral phylogenies like those typically

estimated from serially sampled DNA sequences, we focused on estimating important epidemiological parameters – the reproduction number, $R_0$, the sampling rate, $\delta$, the migration rate, $m$, and the outbreak origin.

Our CNN produced estimates that are as accurate as the mean posterior estimates (MPE) under the true simulating model. We compared the absolute percent error (APE) of the network predictions to the APE of the MPE of the Bayesian location-independent birth-death-sampling (LIBDS) model (Figure 2). The APE is straight-forward to interpret, e.g. an APE of $< 10$ means the estimate is within 10 percentage points (ppts) of the true value. For the three epidemiological rate parameters, $R_0$, $\delta$ and $m$, both methods made very similar predictions for the 100 time tree test set (Figure 2 panel A). The two methods appear to produce estimates that are more similar to each other than to the ground truth labels (compare bottom row scatter plots in orange to the blue and red scatter plots in panel A). Fig. 2 panel B shows that the inferred median difference in APE, $\tilde{\mu}^d$, between the method's estimates for the three parameters is close to zero ($\mid \tilde{\mu}^d \mid$ 95% HPI is $< 4$ ppts; SI Table S1; SI Figure S3).

We also compared the performance of uncertainty quantification using quantile-CNN-based conformalized quantile regression (CQR; Romano et al. 2019) to that of Bayesian HPIs for each of the experiments. We trained seven qCNNs to predict inner-quantiles at seven different levels to compare with the Bayesian HPIs; $\tau = \{0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$. We then used another simulated dataset to calibrate predicted intervals which we refer to as CPIs which theoretically have correct coverage properties (Romano et al. 2019) like the HPIs. For the test dataset of 138 trees, the CPIs had coverages that matched well with expectations to a comparable degree to the Bayesian HPI (Figure 2 panel C) though more variable. To further confirm that our CQR procedure was adequately calibrating the qCNN estimates, we confirmed correct coverages of CPIs for a much larger dataset with 5,000 trees (Figure 3). On average, the widths of CPIs in the set of 138 trees shown in (Figure 2) was about 20 - 40% wider than that of the corresponding

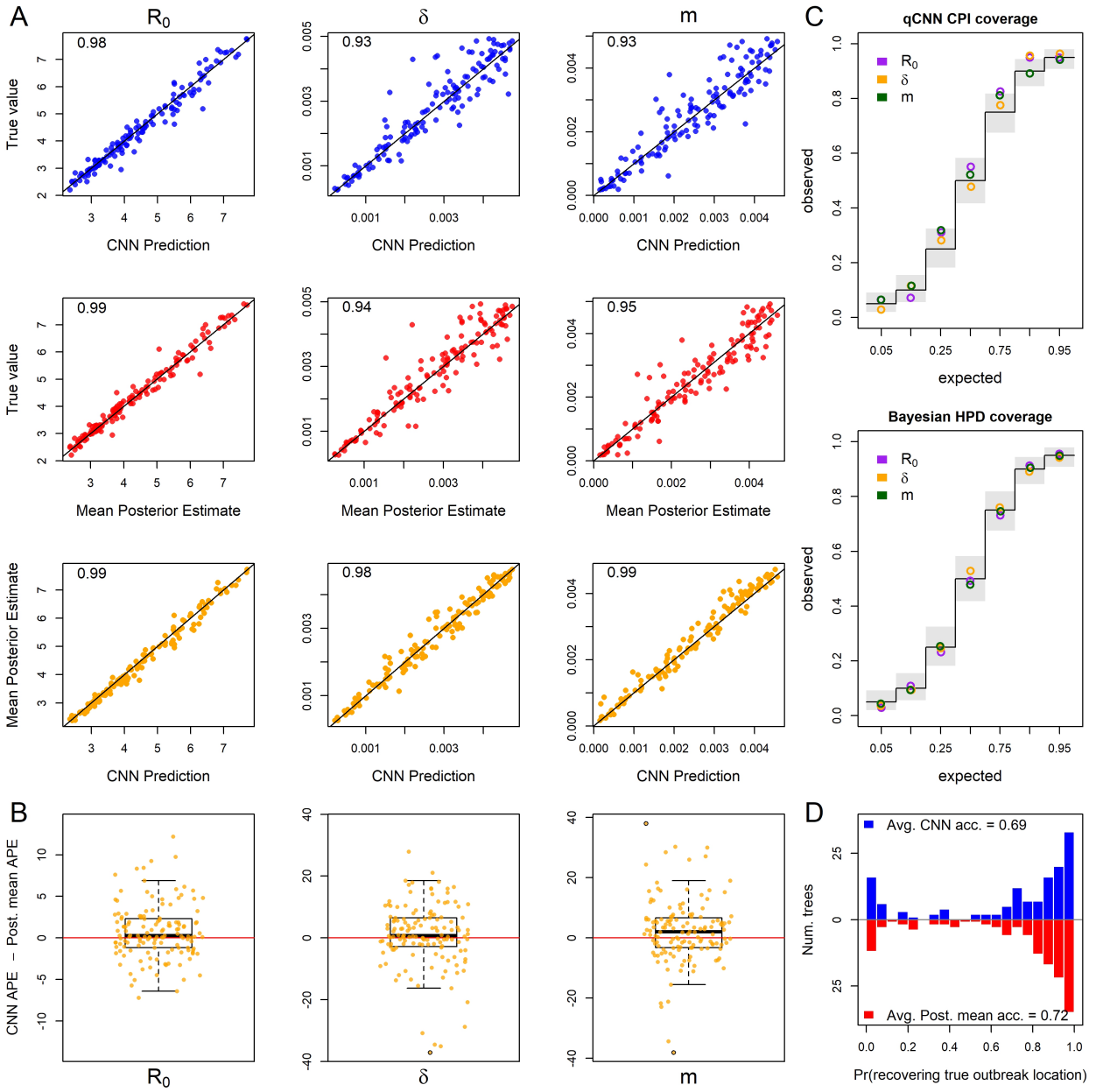


Figure 2: Inference under the true simulating model. (A) Scatterplot of CNN predictions and posterior mean estimates from Bayesian analyses against the true values (top two rows in blue and red respectively) of the basic reproduction number, $R_0$, the sampling rate, $\delta$, and the migration rate, m for 138 test trees. In the upper-left corners of the scatter plots are the correlations of the plotted data. The bottom row in orange shows scatter plots of the CNN estimates against the posterior mean estimates for the same trees. (B) The difference in the absolute percent error (APE) of estimates for the two inference methods. Boxes show the inner 50% quantile of the data while whiskers extend 1.5 IQR. Dots with black circles were truncated to 2× the length of whiskers for visualization purposes. (C) Coverage plots show the expected frequency of coverage for each of the categories and the observed frequencies (black steps and colored circle respectively). Gray boxes are the expected 95% confidence intervals at each of the expected coverage values which follows a Beta$((n+1)q, n-(n+1)q+1)$ distribution. (D) Histograms of the probabilities of inferring the correct outbreak origin location.

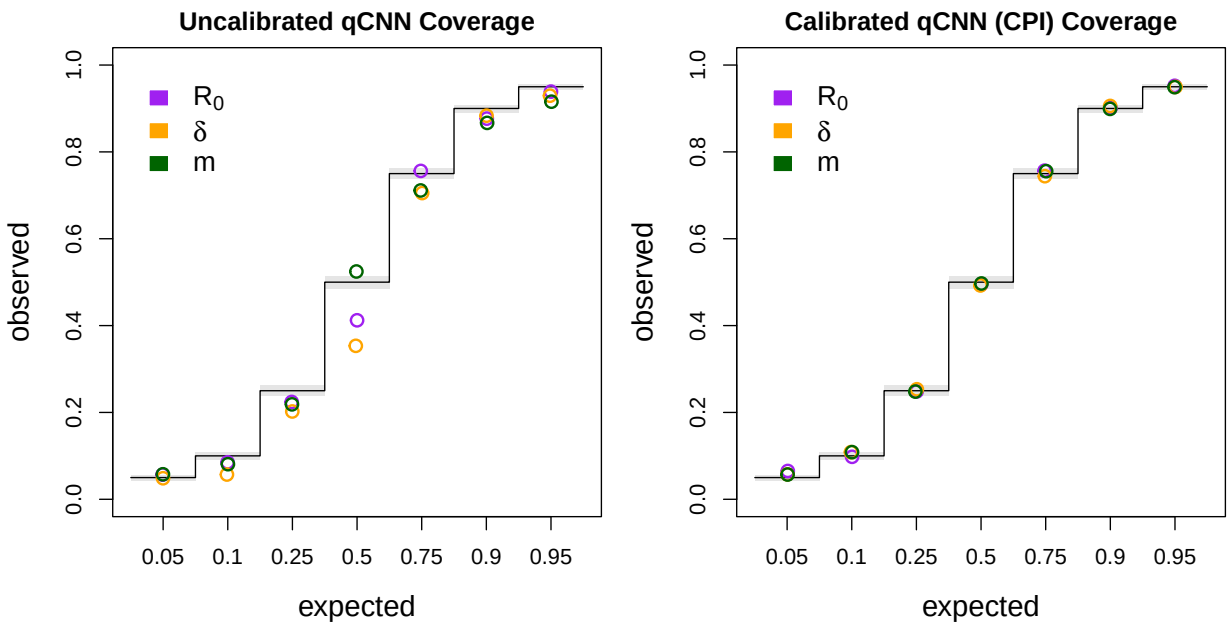


Figure 3: Coverage of uncalibrated qCNN quantile predictions (left) and calibrated qCNN which produce "calibrated probability intervals" (CPI) on the right. The observed coverage of 5,000 samples tested at seven different predicted coverage levels (labeled horizontal). See Figure 2 C for more details on coverage plots.

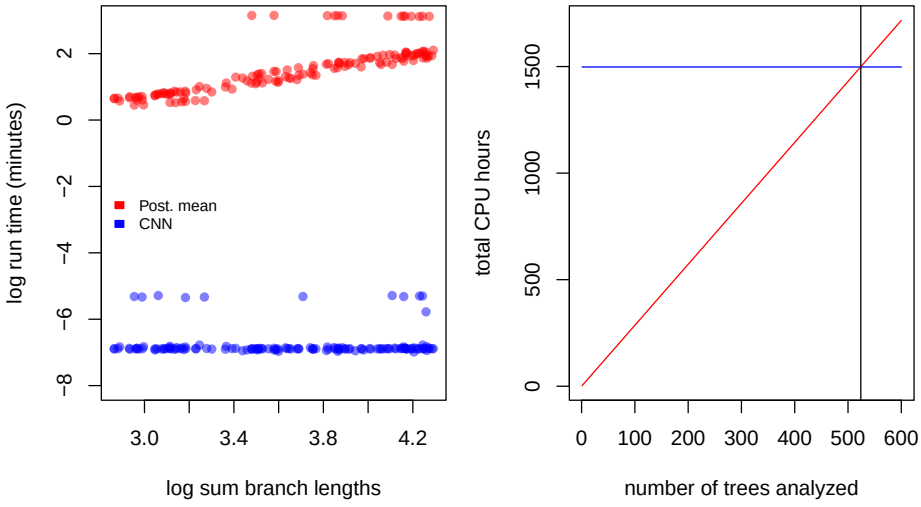


Figure 4: Left: Estimates of time to complete analysis of each of 138 trees relative to tree size. Right: The number of trees (524; gray vertical line) needed to analyze for total analysis time of Bayesian method (red line) to equal that of the entire simulation and CNN training and inference pipeline (blue line).

HPI and Jaccard similarity index ranging from 0.66 to 0.75 suggesting a high degree of overlap between the intervals (SI Figure S4 and SI Table S2). These results indicate the probability level of the CPI, *e.g.* 95%, can be safely interpreted as the probability a parameter falls within the CPI. The wider intervals suggest the basic CQR method employed here is somewhat less precise and thus more conservative than the Bayesian method.

Our trained CNN provides nearly instantaneous estimates of model parameters. While the run time of the likelihood approach employed in this study scales linearly with the size of the tree, the neural network has virtually constant run times that are more than three orders of magnitude faster. Because simulation-trained neural networks have a one-time cost of simulating the training data set and then training the neural network, these methods are often called amortized-approximators (Bürkner et al. 2022). This means the time savings aren't recouped until a certain number of trees have been analyzed. For example, here over 524 trees would need to be analyzed to realize the cost savings of simulating data and training our neural network (Figure 4). This illustrates the importance of simulation optimization and generality for likelihood-free approaches to inference.

## *Comparing sensitivity to model misspecification*

To test the relative sensitivity of CNN estimates and the likelihood-based MPE to model misspecification, we simulated several test data sets under different, more complex epidemic scenarios and compared the decrease in accuracy (increase in APE).

Our first model misspecification experiment tested performance when assuming all locations had the same $R_0$ when, in fact, each location had different $R_{0_i}$ values. The median APE for all three parameters increased to varying degrees (SI Fig. S5 Panel A) compared to the median APE measured in Fig. S3. We found that both methods converged on similar biased estimates for $R_0$. In both the CNN and Bayesian method, estimates of $\delta$ were relatively robust to misspecifying $R_0$. In contrast, the migration rate

showed much more sensitivity to this model violation in both methods with both methods also converging on similarly biased estimates (Figure 5 A). The median difference in error between the two methods is close to zero for all rate parameters ($|\ \tilde{\mu}^d\ |$ 95% HPI $< 6$ ppts; SI Table S1) (SI Figure S5 Panel B). For both methods of uncertainty quantification the coverage declined by similar amounts for all three parameters with $\delta$ showing little to no sensitivity to $R_0$ misspecification (Figure 5 panel C and SI Table S2). The patterns of coverage are also somewhat less regular across the qCNN quantiles than the HPIs for the migration rate parameter likely due in part to the fact that each inner quantile qCNN was trained independently and thus have independent errors. The relative interval widths and Jaccard similarity indexes did not change appreciably from predictions under the true model (SI Figure S4 and SI Table S2). Our CNN appears to be slightly more sensitive than the Bayesian approach when predicting the outbreak location. Nevertheless, their distributions are quite similar (Figure 5 Panel C).

