

Toward a semi-supervised learning approach to phylogenetic estimation

DANIELE SILVESTRO^{1,2}, THIBAUT LATRILLE³, NICOLAS SALAMIN³

¹ *Department of Biology, University of Fribourg and Swiss Institute of Bioinformatics, 1700 Fribourg, Switzerland*

² *Gothenburg Global Biodiversity Centre, Department of Biological and Environmental Sciences, University of Gothenburg, 40530 Gothenburg, Sweden*

³ *Department of Computational Biology, University of Lausanne, 1015 Lausanne, Switzerland*

Correspondence to: daniele.silvestro@unifr.ch, nicolas.salamin@unil.ch

Abstract

Models have always been central to inferring molecular evolution and to reconstructing phylogenetic trees. Their use typically involves the development of a mechanistic framework reflecting our understanding of the underlying biological processes, such as nucleotide substitutions, and the estimation of model parameters by maximum likelihood or Bayesian inference. However, deriving and optimizing the likelihood of the data is not always possible under complex evolutionary scenarios or even tractable for large datasets, often leading to unrealistic simplifying assumptions in the fitted models. To overcome this issue, we coupled stochastic simulations of genome evolution with a new supervised deep learning model to infer key parameters of molecular evolution. Our model is designed to directly analyze multiple sequence alignments and estimate per-site evolutionary rates and divergence, without requiring a known phylogenetic tree. The accuracy of our predictions matched that of likelihood-based phylogenetic inference, when rate heterogeneity followed a simple gamma distribution, but it strongly exceeded it under more complex patterns of rate variation, such as codon models. Our approach is highly scalable and can be efficiently applied to genomic data, as we showed on a dataset of 26 million nucleotides from the clownfish clade. Our simulations also showed that the integration of per-site rates obtained by deep learning within a Bayesian framework led to significantly more accurate phylogenetic inference, particularly with respect to the estimated branch lengths. We thus propose that future advancements in phylogenetic analysis will benefit from a semi-supervised learning approach that combines deep-learning estimation of substitution rates, which allows for more flexible models of rate variation, and probabilistic inference of the phylogenetic tree, which guarantees interpretability and a rigorous assessment of statistical support.

Keywords: phylogenetic inference, molecular evolution, recurrent neural networks, simulations, substitution rates

Introduction

Since the seminal work by J. Felsenstein (1973) to infer phylogenetic trees by maximum likelihood, evolutionary models based on probabilistic approaches have been the central modeling framework in phylogenetics. This has led to a tremendous increase in our ability to infer evolutionary relationships, to investigate the dynamics of molecular evolution, to model the evolution of complex traits across lineages and to test evolutionary hypotheses to advance our understanding of the factors shaping the tree of life (Felsenstein, 2003; Lemey et al., 2009). While other methods based on genetic distances or parsimony criteria are still employed, for instance, to initialize phylogenetic tree inference or to provide a fast preliminary description of evolutionary processes, probabilistic approaches are widely seen as the best practice in the field.

One of the main challenges with the development of probabilistic models of evolution is to ensure that the parameters incorporated in the models are valid and identifiable when applied to real biological data. Yet, it is difficult, and in most cases even impossible, to experimentally generate data that would allow us to observe evolution in action and validate the estimates of the model parameters inferred from the outcome of such experiments. Indeed, while experimental evolution is applicable to some organisms with short generation times (e.g. for viruses or bacteria; Hillis et al., 1992; Bull et al., 1997; Lenski, 2017), simultaneously capturing the evolutionary dynamics that result in genome evolution and the traits involved in adaptation remains impossible for the most part. This means that validating evolutionary models is challenging and, when analyzing these evolutionary dynamics, we are inherently unable to test the accuracy of model predictions based on empirical measurements.