Next, we measured method sensitivity when the sampling process of the test trees violates assumptions in the inference model. In this set, each location had a unique and independent sampling rate, $\delta$, rather than a single $\delta$ shared among locations. The median APE only increased for $\delta$ and m (SI Figure S7 Panel A). As expected, estimates of $\delta$ were highly biased for both methods (Figure 6 panel A). Panel A also shows that $R_0$ is virtually insensitive to sampling model misspecification, but that migration rate, again, is highly sensitive in both the CNN and likelihood method. The median difference in error between the two methods is close to zero for all the rate parameters ($|\ \tilde{\mu}^d\ |$ 95% HPI $< 5$ ppts; SI Table S1, SI Figure S7) (Figure 6 panel B). For both methods coverage declined for $\delta$ and m, while $R_0$ showed little to no sensitivity to $\delta$ misspecification (Figure 6 panel C and SI Table S2). The relative widths and degree of overlap was again similar to the experiments above (SI Figure S8, SI Table S2). We again also see greater irregularity among CPI levels in coverage, notably $\delta$ at inner-quantile level 0.9. The location of outbreak prediction is also somewhat sensitive in both methods, with the CNN showing a slightly larger mean

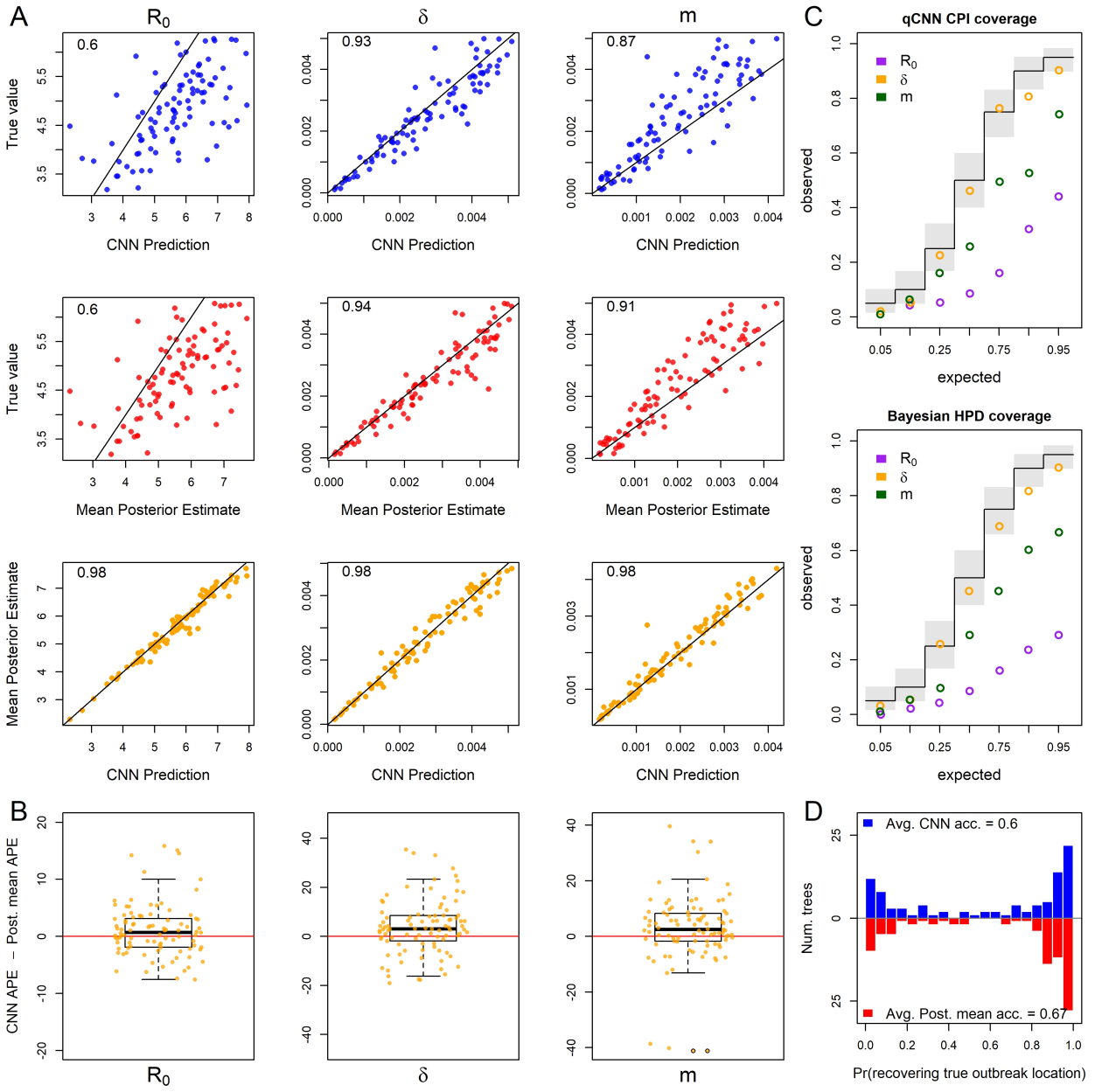


Figure 5: For 93 test trees where the $R_0$ parameter was misspecified: the simulating model for the test data specified 5 unique $R_0$s among the five locations while the inference methods assumed one $R_0$ shared among locations. Because of this, the estimates for $R_0$ are plotted against mean of the five true $R_0$ values. See Figure 2 for general details about plots.

difference, but the overall distribution of accuracy of all the test trees again is similar (Figure 6 panel C).

To explore sensitivity to migration model underspecification, we simulated a test set where the migration rates between locations is free to vary rather than being the same

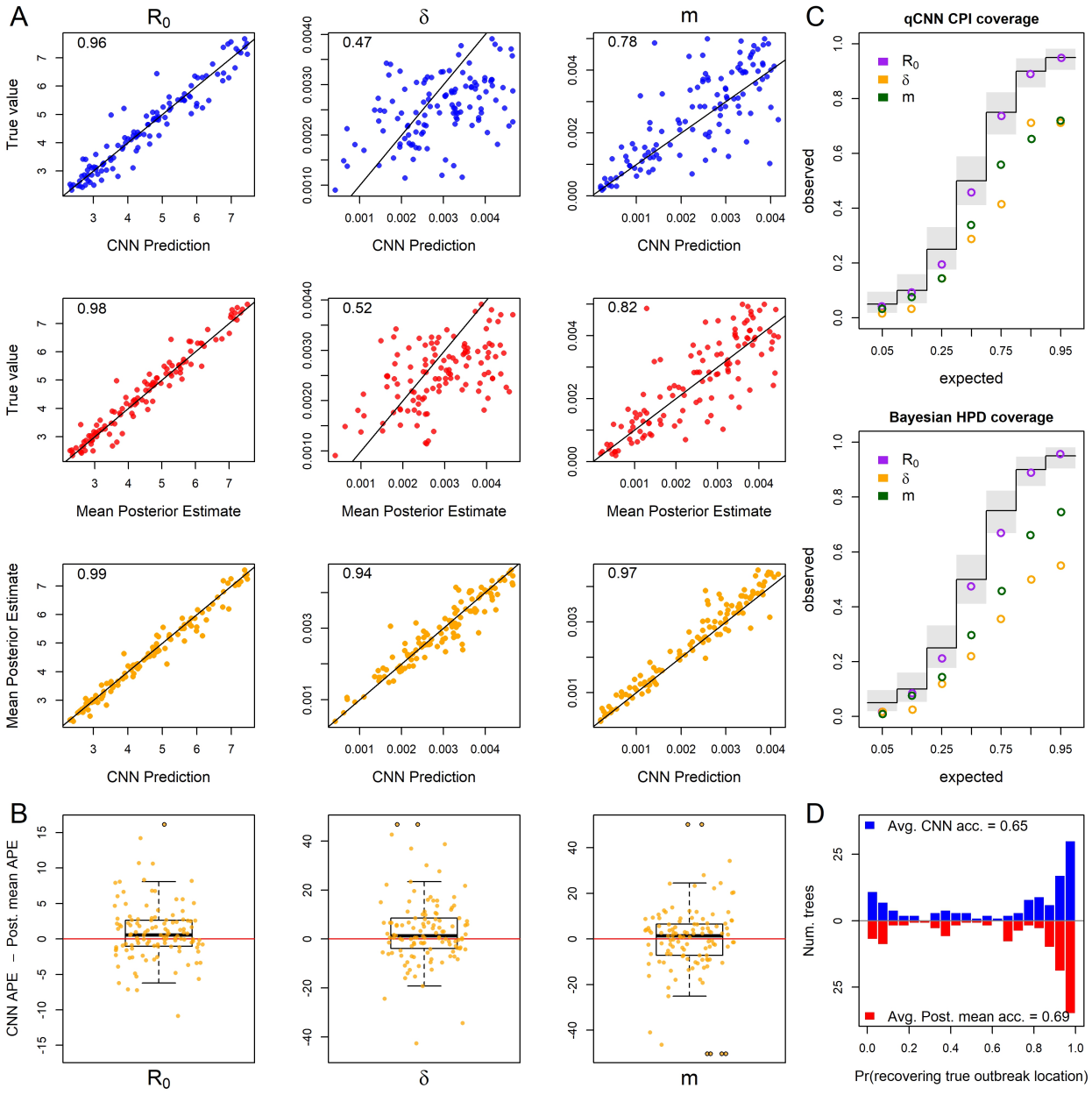


Figure 6: For 118 test trees where the sampling rate parameter was misspecified: the simulating model for the test data specified 5 unique sampling rates among the five locations while the inference methods assumed one sampling rate shared among locations. The estimates of $\delta$ are plotted against the mean true values of $\delta$. See Figure 2 for general details about plots.

among locations as in the inference model. This implies 5! unique location-pairs and thus unique migration rates in the test data set. Results show that for both methods the parameters $R_0$ and $\delta$ are highly robust to this simplification (SI Fig. S9 Panel A). Though

estimates of a single migration rate had a high degree of error compared to a single pair of locations' migration rates (Figure 7 panel A), the two methods still had similar estimates with the difference in APE centered near zero (Figure 7 panel B). The inferred median difference in APE was close to zero ($| \tilde{\mu}^d |$ 95% HPI $< 3$ ppts; SI Table S1; SI Figure S9 Panel B). For both methods the coverage only declined significantly for the migration rate and the decrease was again similar in magnitude across quantiles (Figure 7 panel C and SI Table S2). Again, relative widths and degree of overlap of CPI and HPI were similar to previous experiments (SI Figure S10, SI Table S2) There was a slight but similar decrease in accuracy in predicting the outbreak location for both methods (Figure 7 panel C).

When testing the sensitivity of the two methods to arbitrary groupings of locations, we found that both methods showed equal sensitivity to the same parameters (Fig. 8 Panels A and B). In particular, the migration rate showed a modest increase in median APE and $R_0$ and sample rate showed virtually no sensitivity to arbitrary grouping of locations (SI Figure S11 Panel A). The inferred median difference between method APE's was again close to zero ($| \tilde{\mu}^d |$ 95% HPI $< 4$ ppts; SI Table S1; SI Figure S11 Panel B). For both methods the coverage declined modestly only for the migration rate (Figure 5 panel C and SI Table S2). Relative widths and interval overlap showed virtually no change (SI Figure S12 and SI Table S1). These results suggest that for at least the exponential phase of outbreaks where rate parameters do not vary among locations, these models have a fair amount of robustness to the decisions leading to geographical division of continuous space into discrete space. The outbreak location showed higher accuracy in both methods due to the fact that the test data was no longer a flat distribution; the 6 combined locations should contain 60% of the outbreak locations (Figure 8 panel C).

Finally, we explored the relative sensitivity of our CNN to amounts of phylogenetic error that are present in typical phylogeographic analyses. Our simulated phylogenetic error produced trees with average Jaccard similarity indexes between the inferred tree and the true tree of about 0.5 with 95% of simulated trees having distances within 0.36 and 0.72.

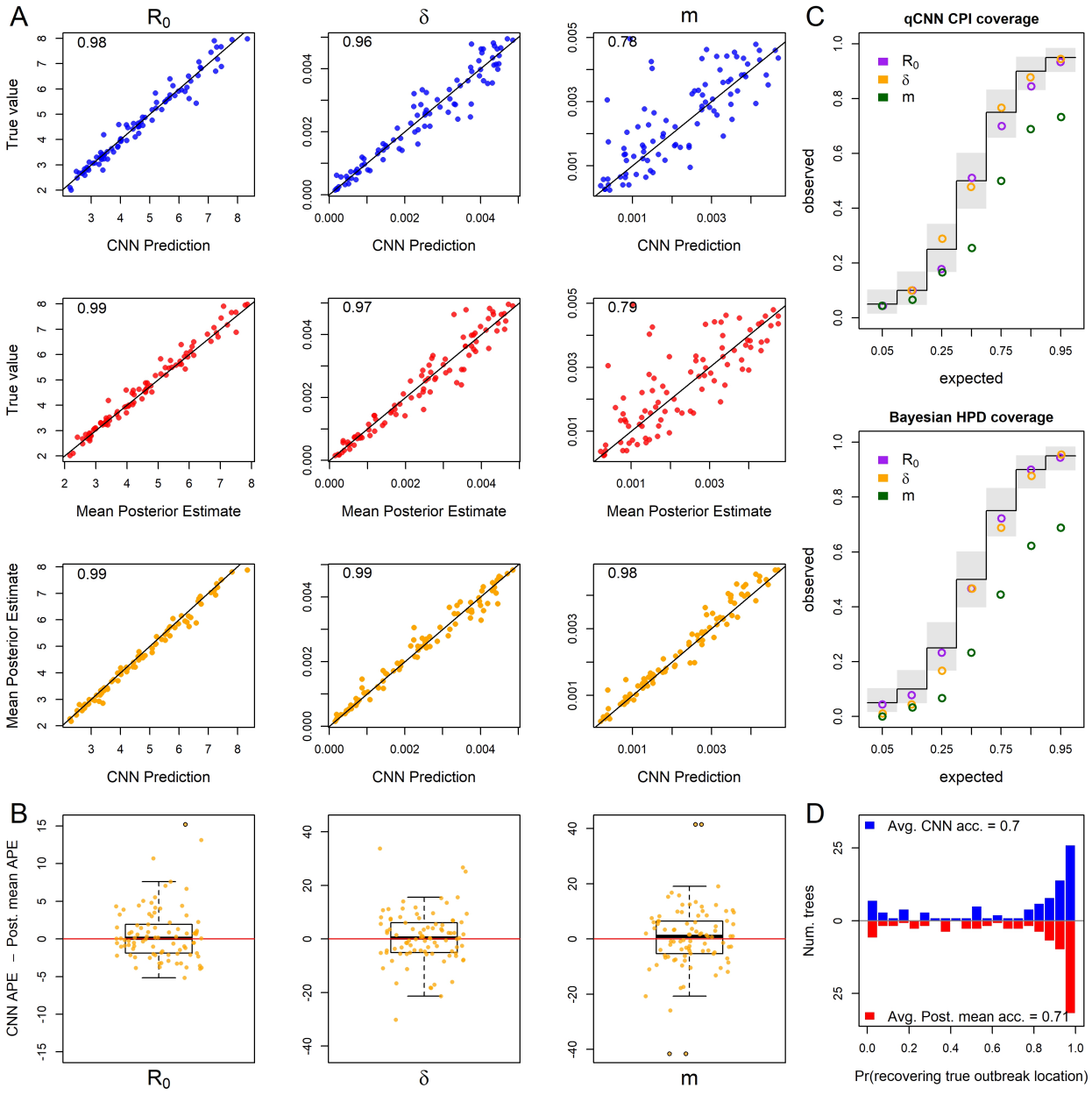


Figure 7: For 90 test trees where the migration rate parameter was misspecified: the simulating model for the test data specified 5! (120) unique migration rates among the unique pairs of the five locations while the inference methods assumed all migration rates were equal. The infered migration rate is plotted against the mean pairwise migraiton rates of test data set. See Figure 2 for general details about plots.

We again compared inferences derived from the true tree and the tree with errors using the CNN and the Bayesian LIBDS methods. Results show that migration rate was minimally affected but $R_0$ and $\delta$ were to a some degree sensitive to phylogenetic error (Figure 9 panel

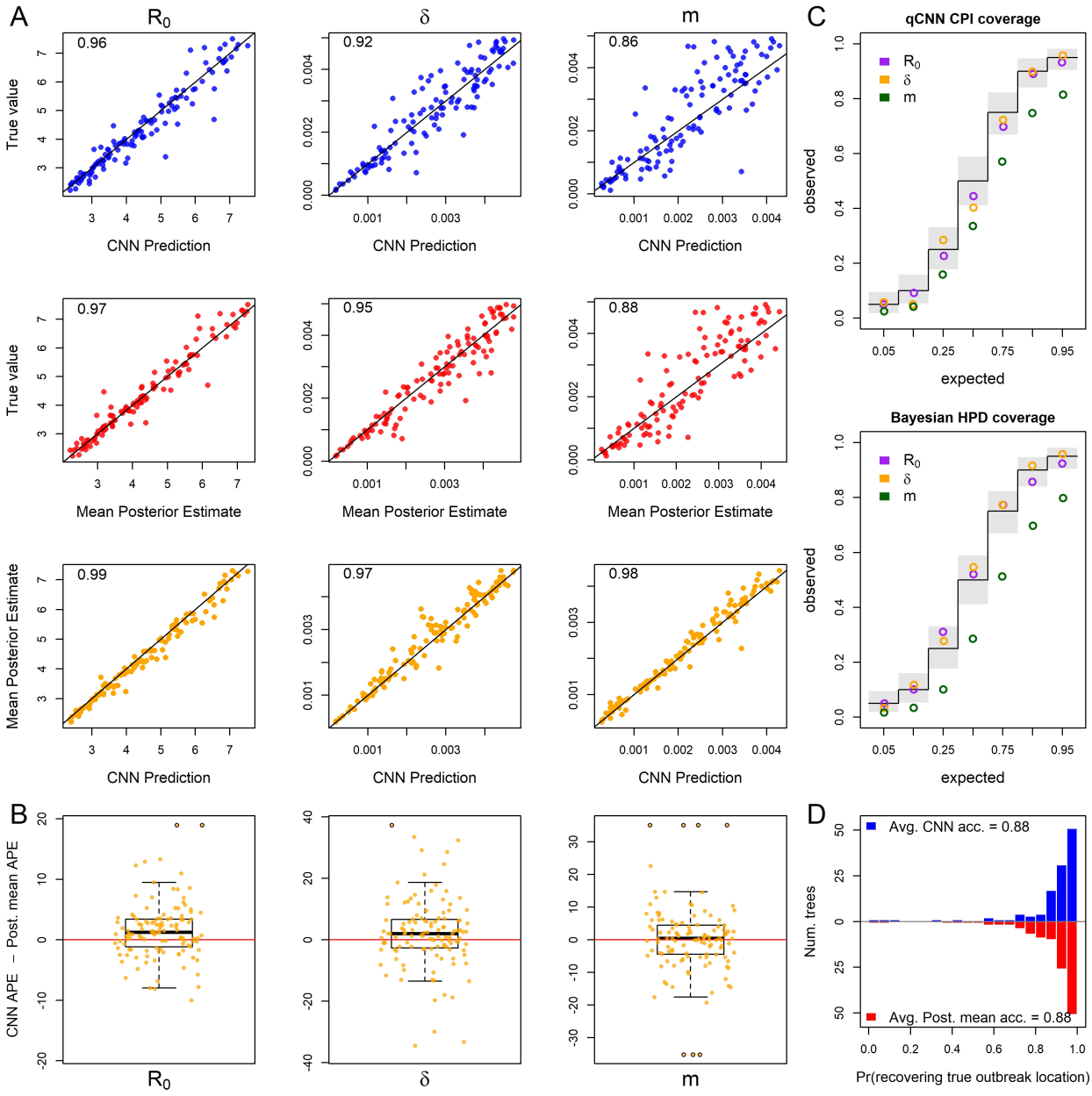


Figure 8: For 101 test trees where the number of locations was misspecified: the simulating model for the test data specified an outbreak among 10 locations with 6 locations subsequently combined into a single location while the inference methods assumed 5 locations with no arbitrary combining of locations. See Figure 2 for general details about plots.

A; SI Figure S13 Panel A), with both methods again showing similar degrees of sensitivity (Figure 9 panel B). The inferred median difference was, yet again, small ($|\ \tilde{\mu}^d\ |$ 95% HPI $< 6$ ppts. SI Table S1, SI Figure S13 Panel B). Coverages of $\delta$ declined for both methods in a similar way across quantiles. Again the 90% inner quantile showed some inconsistency

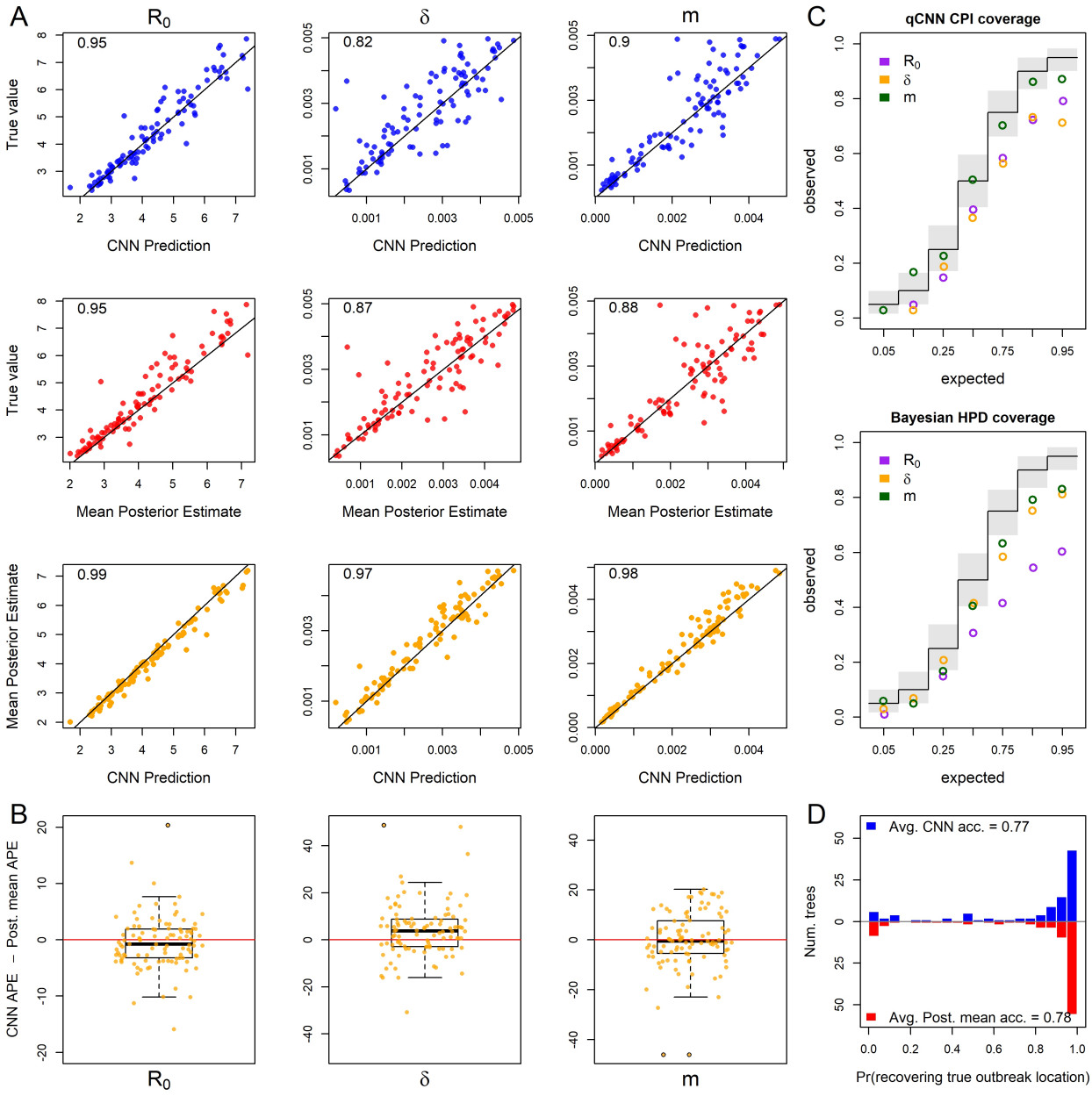


Figure 9: For 118 test trees where the time tree was misspecified: the true tree from the simulated test set was replaced with an inferred tree from simulated DNA alignments under the true tree. See Figure 2 for general details about plots.

with its nieghboring quantiles. In this case its coverage for $\delta$ was slightly higher than the 95th inner quantile. The CPIs for $R_0$ appear much less sensitive (Figure 9 panel C and SI Table S2). Although the relative widths of the CPIs and HPIs were similar to previous experiments, the degree of overlap decreased somewhat by about 5 - 10% (SI Figure S14

and SI Table S2). One difference between this experiment and the others, is that trees are data instead of model parameters. It is interesting that the point estimates from the two methods show similar biases while the coverages seem to depart somewhat. Inference of the origin location, were very similar for both methods (Fig. 9 Panel C).

## *Analysis of SARS CoV-2 tree*

We next compared our likelihood-free method to a recent study investigating the phylodynamics of the first wave of the SARS CoV-2 pandemic in Europe (Nadeau et al. 2021). Despite simulating the migration and the sampling processes differently from Nadeau et al. (2021), our CNN produces similar estimates for the location-specific $R_0$ and the origin of the A2 clade (Figure 10). Whether the full tree or just the A2 clade is fed into the network, the predicted $R_0$ for each location was not far from the posterior estimates of Nadeau et al. (2021). For the most part the $R_0$ 95% CPI for each location overlaps to a high degree with the 95% HPI and is roughly 1.5 times wider indicating that our CNN estimates are relatively conservative. For Hubei the interval width of the a2 clade is much wider than the estimate using the whole tree. This is not surprising because there are no samples from Hubei in the a2 clade. We also obtained estimates for a single sampling rate and a single migration rate from our CNN and CPIs from our calibrated qCNN. Among the five location-specific estimates of the sampling proportion and the migration proportion from Nadeau et al. (2021), our CNN's point estimates and interval estimates fall well within the their combined ranges.

The spillover location prediction CNN produced probability estimates of the A2 clade ancestral location the mostly agreed with that of Nadeau and colleagues (Figure 10, right histograms). The only significant discrepancy in the European origin prediction is that Nadeau and colleague's analysis suggests a much higher probability that the most recent common ancestor of the A2 clade was in Hubei than our CNN predicts. This is likely because our CNN only used the A2 clade to predict A2 origins which has no Hubei

samples to infer the origin of the A2 clade while Nadeau et al. (2021) used the whole tree. Notwithstanding this difference, among European locations, both methods predict Germany is the most likely location of the most recent common ancestor followed by Italy.

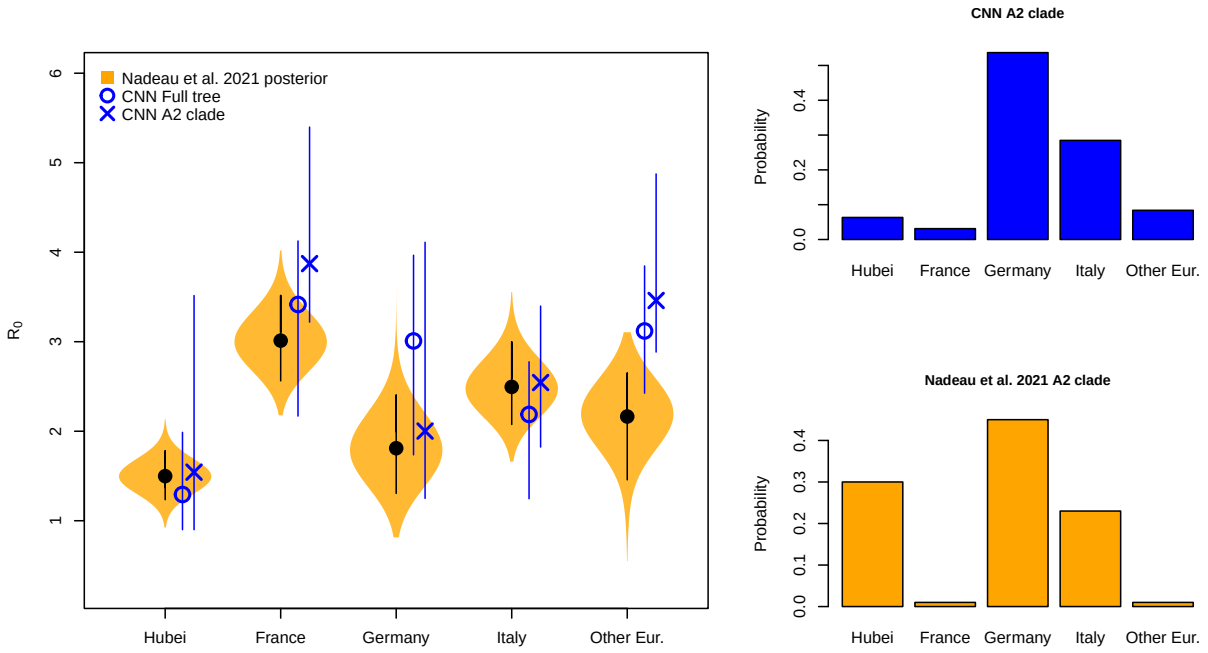


Figure 10: Location-dependent birth-death-sampling model (LDBDS) CNN comparison to (Nadeau et al. 2021) inference. Left violin plots show the posterior distributions of $R_0$ for each location in Europe as well as Hubei, China (orange). The block dot and line within each violin plot shows the posterior mean and 95% HPI respectively. The blue X and O marks the LDBDS CNN prediction from analyzing the full tree and the A2 (European) clade respectively. Vertical blue lines give the 95% CPI for the CNN estimates of $R_0$. Right barplots show the LDBDS CNN prediction (blue) and posterior inference (orange) from (Nadeau et al. 2021) of the ancestral location of the A2 (European) clade (see Figure 1 (Nadeau et al. 2021)).

# Discussion and Conclusions

Inference models are necessarily simplified approximations of the real world. Both simulation-trained neural networks and likelihood-based inference approaches suffer from model under-specification and/or misspecification. When comparing inference methods it is

important to assess the sensitivity of model inference to simplifying assumptions. In this study we show that newer deep learning approaches and standard Bayesian approaches behave and misbehave in similar ways under a panel of phylodynamic estimation tasks where the inference model is correct as well as when it is misspecified.

By extending new approaches to encode phylogenetic trees in a compact data structure (Voznica et al. 2022; Lambert et al. 2022), we have developed the first application of phylodynamic deep learning applied to phylogeography with serial sampling. Our approach is similar to that of Lambert et al. (2022) in which they analyzed a binary SSE model with exclusively extant sampling. By training a neural network on phylogenetic trees generated by simulated epidemics, we were able to accurately estimate key epidemiological parameters, such as the reproduction number and migration rate, in a fraction of the time it would take with likelihood-based methods. Like Voznica et al. (2022) and Lambert et al. (2022), we found that CNN estimators perform as well or nearly as well as likelihood-based estimators under conditions where the inference model is correctly specified to match the simulation model. The success of these separate applications of deep learning to different phylodynamic problems is a testament to the versatility of the CBLV encoding of trees.

We compared the sensitivity of deep learning and likelihood-based inference to model misspecification. Because deep-learning methods of phylogenetic and phylodynamic inference are new, few studies compare how simulation-trained deep learning methods fail in comparison to likelihood methods in this way (Flagel et al. 2019). We assume that when the inference model is correctly specified to match the simulation model, the trained CNN will, at best, produce noisy approximations of likelihood-based parameter estimates. In reality, issues related to training data set size, learning efficiency, and network overfitting may cause our CNN-based estimates to contain excess variance or bias when compared to Bayesian likelihood-based estimators. Our results from five model misspecification experiments show that both methods of inference perform similarly when the simulating model and the inference model assumptions do not perfectly match. These similarities

exist not only in aggregate, when comparing method performance across datasets, but also when comparing performance for each individual dataset. This suggests that the CNN and likelihood methods are truly estimating parameters using isomorphic criteria, despite the fact that CNN heuristically learns these criteria through data patterns, while likelihood precisely and mathematically defines these criteria through the model definition itself.

Results of comparative sensitivity experiments like this are important because if likelihood-free methods using deep neural networks can easily be trained to yield estimates that are as robust to model misspecification as likelihood-based methods, then analysis of a large space of more complex outbreak scenarios for which tractable likelihood functions are not available can be developed and applied to real world data. Additionally, sufficiently realistic, pre-trained neural networks can yield nearly instantaneous inferences from data in real time to inform analysts and policy makers.

We also tested location-dependent SIR simulation trained neural network against a previous publication fitting a similar model – location-dependent birth-death-sampling (LDBDS) model – on real-world data using a Bayesian method. Our CNN predicted location-specific $R_{0_i}$ and outbreak origin in Europe were similar to that inferred in (Nadeau et al. 2021). This result and our model misspecification experiments suggest that simulation-trained deep neural networks trained on phylogenetic trees can find patterns in the training data that generalize well beyond the training data set.

Our study extends the results of Voznica et al. (2022) and Lambert et al. (2022) in several important ways. Our work showed that the new compact bijective ladderized vector encoding of phylogenetic trees can easily be extended with one-hot encoding to include metadata about viral samples. Using this strategy, we trained a neural network to not only predict important epidemiological parameters such as $R_{0_i}$ and the sampling rate, but also geographic parameters such as the migration rate and the location of outbreak origination or spillover. We anticipate that more diverse and complex metadata can be incorporated to train neural networks to make predictions about many important aspects of

763 epidemiological spread such as the relative roles of different demographic groups and the
764 overlap of different species' ranges.