To overcome this limitation, most methods to infer evolutionary processes use simulations to assess the identifiability of model parameters and/or the robustness of the estimation. These simulations are synthetic realizations of evolutionary processes that are obtained through stochastic simulations. In a phylogenetic framework, we can use birth-death processes to generate phylogenetic trees and apply Markov models of nucleotide substitutions to simulate the evolution of a DNA sequence along the tree. Then, the same stochastic processes are typically used to also derive a likelihood-based model to estimate the generating parameter values from the simulation outcomes. For instance, we can derive the likelihood of a phylogenetic tree under a birth-death process and use it to estimate the speciation and extinction rates (Nee et al., 1994; Gernhard,

2008). We can also derive the likelihood of a DNA sequence alignment under a Markov process of evolution and use it to infer the underlying tree (Felsenstein, 1981). Comparisons between simulated and estimated parameter values (e.g., the true vs inferred phylogenetic tree) are then used to assess the accuracy of likelihood-based inference. This approach is routinely used for molecular evolution (e.g. Zaheri et al., 2014), phenotypic evolution of quantitative and discrete traits (e.g. Harmon et al., 2010; Maddison and FitzJohn, 2015), phylogenetic inference (e.g. Salamin et al., 2005), species diversification (Rabosky, 2006; Stadler, 2011), fossil preservation (Heath et al., 2014; Silvestro et al., 2019), and biogeographic inference (Landis et al., 2013; Hauffe et al., 2022).

Despite its apparent circularity, this is a robust approach to validate the ability of a likelihood-based model to recover the parameters of the generative process correctly and can be used to verify their identifiability (e.g. Ree and Sanmartín, 2018; Silvestro et al., 2018; Louca and Pennell, 2020). Further, simulations generated while violating model assumptions can help to quantify the limits of our models and the conditions where the models will fail. A potential limitation of this use of simulations is that they tend to be oversimplified realizations of the biological process, which can impact our assessments of model accuracy (Nute et al., 2019).

Beside the use of likelihood-based models to infer the biological processes of interest, there is a growing interest in machine learning approaches to detect patterns associated with evolutionary processes. The use of Deep Learning (DL) has quickly expanded into a wealth of applications across scientific fields (e.g., Jumper et al., 2021). In evolutionary biology, deep neural networks have been used in population genetics (Chan et al., 2018; Flagel et al., 2019), to infer species diversification dynamics (Silvestro et al., 2020; Lambert et al., 2023; Cooper et al., 2024), to study coevolution (see Sapoval et al., 2022, for a review), but also, for example, to infer phylogenetic trees using quartets (Zou et al., 2019; Suvorov et al., 2019; Kulikov et al., 2023), perform substitution model testing (Abadi et al., 2020) or place new samples on an existing phylogenetic tree (Jiang et al., 2022). However, the development of DL in phylogenetics and evolutionary models is restricted because empirical training datasets are scarce or cannot be generated unless we restrict our focus to fast-evolving organisms or short-term evolutionary processes. Further, DL approaches often consist of (or are viewed as) over-parameterized black-box models that do not allow a direct interpretation of the parameters, contrary to probabilistic approaches (Sapoval et al., 2022).

Although probabilistic inference and DL can be seen as very different methodologies to analyse data, there are analogies in how model validation is performed. Indeed, generative models of evolution based on stochastic simulations can be coupled with supervised DL models, just like the same simulations are used to benchmark likelihood-based models (Fig. 1a, b).

Figure 1: Schematic representation of the workflow used to estimate parameters of interest within an unsupervised model used within a likelihood framework (a) and a supervised model used with a DL framework (b). Both models use alignments of orthologous sequences of nucleotides (empirical or simulated data) to infer substitution rates and genetic distances. The simulations are used differently in the two cases, to either validate the model in a likelihood framework or to simulate training datasets in the DL framework. In panel (c) we illustrate the implementation of a semi-supervised model to infer phylogenetic trees combining per-site rates obtained from our phyloRNN DL model and Bayesian phylogenetic inference as implemented in RevBayes (Höhna et al., 2016).