765 This approach can be readily applied to numerous compartment models used to
766 describe the spread of different pathogens among different species, locations, and
767 demographic groups, e.g. SEIR, SIRS, SIS, etc. (Ponciano and Capistrán 2011; Volz and
768 Siveroni 2018; Bjørnstad et al. 2020; Chang et al. 2020; O'Dea and Drake 2021) as well as
769 modeling super-spreader dynamics as in (Voznica et al. 2022). Here we focused on one
770 phase of outbreaks (the exponential phase), but there are many other scenarios to be
771 investigated, such as when the stage of an epidemic differs among locations (e.g.
772 exponential, peaked, declining). With likelihood-free methods, the link between the
773 underlying population dynamics from which viral genomes are sampled and inferred
774 phylogenetic trees can easily be interrogated. More complex models will require larger trees
775 to infer model parameters. In this study we explored trees that contained fewer than 500
776 tips, but anticipate that larger trees will demonstrate even greater speed advantages of
777 neural networks over likelihood-based methods either through subsampling regimes
778 (Voznica et al. 2022) or by including larger trees in training datasets.

779 With fast, likelihood-free inference afforded by deep learning, the technical
780 challenges shift from exploring models for which tractable likelihood functions can be
781 derived towards models that produce realistic empirical data patterns, have parameters
782 that control variation of those patterns, and are efficient enough to generate large training
783 data sets. A growing number of advanced simulators are rapidly expanding the possibilities
784 for deep learning in phylogenetics. For example, FAVITES (Moshiri et al. 2019) is a
785 simulator of disease spread through large contact networks that tracks transmission trees
786 and simulates sequence evolution. Gen3sis, MASTER, SLiM, and VGsim are flexible
787 simulation engines for generating complex ecological, evolutionary, and disease
788 transmission simulations (Hagen et al. 2021; Vaughan and Drummond 2013; Shchur et al.
789 2022; Haller and Messer 2019; Overcast et al. 2021). Continued advances in epidemic

simulation speed and flexibility will be essential for likelihood-free methods to push the boundaries of epidemic modeling sophistication and usefulness.

There are several avenues of development still needed to realize the potential of likelihood-free inference in phylogeography using deep learning. The current setup is ideal for simulation experiments, but it is more difficult to ensure that the optimal parameter values for empirical data sets are within the range of training data parameters. Standardizing input tree height, geographical distance, and other parameters help make training data more universally applicable. Simulation-trained neural networks are often called amortized methods (Bürkner et al. 2022; Schmitt et al. 2022) because the cost of inference is front-loaded, *i.e.* it takes time to simulate a training set and train a neural network. The total cost in time per phylogenetic tree amortizes as the number of trees analyzed by the trained model increases. These methods are therefore important when a model is intended to be widely deployed or be responsive to an emerging outbreak where policy decisions must be formulated rapidly. Because amortized approximate methods require multiple analyses to realize time savings, researchers need to generate training data sets over a broad parameter and model space so that trained networks can be applied to new and diverse data sets.

Our analysis introduces a simple approach to estimate the ancestral state corresponding to the root node or stem node of a phylogeny. More sophisticated supervised learning approaches will be needed to train neural networks to predict the ancestral locations for internal nodes other than the root. The topologies and branch lengths of random phylogenies in the training and test datasets will vary from tree to tree. Our approach relies on the fact that all trees contain a root node, meaning all trees can help predict the root node's state. However, few (if any) trees in the training dataset will contain an arbitrary clade of interest within a test dataset, suggesting to us that naive approaches to train networks to estimate ancestral states for all internal nodes will probably fail. We are unaware of any existing solutions for generalized ancestral state estimation using deep

learning, and expect the problem will gather more attention as the field matures.

Quantifying uncertainty is crucial to data analysis and decision making, and Bayesian statistics provides a framework for doing so in a rigorous way. It is essential to understand how uncertainty estimation with likelihood-free methods compare to likelihood-based methods when confronted with the mismatch of models and real-world data-generating processes. We quantified uncertainty using conformalized quantile regression (CQR; Romano et al. 2019) by training neural networks to predict quantiles and then calibrating those quantiles to produce the expected coverage. We refer to the resulting intervals as CPI and demonstrate that they predict well the coverage of true values on a test dataset (Figure 3) and behave in similar ways to Bayesian methods when the model is or is not misspecified (Figures 2 - 9). Despite having the same (correct) coverage as the Bayesian HPI, the interval length was 20-50% wider on average making them a more conservative (less precise) estimation procedure. Though this can likely be improved with more training data for qCNNs, there are more fundamental challenges for uncertainty quantification with quantile regression and conformalization.

Methods for estimating more precise intervals is an active vein of research among machine learning researchers and statisticians (Barber et al. 2020; Chung et al. 2021; Sousa et al. 2022; Gibbs et al. 2023). For example, although intervals estimated by the qCNN are conditional on each data point, the calibration of quantiles through CQR involves estimating marginal calibration terms that shift all quantiles by the same amount. If the error in the quantile coverage is not constant across the prediction range, then a more adaptive procedure should yeild more precise intervals (Sousa et al. 2022; Gibbs et al. 2023).

We also compared the consistency among CPI estimates at different inner-quantiles to that of HPIs at those same quantiles. We find that independently trained neural networks for each $\alpha$ level can potentially lead to inconsistencies where narrower, nested inner quantiles can have close to or higher coverage than wider quantiles (*e.g.* Figure 9 C).

Overall, our results suggest CQR is approximately consistent with likelihood-based methods and similarly sensitive to model misspecification, while there is room for improvement. Methods where all quantiles of interest can be estimated jointly (Chung et al. 2021) may be a fruitful avenue of research for such improvements.

Another important challenge of inference with deep learning is the problem of convergence to a location on the loss function surface that approximates the maximum likelihood well. There are a number of basic heuristics that can help such as learning curves but more rigorous methods of ascertaining convergence is the subject of active research (Bürkner et al. 2022; Schmitt et al. 2022).

With recent advances in deep learning in epidemiology, evolution, and ecology (Battey et al. 2020; Schrider and Kern 2018; Voznica et al. 2022; Radev et al. 2021; Lambert et al. 2022; Rosenzweig et al. 2022; Suvorov and Schrider 2022) biologists can now explore the behavior of entire classes of stochastic branching models that are biologically interesting but mathematically or statistically prohibitive for use with traditional likelihood-based inference techniques. Beyond epidemiology, we anticipate that deep learning approaches will be useful for a wide range of currently intractable phylogenetic modeling problems. Many phylogenetic scenarios – such as the adaptive radiation of anoles (**?**) or the global spread of the grasses (**?**) – involve the evolution of discrete traits, continuous traits, speciation, and extinction within an ecological or spatial context across a set of co-evolving species. Deriving fully mechanistic yet tractable phylogenetic model likelihoods for such complex scenarios is difficult, if not impossible. Careful development and applications of likelihood-free modeling methods might bring these phylogenetic scenarios into renewed focus for more detailed study. Although we are cautiously optimistic about the future of deep learning methods for phylogenetics, it will become increasingly important for the field to diagnose the conditions where phylogenetic deep learning underperforms relative to likelihood-based approaches, and to devise general solutions to benefit the field.

# Funding

This research was supported in part by an appointment to the Department of Defense (DOD) Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and the DOD. ORISE is managed by ORAU under DOE contract number DE-SC0014664. All opinions expressed in this paper are the author's and do not necessarily reflect the policies and views of DOD, DOE, or ORAU/ORISE. MJL was supported by the National Science Foundation (DEB 2040347) and by an internal grant awarded by the Incubator for Transdisciplinary Futures at Washington University.

# Acknowledgements

We are grateful to Jacob McCord, Mark Lowell, Michael May, Fábio Mendes, Sarah Swiston, Sean McHugh, Walker Sexton, and Mariana Braga for helpful comments on the research.

Data available from the Dryad Digital Repository: https://doi.org/10.25338/B8SH2J (Thompson et al. 2023) and code is available on github: https://github.com/ammonthompson/phylogeo_epi_cnn

887 *

888 References

889 Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro,
890 Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian
891 Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal
892 Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat
893 Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens,
894 Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay
895 Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin
896 Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on
897 Heterogeneous Distributed Systems, March 2016.

898 Laith Alzubaidi, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-Dujaili, Ye Duan, Omran
899 Al-Shamma, J. Santamaría, Mohammed A. Fadhel, Muthana Al-Amidie, and Laith
900 Farhan. Review of deep learning: Concepts, CNN architectures, challenges, applications,
901 future directions. *Journal of Big Data*, 8(1):53, 2021. ISSN 2196-1115. doi:
902 10.1186/s40537-021-00444-8.

903 Roy M Anderson and Robert M May. Population biology of infectious diseases: Part i.
904 *Nature*, 280(5721):361–367, 1979.

905 Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and
906 Tijana Zrnic. Prediction-Powered Inference, February 2023.

907 Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The
908 limits of distribution-free conditional predictive inference, April 2020.

909 CJ Battey, Peter L Ralph, and Andrew D Kern. Predicting geographic location from
910 genetic variation with deep neural networks. *eLife*, 9:e54507, June 2020. ISSN
911 2050-084X. doi: 10.7554/eLife.54507.

912 Jeremy M. Beaulieu and Brian C. O'Meara. Detecting Hidden Diversification Shifts in
913 Models of Trait-Dependent Speciation and Extinction. *Systematic Biology*, 65(4):
914 583–601, July 2016. ISSN 1063-5157, 1076-836X. doi: 10.1093/sysbio/syw022.

915 Ottar N. Bjørnstad, Katriona Shea, Martin Krzywinski, and Naomi Altman. The SEIRS
916 model for infectious disease dynamics. *Nature Methods*, 17(6):557–558, June 2020. ISSN
917 1548-7091, 1548-7105. doi: 10.1038/s41592-020-0856-2.

918 Folmer Bokma. Artificial neural networks can learn to estimate extinction rates from
919 molecular phylogenies. *Journal of theoretical biology*, 243(3):449–454, 2006.

920 Remco Bouckaert, Timothy G. Vaughan, Joëlle Barido-Sottani, Sebastián Duchêne,
921 Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, Graham Jones, Denise
922 Kühnert, Nicola De Maio, Michael Matschiner, Fábio K. Mendes, Nicola F. Müller,
923 Huw A. Ogilvie, Louis du Plessis, Alex Popinga, Andrew Rambaut, David Rasmussen,
924 Igor Siveroni, Marc A. Suchard, Chieh-Hsi Wu, Dong Xie, Chi Zhang, Tanja Stadler, and
925 Alexei J. Drummond. BEAST 2.5: An advanced software platform for Bayesian
926 evolutionary analysis. *PLOS Computational Biology*, 15(4):e1006650, April 2019. ISSN
927 1553-7358. doi: 10.1371/journal.pcbi.1006650.

928 Paul-Christian Bürkner, Maximilian Scholz, and Stefan Radev. Some models are useful,
929 but how do we know which ones? Towards a unified Bayesian model taxonomy,
930 September 2022.

931 Sheryl L. Chang, Mahendra Piraveenan, Philippa Pattison, and Mikhail Prokopenko.
932 Game theoretic modelling of infectious disease dynamics and intervention methods: A
933 review. *Journal of Biological Dynamics*, 14(1):57–89, January 2020. ISSN 1751-3758,
934 1751-3766. doi: 10.1080/17513758.2020.1720322.

935 F. K. Chollet. Keras: The Python deep learning API. https://keras.io/.

Youngseog Chung, Willie Neiswanger, Ian Char, and Jeff Schneider. Beyond Pinball Loss: Quantile Methods for Calibrated Uncertainty Quantification, December 2021.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, December 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1912789117.

Emanuel Masiero da Fonseca, Guarino R. Colli, Fernanda P. Werneck, and Bryan C. Carstens. Phylogeographic model selection using convolutional neural networks, September 2020.

Jordan Douglas, Fábio K Mendes, Remco Bouckaert, Dong Xie, Cinthy L Jiménez-Silva, Christiaan Swanepoel, Joep de Ligt, Xiaoyun Ren, Matt Storey, James Hadfield, Colin R Simpson, Jemma L Geoghegan, Alexei J Drummond, and David Welch. Phylodynamics reveals the role of human travel and contact tracing in controlling the first wave of COVID-19 in four island nations. *Virus Evolution*, 7(2), September 2021. ISSN 2057-1577. doi: 10.1093/ve/veab052.

Richard G. FitzJohn. Diversitree: Comparative phylogenetic analyses of diversification in R. *Methods in Ecology and Evolution*, 3(6):1084–1092, 2012. ISSN 2041-210X. doi: 10.1111/j.2041-210X.2012.00234.x.

Lex Flagel, Yaniv Brandvain, and Daniel R Schrider. The Unreasonable Effectiveness of Convolutional Neural Networks in Population Genetic Inference. *Molecular Biology and Evolution*, 36(2):220–238, February 2019. ISSN 0737-4038, 1537-1719. doi: 10.1093/molbev/msy224.

Jiansi Gao, Michael R May, Bruce Rannala, and Brian R Moore. New Phylogenetic Models Incorporating Interval-Specific Dispersal Dynamics Improve Inference of Disease Spread. *Molecular Biology and Evolution*, 39(8):msac159, August 2022. ISSN 1537-1719. doi: 10.1093/molbev/msac159.

961 Jiansi Gao, Michael R. May, Bruce Rannala, and Brian R. Moore. Model misspecification
962 misleads inference of the spatial dynamics of disease outbreaks. *Proceedings of the*
963 *National Academy of Sciences*, 120(11):e2213913120, March 2023. doi:
964 10.1073/pnas.2213913120.

965 Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. Conformal Prediction With
966 Conditional Guarantees, May 2023.

967 James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton
968 Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. Nextstrain: real-time
969 tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, December 2018. ISSN
970 1367-4803. doi: 10.1093/bioinformatics/bty407. URL
971 `https://doi.org/10.1093/bioinformatics/bty407`.

972 Oskar Hagen, Benjamin Flück, Fabian Fopp, Juliano S. Cabral, Florian Hartig, Mikael
973 Pontarp, Thiago F. Rangel, and Loïc Pellissier. Gen3sis: A general engine for
974 eco-evolutionary simulations of the processes that shape Earth's biodiversity. *PLOS*
975 *Biology*, 19(7):e3001340, July 2021. ISSN 1545-7885. doi: 10.1371/journal.pbio.3001340.

976 Benjamin C Haller and Philipp W Messer. SLiM 3: Forward Genetic Simulations Beyond
977 the Wright–Fisher Model. *Molecular Biology and Evolution*, 36(3):632–637, March 2019.
978 ISSN 0737-4038. doi: 10.1093/molbev/msy228.

979 Sebastian Höhna, Michael J. Landis, Tracy A. Heath, Bastien Boussau, Nicolas Lartillot,
980 Brian R. Moore, John P. Huelsenbeck, and Fredrik Ronquist. RevBayes: Bayesian
981 Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification
982 Language. *Systematic Biology*, 65(4):726–736, July 2016. ISSN 1063-5157, 1076-836X.
983 doi: 10.1093/sysbio/syw021.

984 Eddie C Holmes and Geoff P Garnett. Genes, trees and infections: molecular evidence in
985 epidemiology. *Trends in Ecology & Evolution*, 9(7):256–260, 1994.

Eddie C Holmes, Sean Nee, Andrew Rambaut, Geoff P Garnett, and Paul H Harvey. Revealing the history of infectious disease epidemics through phylogenetic trees. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 349(1327):33–40, 1995.

Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53(8):5455–5516, December 2020. ISSN 1573-7462. doi: 10.1007/s10462-020-09825-6.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017.

Denise Kühnert, Tanja Stadler, Timothy G. Vaughan, and Alexei J. Drummond. Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth–death SIR model. *Journal of The Royal Society Interface*, 11(94):20131106, May 2014. ISSN 1742-5689, 1742-5662. doi: 10.1098/rsif.2013.1106.

Denise Kühnert, Tanja Stadler, Timothy G. Vaughan, and Alexei J. Drummond. Phylodynamics with Migration: A Computational Framework to Quantify Population Structure from Genomic Data. *Molecular Biology and Evolution*, 33(8):2102–2116, August 2016. ISSN 0737-4038. doi: 10.1093/molbev/msw064.

Sophia Lambert, Jakub Voznica, and Hélène Morlon. Deep Learning from Phylogenies for Diversification Analyses, September 2022.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, 113(523):1094–1111, July 2018. ISSN 0162-1459. doi: 10.1080/01621459.2017.1307116.

Philippe Lemey, Andrew Rambaut, Alexei J. Drummond, and Marc A. Suchard. Bayesian

Phylogeography Finds Its Roots. *PLoS Computational Biology*, 5(9):e1000520, September 2009. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000520.

Philippe Lemey, Nick Ruktanonchai, Samuel L. Hong, Vittoria Colizza, Chiara Poletto, Frederik Van den Broeck, Mandev S. Gill, Xiang Ji, Anthony Levasseur, Bas B. Oude Munnink, Marion Koopmans, Adam Sadilek, Shengjie Lai, Andrew J. Tatem, Guy Baele, Marc A. Suchard, and Simon Dellicour. Untangling introductions and persistence in COVID-19 resurgence in Europe. *Nature*, June 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-03754-2.

Frédéric Lemoine and Olivier Gascuel. Gotree/Goalign: Toolkit and Go API to facilitate the development of phylogenetic workflows. *NAR Genomics and Bioinformatics*, 3(3): lqab075, September 2021. ISSN 2631-9268. doi: 10.1093/nargab/lqab075.

Wayne P. Maddison, Peter E. Midford, and Sarah P. Otto. Estimating a Binary Character's Effect on Speciation and Extinction. *Systematic Biology*, 56(5):701–710, October 2007. ISSN 1076-836X, 1063-5157. doi: 10.1080/10635150701607033.

Mike Meredith and John Kruschke. Bayesian Estimation Supersedes the t-Test. page 13.

Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5):1530–1534, May 2020. ISSN 0737-4038. doi: 10.1093/molbev/msaa015.

Niema Moshiri, Manon Ragonnet-Cronin, Joel O Wertheim, and Siavash Mirarab. FAVITES: Simultaneous simulation of transmission networks, phylogenetic trees and sequences. *Bioinformatics*, 35(11):1852–1861, June 2019. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bty921.

Sarah A. Nadeau, Timothy G. Vaughan, Jérémie Scire, Jana S. Huisman, and Tanja Stadler. The origin and early spread of SARS-CoV-2 in Europe. *Proceedings of the*

*National Academy of Sciences*, 118(9):e2012008118, March 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2012008118.

Luca Nesterenko, Bastien Boussau, and Laurent Jacob. Phyloformer: Towards fast and accurate phylogeny estimation with self-attention networks, June 2022.

Eamon B O'Dea and John M Drake. A semi-parametric, state-space compartmental model with time-dependent parameters for forecasting COVID-19 cases, hospitalizations, and deaths. page 32, 2021.

Isaac Overcast, Megan Ruffley, James Rosindell, Luke Harmon, Paulo AV Borges, Brent C Emerson, Rampal S Etienne, Rosemary Gillespie, Henrik Krehenwinkel, D Luke Mahler, et al. A unified model of species abundance, genetic diversity, and functional diversity reveals the mechanisms structuring ecological communities. *Molecular Ecology Resources*, 21(8):2782–2800, 2021.

Jonathan E. Pekar, Andrew Magee, Edyth Parker, Niema Moshiri, Katherine Izhikevich, Jennifer L. Havens, Karthik Gangavarapu, Lorena Mariana Malpica Serrano, Alexander Crits-Christoph, Nathaniel L. Matteson, Mark Zeller, Joshua I. Levy, Jade C. Wang, Scott Hughes, Jungmin Lee, Heedo Park, Man-Seong Park, Katherine Zi Yan Ching, Raymond Tzer Pin Lin, Mohd Noor Mat Isa, Yusuf Muhammad Noor, Tetyana I. Vasylyeva, Robert F. Garry, Edward C. Holmes, Andrew Rambaut, Marc A. Suchard, Kristian G. Andersen, Michael Worobey, and Joel O. Wertheim. The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2. *Science*, 0(0):eabp8337, July 2022. doi: 10.1126/science.abp8337.

José M. Ponciano and Marcos A. Capistrán. First Principles Modeling of Nonlinear Incidence Rates in Seasonal Epidemics. *PLOS Computational Biology*, 7(2):e1001079, February 2011. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1001079.

O. G. Pybus, M. A. Suchard, P. Lemey, F. J. Bernardin, A. Rambaut, F. W. Crawford,

R. R. Gray, N. Arinaminpathy, S. L. Stramer, M. P. Busch, and E. L. Delwart. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proceedings of the National Academy of Sciences*, 109(37):15066–15071, September 2012. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1206598109.

Stefan T. Radev, Frederik Graw, Simiao Chen, Nico T. Mutters, Vanessa M. Eichel, Till Bärnighausen, and Ullrich Köthe. OutbreakFlow: Model-based Bayesian inference of disease outbreak dynamics with invertible neural networks and its application to the COVID-19 pandemics in Germany. *PLOS Computational Biology*, 17(10):e1009472, October 2021. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1009472.

A. Rambaut and N. C. Grassly. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences*, 13:235–238, 1997.

Andrew Rambaut, Oliver G Pybus, Martha I Nelson, Cecile Viboud, Jeffery K Taubenberger, and Edward C Holmes. The genomic and epidemiological dynamics of human influenza a virus. *Nature*, 453(7195):615–619, 2008.

Liam J. Revell. Phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2):217–223, 2012. ISSN 2041-210X. doi: 10.1111/j.2041-210X.2011.00169.x.

Francisco Richter, Bart Haegeman, Rampal S. Etienne, and Ernst C. Wit. Introducing a general class of species diversification models for phylogenetic trees. *Statistica Neerlandica*, 74(3):261–274, 2020. ISSN 1467-9574. doi: 10.1111/stan.12205.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized Quantile Regression. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Benjamin K. Rosenzweig, Matthew W. Hahn, and Andrew Kern. Accurate Detection of Incomplete Lineage Sorting via Supervised Machine Learning, November 2022.

Marvin Schmitt, Paul-Christian Bürkner, Ullrich Köthe, and Stefan T. Radev. Detecting Model Misspecification in Amortized Bayesian Inference with Neural Networks, May 2022.

Daniel R. Schrider and Andrew D. Kern. Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends in Genetics*, 34(4):301–312, April 2018. ISSN 01689525. doi: 10.1016/j.tig.2017.12.005.

Jérémie Scire, Joëlle Barido-Sottani, Denise Kühnert, Timothy G. Vaughan, and Tanja Stadler. Improved multi-type birth-death phylodynamic inference in BEAST 2. Preprint, Evolutionary Biology, January 2020.

Vladimir Shchur, Vadim Spirin, Dmitry Sirotkin, Evgeni Burovski, Nicola De Maio, and Russell Corbett-Detig. VGsim: Scalable viral genealogy simulator for global pandemic. *PLOS Computational Biology*, 18(8):e1010409, August 2022. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1010409.

Claudia Solis-Lemus, Shengwen Yang, and Leonardo Zepeda-Nunez. Accurate Phylogenetic Inference with a Symmetry-preserving Neural Network Model, January 2022.

Martim Sousa, Ana Maria Tomé, and José Moreira. Improved conformalized quantile regression, November 2022.

Tanja Stadler. Sampling-through-time in birth–death trees. *Journal of Theoretical Biology*, 267(3):396–404, December 2010. ISSN 00225193. doi: 10.1016/j.jtbi.2010.09.010.

Tanja Stadler, Roger Kouyos, Viktor von Wyl, Sabine Yerly, Jürg Böni, Philippe Bürgisser, Thomas Klimkait, Beda Joos, Philip Rieder, Dong Xie, Huldrych F. Günthard, Alexei J. Drummond, Sebastian Bonhoeffer, and the Swiss HIV Cohort Study. Estimating the

Basic Reproductive Number from Viral Sequence Data. *Molecular Biology and Evolution*, 29(1):347–357, January 2012. ISSN 1537-1719, 0737-4038. doi: 10.1093/molbev/msr217.

Ingo Steinwart and Andreas Christmann. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225, February 2011. ISSN 1350-7265. doi: 10.3150/10-BEJ267.

Anton Suvorov and Daniel R. Schrider. Reliable estimation of tree branch lengths using deep neural networks. *bioRxiv*, 2022. doi: 10.1101/2022.11.07.515518. URL `https://www.biorxiv.org/content/early/2023/02/21/2022.11.07.515518`.

Anton Suvorov, Joshua Hochuli, and Daniel R Schrider. Accurate Inference of Tree Topologies from Multiple Sequence Alignments Using Deep Learning. *Systematic Biology*, 69(2):221–233, March 2020. ISSN 1063-5157, 1076-836X. doi: 10.1093/sysbio/syz060.

Ammon Thompson, Benjamin Liebeskind, Erik J. Scully, and Michael J. Landis. Deep learning phylogeography. *Dryad*, 2023. doi: 10.25338/B8SH2J.

Timothy G. Vaughan and Alexei J. Drummond. A Stochastic Simulator of Birth–Death Master Equations with Application to Phylodynamics. *Molecular Biology and Evolution*, 30(6):1480–1493, June 2013. ISSN 0737-4038. doi: 10.1093/molbev/mst057.

Timothy G. Vaughan, Denise Kühnert, Alex Popinga, David Welch, and Alexei J. Drummond. Efficient Bayesian inference under the structured coalescent. *Bioinformatics*, 30(16):2272–2279, August 2014. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btu201.

Erik M. Volz and Igor Siveroni. Bayesian phylodynamic inference with complex models. *PLOS Computational Biology*, 14(11):e1006546, November 2018. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006546.

Erik M. Volz, Katia Koelle, and Trevor Bedford. Viral Phylodynamics. *PLOS Computational Biology*, 9(3):e1002947, March 2013. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002947. URL `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002947`. Publisher: Public Library of Science.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. Conformal Prediction: General Case and Regression. In Vladimir Vovk, Alexander Gammerman, and Glenn Shafer, editors, *Algorithmic Learning in a Random World*, pages 19–69. Springer International Publishing, Cham, 2022. ISBN 978-3-031-06649-8. doi: 10.1007/978-3-031-06649-8_2.

J. Voznica, A. Zhukova, V. Boskova, E. Saulnier, F. Lemoine, M. Moslonka-Lefebvre, and O. Gascuel. Deep learning from phylogenies to uncover the epidemiological dynamics of outbreaks. *Nature Communications*, 13(1):3896, July 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-31511-0.

Nicole L. Washington, Karthik Gangavarapu, Mark Zeller, Alexandre Bolze, Elizabeth T. Cirulli, Kelly M. Schiabor Barrett, Brendan B. Larsen, Catelyn Anderson, Simon White, Tyler Cassens, Sharoni Jacobs, Geraint Levan, Jason Nguyen, Jimmy M. Ramirez, Charlotte Rivera-Garcia, Efren Sandoval, Xueqing Wang, David Wong, Emily Spencer, Refugio Robles-Sikisaka, Ezra Kurzban, Laura D. Hughes, Xianding Deng, Candace Wang, Venice Servellita, Holly Valentine, Peter De Hoff, Phoebe Seaver, Shashank Sathe, Kimberly Gietzen, Brad Sickler, Jay Antico, Kelly Hoon, Jingtao Liu, Aaron Harding, Omid Bakhtar, Tracy Basler, Brett Austin, Duncan MacCannell, Magnus Isaksson, Phillip G. Febbo, David Becker, Marc Laurent, Eric McDonald, Gene W. Yeo, Rob Knight, Louise C. Laurent, Eileen de Feo, Michael Worobey, Charles Y. Chiu, Marc A. Suchard, James T. Lu, William Lee, and Kristian G. Andersen. Emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States. *Cell*, 184(10):2587–2594.e7, May 2021. ISSN 00928674. doi: 10.1016/j.cell.2021.03.052.

Michael Worobey, Thomas D Watts, Richard A McKay, Marc A Suchard, Timothy Granade, Dirk E Teuwen, Beryl A Koblin, Walid Heneine, Philippe Lemey, and Harold W Jaffe. 1970s and 'patient 0'hiv-1 genomes illuminate early hiv/aids history in north america. *Nature*, 539(7627):98–101, 2016.

Michael Worobey, Jonathan Pekar, Brendan B. Larsen, Martha I. Nelson, Verity Hill, Jeffrey B. Joy, Andrew Rambaut, Marc A. Suchard, Joel O. Wertheim, and Philippe Lemey. The emergence of SARS-CoV-2 in Europe and North America. *Science*, 370 (6516):564–570, October 2020. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abc8169.

# Supplemental Tables

Table S1: BEST comparisons between CNN and Bayesian absolute percent errors (APEs) for model parameters across all experiments.

| 95%HPD intervals of average relative error from BEST analysis | | | |
|---|---|---|---|
| True inference model (Reference for misspecication experiments) | CNN APE | Posterior mean APE | CNN APE - Posterior mean APE |
| $R_0$ | 2.4, 3.5 | 2.1, 3.1 | 0.1,1.2 |
| $\delta$ | 7.0, 10.5 | 5.7, 8.9 | 0.2, 3.0 |
| m | 9.5, 14.1 | 8.4, 12.1 | 0.4, 3.2 |
| | | | |
| Misspecied $R_0$ experiment | CNN APE - CNN Reference APE | Posterior mean APE - Post. mean Reference APE | CNN APE - Posterior mean APE |
| $R_0$ | 11.8, 17.8 | 11.0, 16.9 | -0.1, 1.6 |
| $\delta$ | 0.8, 7.6 | -0.6, 5.3 | 1.3, 5.8 |
| m | 8.2, 17.9 | 6.5, 15.9 | 1.3, 4.7 |
| | | | |
| Misspecied sample rate experiment | CNN APE - CNN Reference APE | Posterior mean APE - Post. mean Reference APE | CNN APE - Posterior mean APE |
| $R_0$ | -0.3, 1.7 | 0.03, 1.7 | 0.1, 1.3 |
| $\delta$ | 12.0, 21.2 | 12.6, 21.4 | 0.1, 4.0 |
| m | 3.3, 12.0 | 5.6, 14.4 | -1.2, 2.7 |
| | | | |
| Misspecied migration rate experiment | CNN APE - CNN Reference APE | Posterior mean APE - Post. mean Reference APE | CNN APE - Posterior mean APE |
| $R_0$ | -0.9, 0.8 | -0.6, 1.0 | -0.5, 0.8 |
| $\delta$ | -2.3, 3.3 | 0.1, 5,8 | -1.4, 2.3 |
| m | 4.0, 15.2 | 5.0, 16.2 | -1.3, 2.6 |
| | | | |
| Misspecied number of locations experiment | CNN APE - CNN Reference APE | Posterior mean APE - Post. mean Reference APE | CNN APE - Posterior mean APE |
| $R_0$ | -0.3, 1.5 | -0.7, 0.8 | 0.5, 1.9 |
| $\delta$ | -0.3, 4.9 | -0.5, 4.2 | 0.4, 3.5 |
| m | 3.4, 11.1 | 5.8, 13.5 | -0.9, 1.6 |
| | | | |
| Phylogenetic error experiment | CNN APE - CNN Reference APE | Posterior mean APE - Post. mean Reference APE | CNN APE - Posterior mean APE |
| $R_0$ | 0.7, 3.0 | 1.7, 4.4 | -1.4, 0.1 |
| $\delta$ | 2.3, 9.6 | 1.5, 7.2 | 1.4, 5.3 |
| m | -1.2, 6.0 | -1.8, 5.4 | -1.7, 2.4 |

Table S2: Comparison 95% CPI and HPI for all experiments.