In this paper, we developed a DL model to infer the rates of molecular evolution from a multiple species alignment of DNA sequences. We coupled stochastic simulations with a new supervised learning model based on recurrent neural networks and sparse networks with parameter sharing. Our model and results showed that the conceptual differences between standard unsupervised likelihood-based models and supervised DL are smaller than generally assumed in the context of molecular evolution. Predictions of site-specific substitution rates were robust across a range of evolutionary scenarios, with accuracy matching or exceeding that of state-of-the-art likelihood estimations. Our approach can efficiently analyze millions of sites to compare evolutionary rates at genomic scales. We showed that the predicted rates can improve the likelihood-based estimation of phylogenetic trees, indicating that the application of this approach might have broader implications in phylogenetic inference. We illustrated this by implementing a semi-supervised approach to use per-site rates estimated by DL within likelihood-based inference of tree topology and branch lengths. Our results showed that DL methods can be integrated within unsupervised likelihood approaches to help incorporate more realistic evolutionary scenarios in phylogenetic inference.

METHODS

A deep learning model to infer molecular evolution

We developed a DL framework to estimate the total amount of divergence across a set of nucleotide sequences and site-specific substitution rates. The sequences are assumed to be part of an alignment of orthologous genes or genomic regions sampled across multiple species.

Our implementation includes two modules: a simulator that can efficiently generate realistic datasets of aligned nucleotide sequences and a DL module that can be trained on these datasets to make predictions from empirical data. Although we employed standard substitution models for generating nucleotide alignments, we incorporated a diverse array of modes of rate heterogeneity across sites, only a few of which are currently available in likelihood-based phylogenetic software. We used this to show that our framework can help to identify patterns that would otherwise be difficult to parameterize in a likelihood context.

Simulating molecular evolution

We simulated the evolution of orthologous sequences within a phylogenetic framework assuming an independent Markov process of substitution at each site. We first generated a phylogenetic tree with a random topology and assigned exponentially distributed branch lengths sampled from $v \sim \text{Exp}(\lambda)$ with scale parameter λ randomly drawn for each tree and $\log(\lambda) \sim \mathcal{U}(\log(0.0002), \log(0.2))$. This sampling approach allowed us to generate a similar number of simulations across trees despite different orders of magnitudes of branch lengths (i.e., as many trees will be simulated with mean branch lengths between 0.0002–0.002 and 0.002–0.02). The total length of the phylogenetic tree was therefore $T = \sum_i(v_i)$, where $i \in \{1, \dots, 2N - 1\}$ was the index of each branch in a tree of N species. We simulated the evolution of nucleotides based on three substitution models (JC, HKY, GTR; Jukes and Cantor, 1969; Hasegawa et al., 1985; Tavaré, 1986) using the program Seq-Gen (Rambaut and Grass, 1997) through its Python interface implemented in Dendropy (v.4.5.2 Sukumaran and Holder, 2010). Across simulations, we varied the model parameters, i.e. base frequencies and instantaneous substitution rates, by sampling them from distributions chosen to reflect a broad range of evolutionary scenarios (Table S1; DOI: 10.5061/dryad.qz612jmn6). We chose these parameter values to reflect general expectations,

for instance the fact that transitions are likely to be more common than transversions, but our Python implementation is flexible and easily allows for user-defined variations.