| Coverage, width, and overlap of 95% Intervals | $R_0$ | | | | δ | | | | m | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CNN CPI | Bayes HPI | Mean CPI width / HPI width | Mean Jaccard index | CNN CPI | Bayes HPI | Mean CPI width / HPI width | Mean Jaccard index | CNN CPI | Bayes HPI | Mean CPI width / HPI width | Mean Jaccard index |
| **True model** | 0.95 | 0.96 | 1.4 | 0.67 | 0.96 | 0.94 | 1.4 | 0.66 | 0.94 | 0.95 | 1.2 | 0.75 |
| **Misspecified $R_0$** | 0.44 | 0.29 | 1.5 | 0.63 | 0.9 | 0.90 | 1.5 | 0.63 | 0.74 | 0.67 | 1.2 | 0.76 |
| **Misspecified δ** | 0.95 | 0.96 | 1.4 | 0.67 | 0.71 | 0.55 | 1.3 | 0.69 | 0.72 | 0.75 | 1.2 | 0.75 |
| **Misspecified m** | 0.93 | 0.94 | 1.5 | 0.63 | 0.94 | 0.96 | 1.5 | 0.65 | 0.73 | 0.69 | 1.3 | 0.73 |
| **Misspecified. Number of locations** | 0.93 | 0.92 | 1.4 | 0.65 | 0.96 | 0.96 | 1.4 | 0.68 | 0.82 | 0.80 | 1.2 | 0.76 |
| **Phylogenetic error** | 0.79 | 0.60 | 1.4 | 0.59 | 0.71 | 0.81 | 1.3 | 0.59 | 0.87 | 0.83 | 1.3 | 0.71 |

1166

# Supplemental Figures

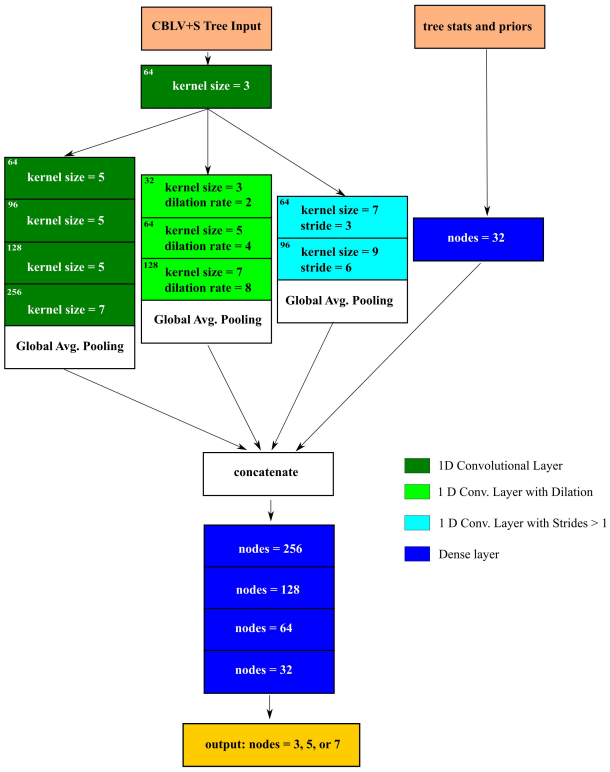


Figure S1: Diagram of deep neural network trained to make 2 kinds of predictions (rates and origin location) under two models (LIBDS and LDBDS).

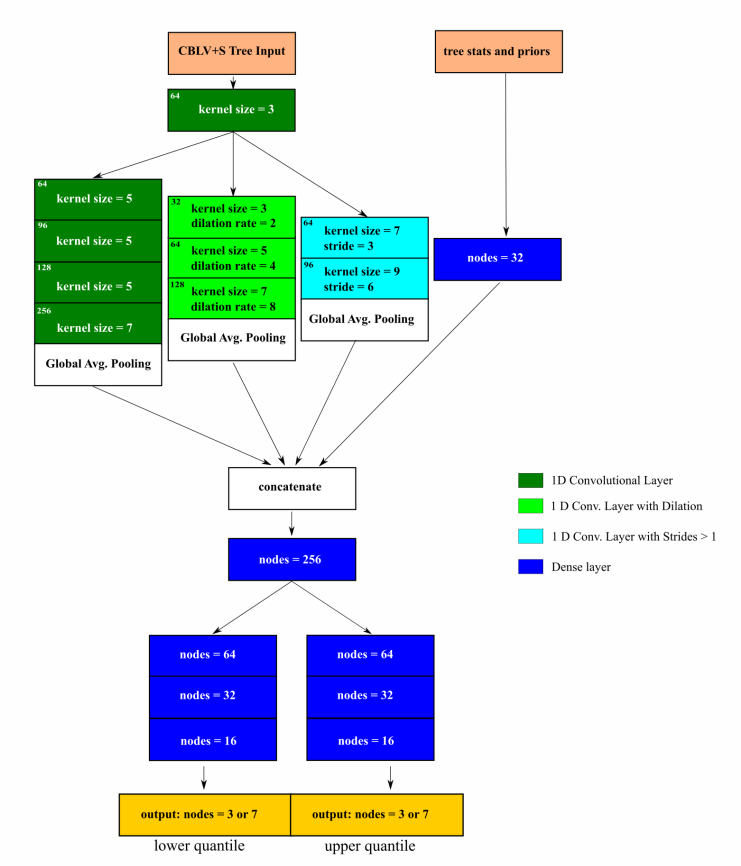


Figure S2: Diagram of deep neural network trained to predict the upper and lower quantiles for a specified $\alpha$ level under two models (LIBDS and LDBDS).

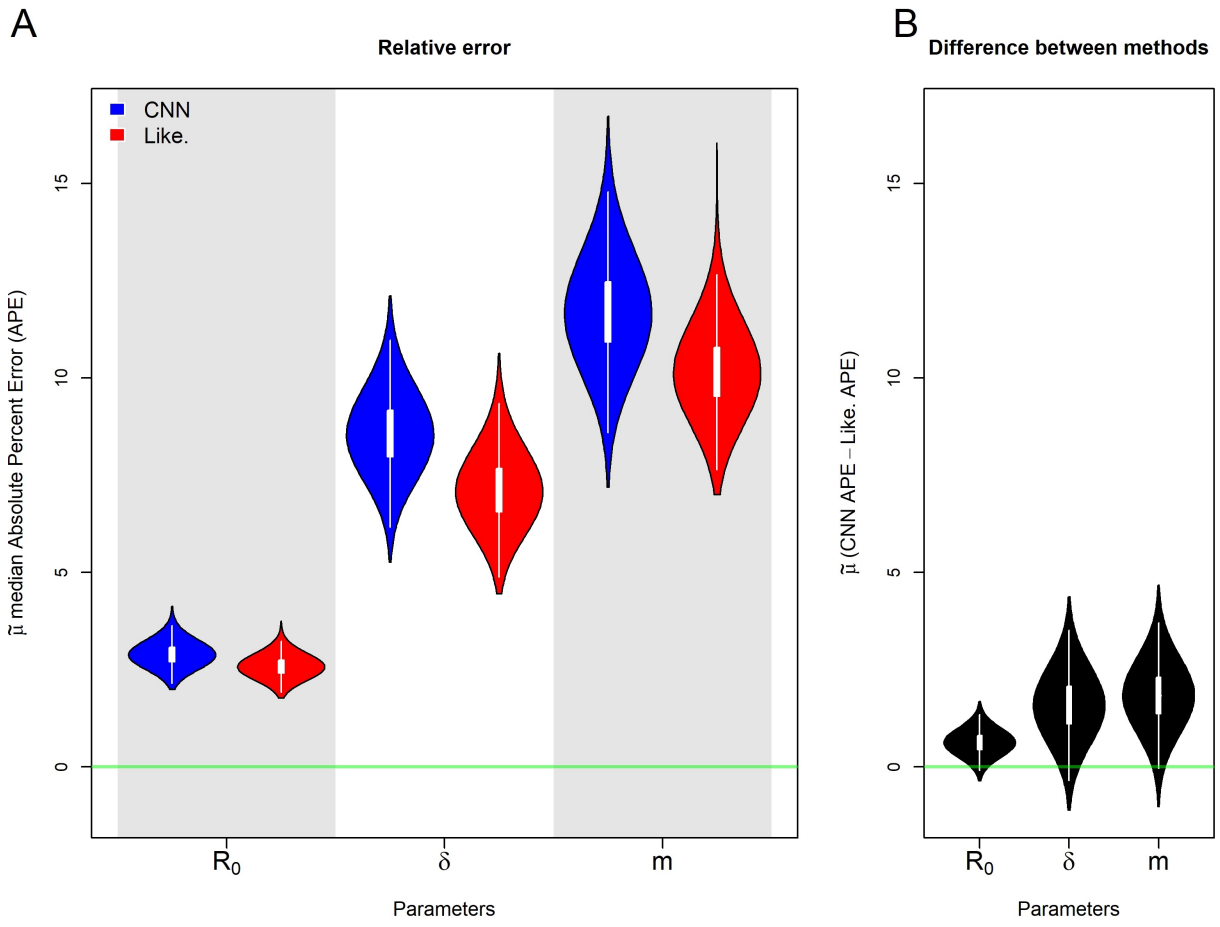


Figure S3: Posterior distributions of the population median, $\tilde{\mu}$, APE estimates of the rate parameters $R_0$, $\delta$, and m under the true model. A) shows posterior distribution of the median APE for each of the 3 rate parameters estimated by the CNN (blue) and the likelihood-based method (red). The green line indicates no error. B) shows the posterior distribution for the median difference between the CNN estimate's APE and the likelihood-based estimate's APE. The green line indicates the median APE difference is zero.

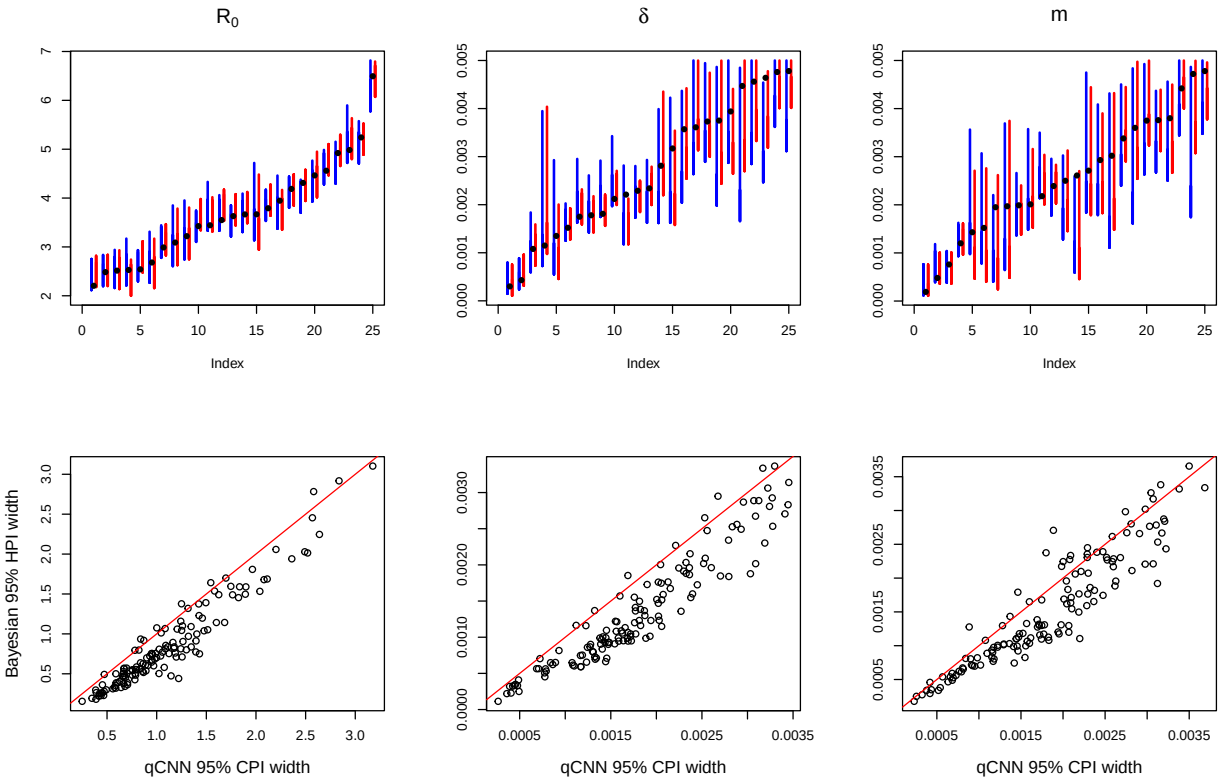


Figure S4: Comparison of interval overlap and relative widths of qCNN and Bayesian methods of uncertainty quantification under the true simulating model. Top row: 95% CPI from CNN conformalized quantile regression (blue). and 95% HPI from Bayesian phylogenetic analysis (red) from a random subset of the data for visualization purposes. Bottom row: scatterplots of the lengths of CPI and HPI intervals of all experiment data. The red diagonal y = x line is for reference.

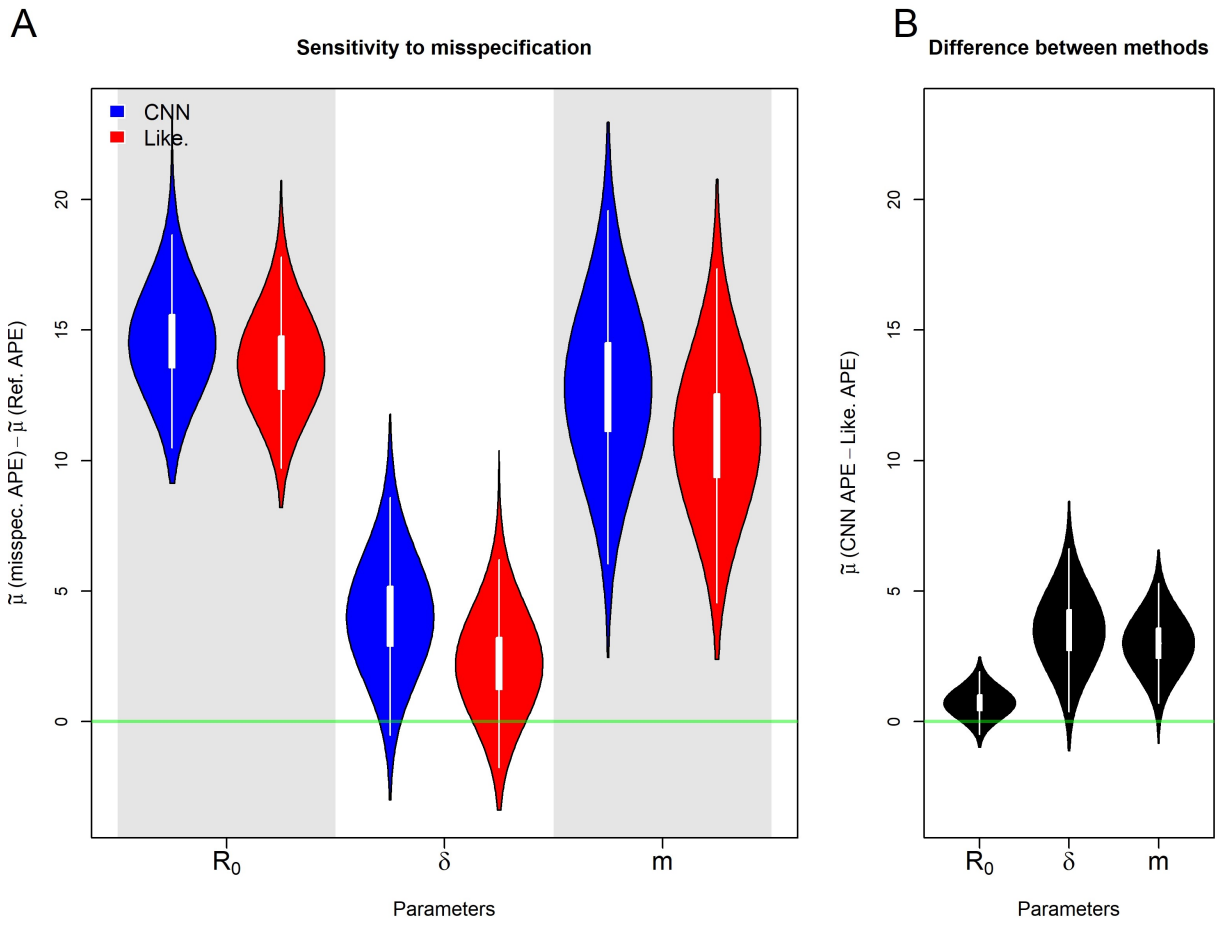


Figure S5: Posterior distributions of median, $\tilde{\mu}$, APE for the misspecified $R_0$ experiment. A) shows posterior distribution of the difference between the median error under the misspecified model and the the median error under the true, reference model. B) shows the posterior distribution for the population median difference between the CNN estimate's APE and the likelihood-based estimate's APE.

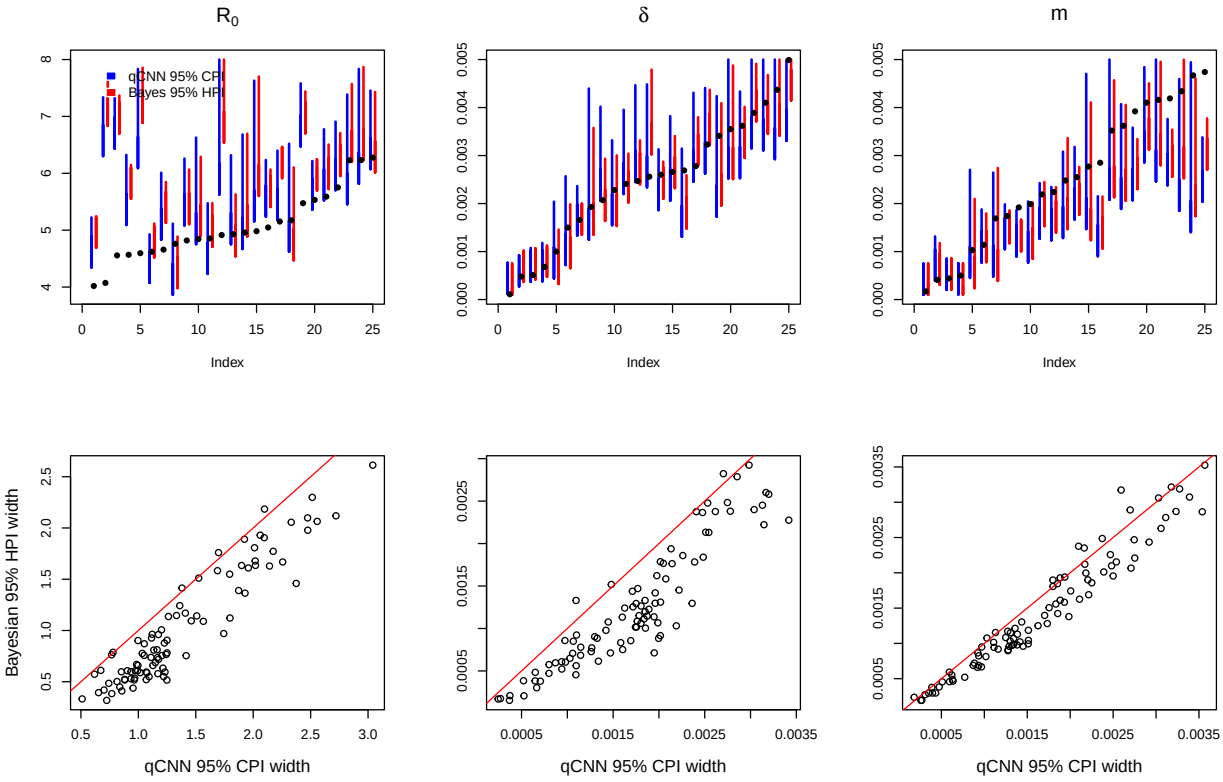


Figure S6: Comparison of CPI and HPI intervals for misspecified $R_0$ experiment. See SI Figure S4 for general details about plot.

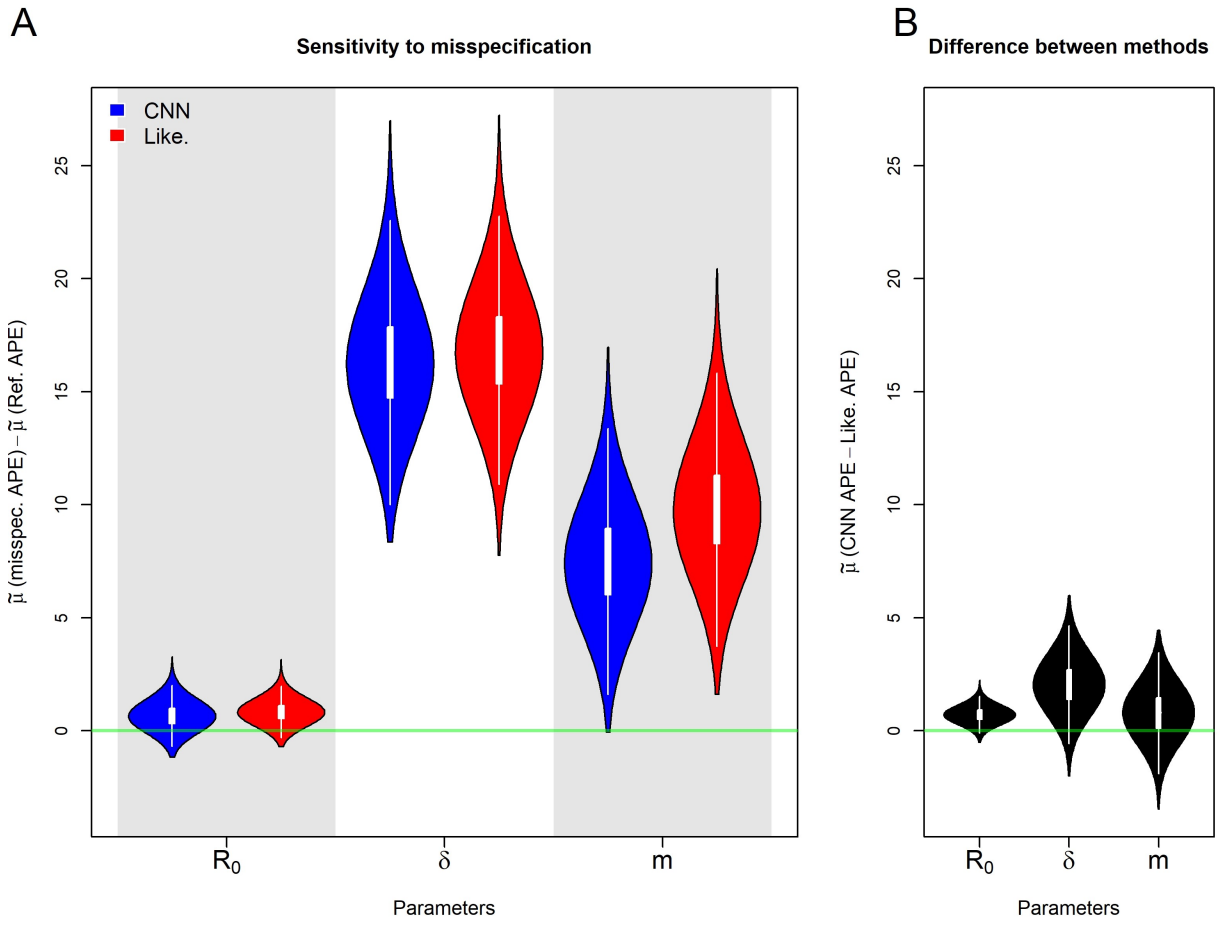


Figure S7: Posterior distributions of median, $\tilde{\mu}$, APE for the misspecified sampling rate, $\delta$, experiment. Details are the same as in S5

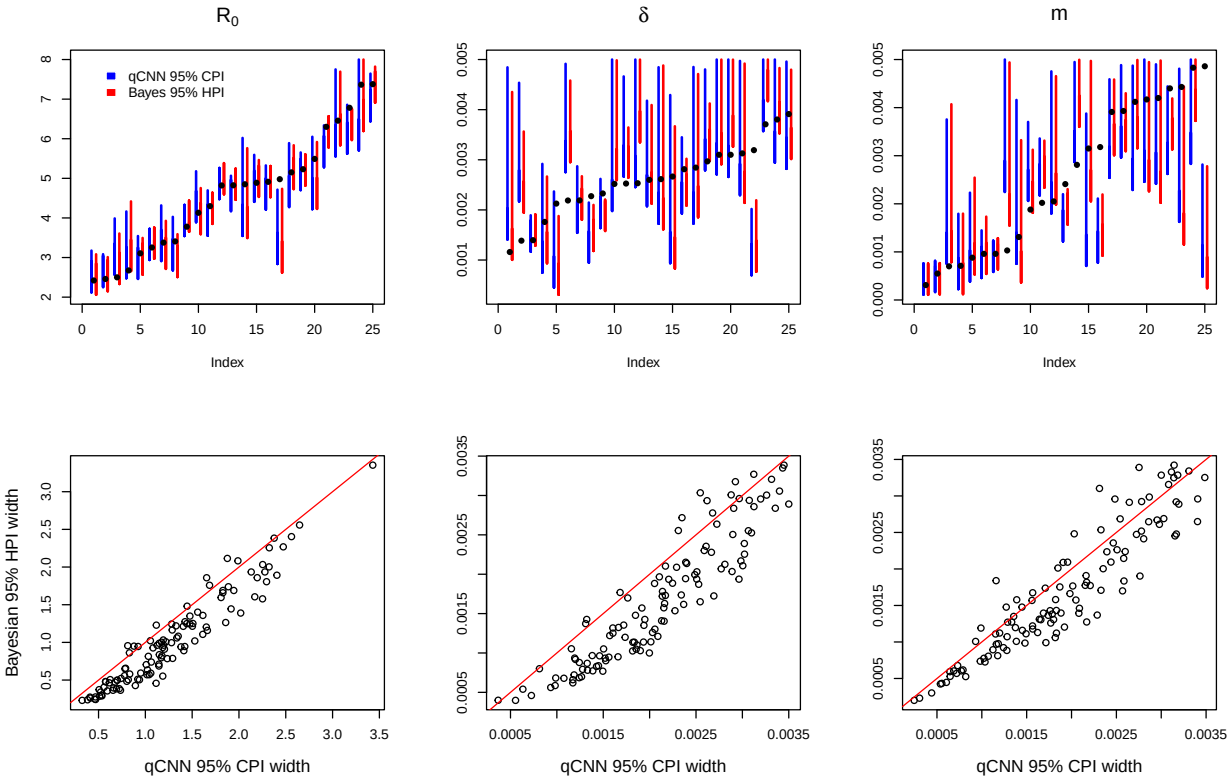


Figure S8: Comparison of CPI and HPI intervals for misspecified $\delta$ experiment. See SI Figure S4 for general details about plot.

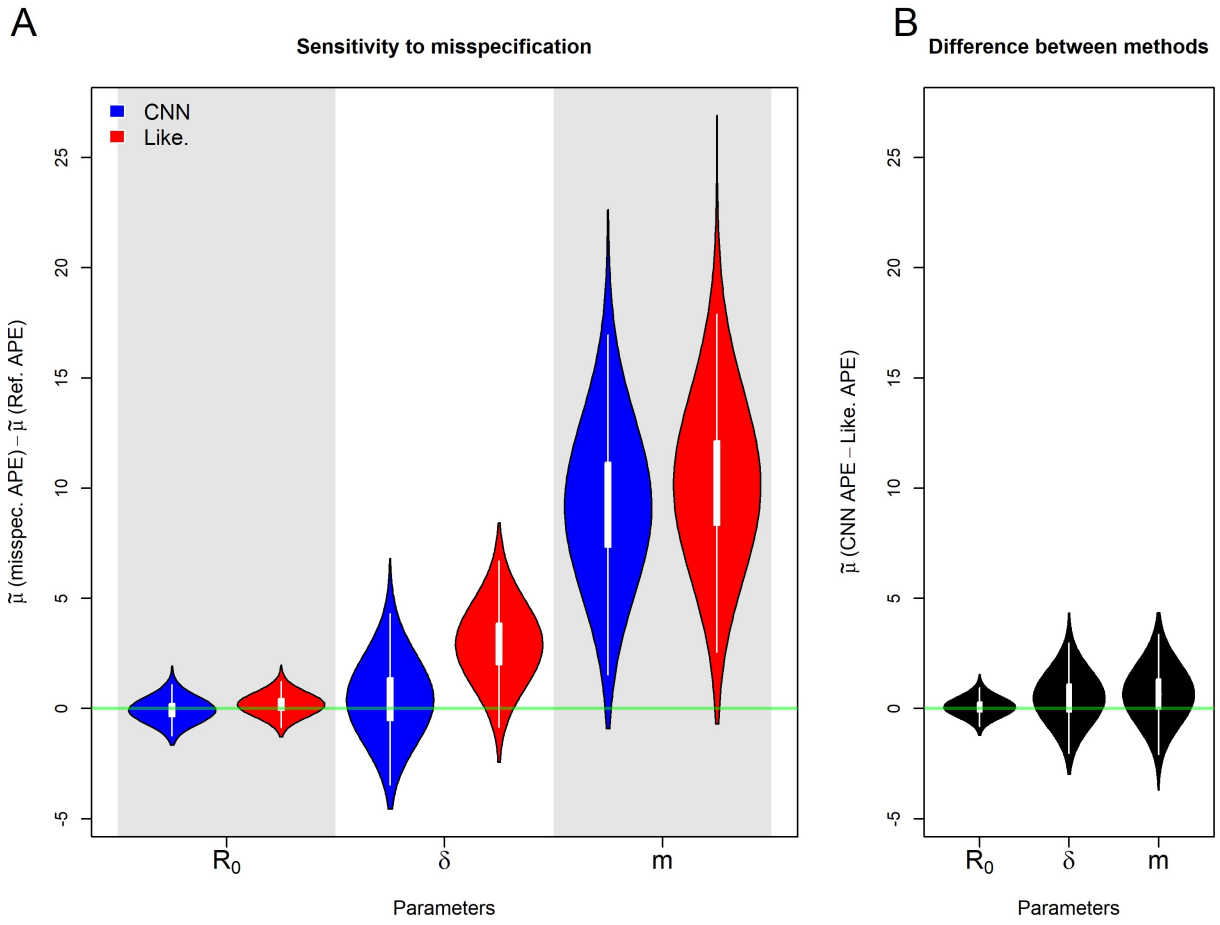


Figure S9: Posterior distributions of median, $\tilde{\mu}$, APE for the misspecified migration rate, m, experiment. Details are the same as in S5

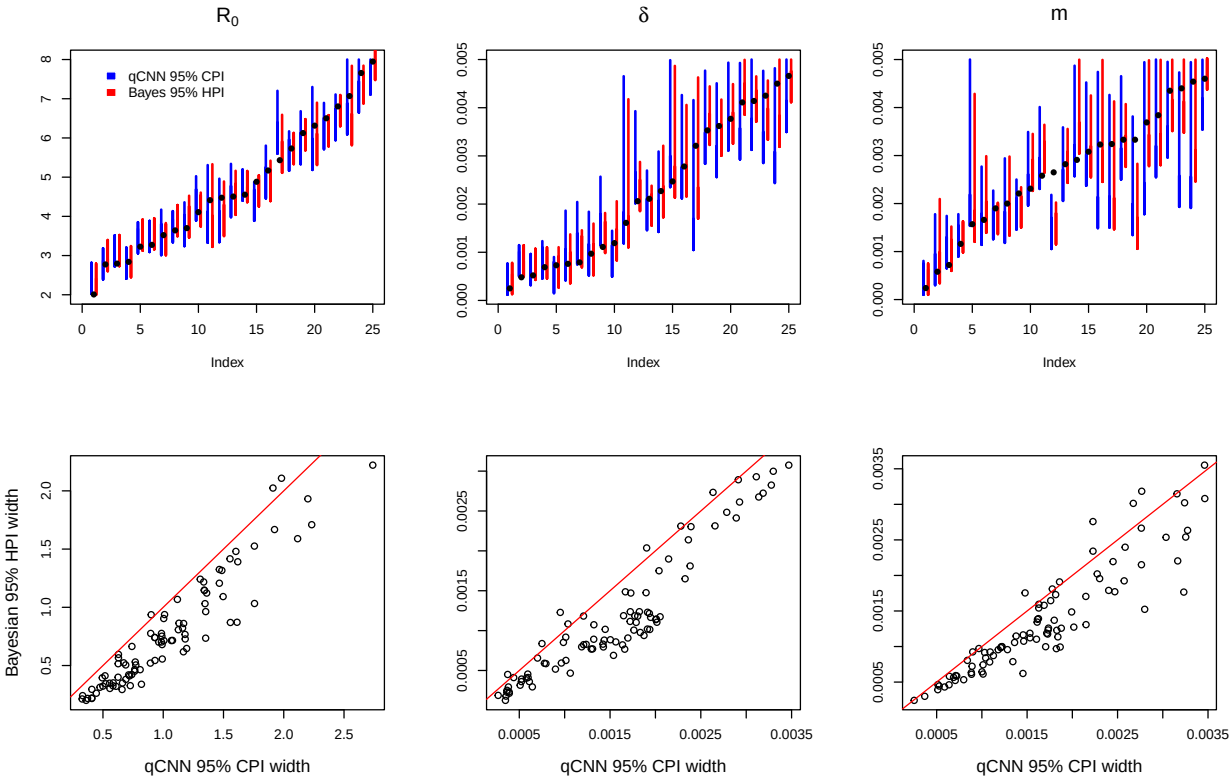


Figure S10: Comparison of CPI and HPI intervals for misspecified migration rate experiment. See SI Figure S4 for general details about plot.

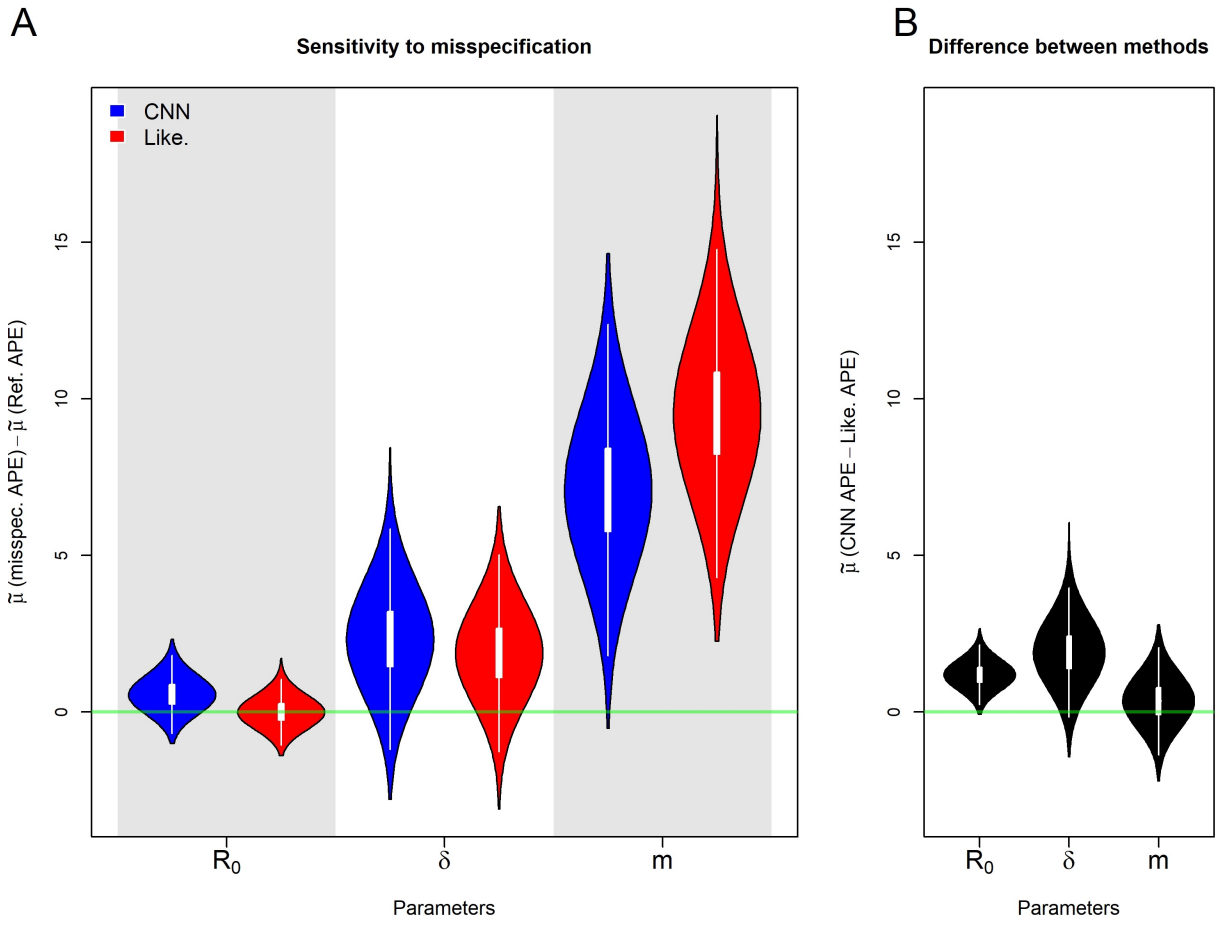


Figure S11: Posterior distributions of the median APE when the model is misspecified for the number of locations. Details are the same as in S5

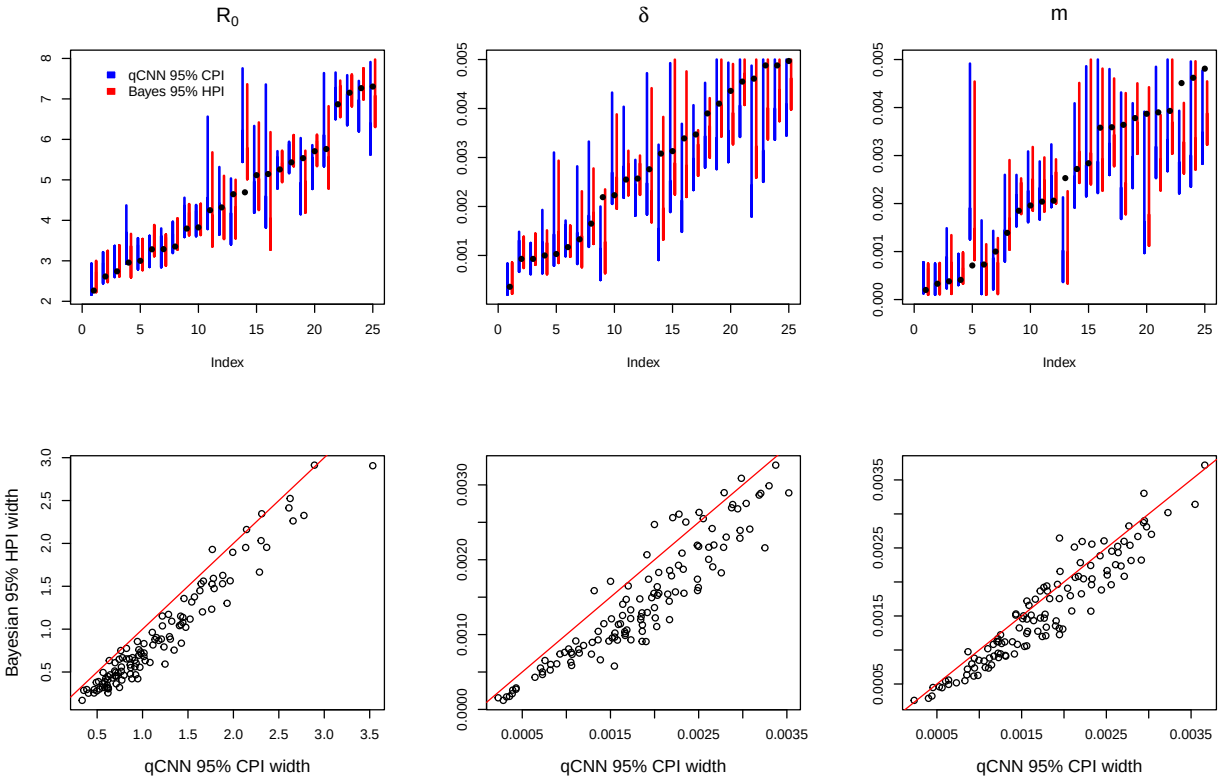


Figure S12: Comparison of CPI and HPI intervals for misspecified number of locations experiment. See SI Figure S4 for general details about plot.

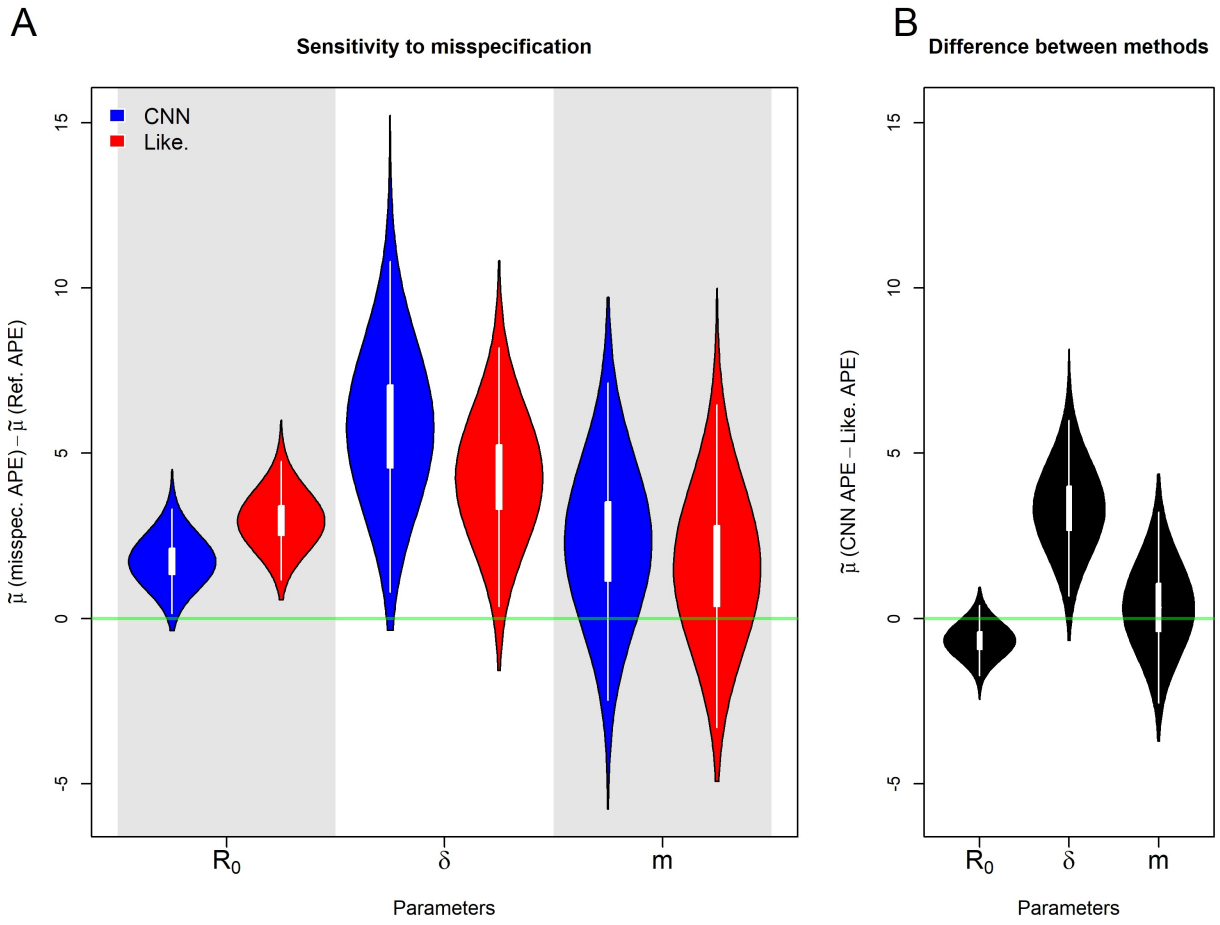


Figure S13: Posterior distributions of the median APE when the phylogenetic tree is incorrect. Details are the same as in S5

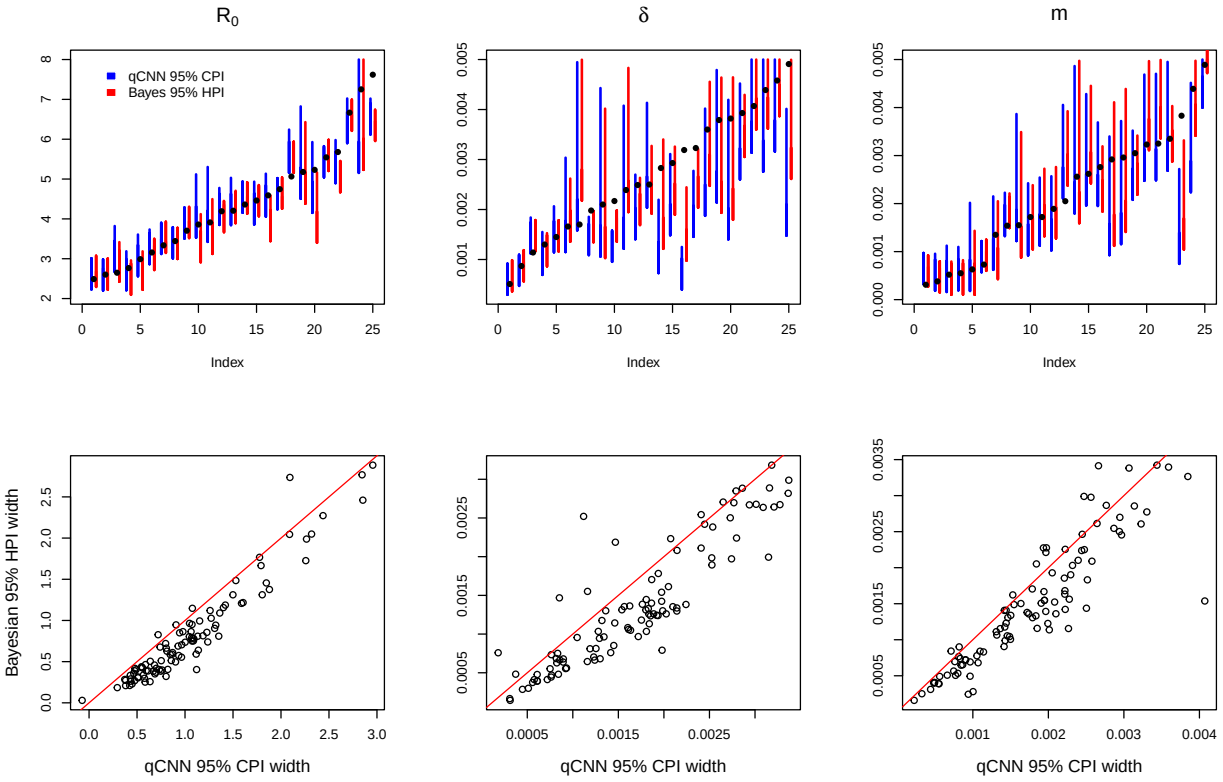


Figure S14: Comparison of CPI and HPI intervals for phylogeny error experiment. See SI Figure S4 for general details about plot.