Since we focused our DL model on the inference of site-specific evolutionary rates, we implemented different distributions of rate heterogeneity across sites. Regardless of the mode of rate heterogeneity (Figs. S1, S2), site-specific evolutionary rates were always rescaled to relative rates, such that their mean across all sites was equal to 1 (Yang, 1994). First, we implemented a gamma mode, where site-specific relative rates were drawn from a gamma distribution, $r_i \sim \Gamma(\alpha, \beta)$ with the shape and rate parameters set equal and drawn from $\log(\alpha) = \log(\beta) \sim \mathcal{U}(\log(0.1), \log(2))$. This setting generated rates with an average value of 1, with increasing heterogeneity when the shape and rate parameters were small (Yang, 1993). This mode of rate heterogeneity reflects the standard gamma model of rate heterogeneity, which is however typically discretized in four or more rate classes (Yang, 1994), and which is almost ubiquitously used in phylogenetic inference. Second, we implemented a bimodal mode of rate heterogeneity where sites are randomly assigned a high or a low rate based on $\log(r_i) \sim \{-m, m\}$ with m sampled from an exponential distribution $\text{Exp}(1)$. Third, we implemented a spike-and-slab mode of rate heterogeneity as a variation of the bimodal model, in which most sites evolve under low rates and few sites evolve under high rates. The low background rates were drawn from a log-normal distribution such that $\log(r_i) \sim \mathcal{N}(0, 0.1)$, while high rates were obtained by multiplying the background rates by a factor $m \sim \mathcal{U}(2, 10)$. Sites were assigned to a high rate randomly with probability r , with $\log(r) \sim \mathcal{U}(\log(0.01), \log(0.1))$. Fourth, we simulated rates based on a non-stationary distribution obtained through a geometric Brownian motion process. In this case, a vector of rates was sampled from a geometric Brownian process such that $\log(r_{i+1}) \sim \mathcal{N}(\log(r_i), \sigma)$, with $\sigma \sim \mathcal{U}(0.02, 0.2)$. Fifth, we implemented a codon mode of rate heterogeneity, in which triplets of nucleotides were assigned low, very low, and high rates, for the first, second and third positions, respectively (Nielsen and Yang, 1998). We sampled the rate of the second position from a log-normal distribution, such that for triplet i , $\log(r_i^{(2)}) \sim \mathcal{N}(0, 0.1)$. Rates for the first and third positions were then obtained as $r_i^{(1)} = m \times r_i^{(2)}$, with $m \sim \mathcal{U}(1, 5)$ and $r_i^{(3)} = n \times r_i^{(2)}$, with $n \sim \mathcal{U}(5, 15)$, respectively.

We additionally simulated datasets with rates varying among a variable number of blocks of adjacent sites, thus introducing auto-correlation in rate heterogeneity (Fig. S2). We drew the number of blocks from a geometric distribution with a mean of 100 and truncated at 1000 (i.e. the

number of sites in the alignments), and randomly sampled the sizes of the blocks. We then applied rate variation among blocks using the gamma, bimodal, spike-and-slab and geometric Brownian modes described above. Finally, we included datasets generated under mixed modes of rate heterogeneity (Fig. S2), in which the alignment was split into two blocks of random size, each with its own randomly selected rate heterogeneity mode.

Architecture and training of the deep learning model

We implemented a DL model (hereafter called phyloRNN) that takes an alignment of nucleotides as input and returns two outputs: site-specific relative rates of evolution and the expected number of substitutions per site, which is equivalent to the sum of all branch lengths in a phylogenetic inference framework. For simplicity, hereafter we will refer to this second output as *total tree length*, even though our model does not use any tree representation. We built our model based on a recurrent neural network (RNN) to capture the sequential nature of the input data. Specifically, we used a bidirectional long short-term memory architecture (bLSTM; Hochreiter and Schmidhuber, 1997; Gers et al., 2000; Graves and Schmidhuber, 2005), with a site-specific output, which is a multidimensional representation of the initial alignment. Each site-specific output of the bLSTM layer fed into two individual fully connected deep neural networks with parameters shared across all sites. The first network returned a site-specific relative rate. The output of the second network was instead concatenated across all sites and fed into a fully connected layer to return a prediction of the total tree length. The model is based on a specific number of sequences and number of sites, although it can be applied as a sliding windows to longer alignments, as demonstrated in our empirical analyses described later (see Fig. S3 for a schematic representation of the model).

After preliminary testing based on validation accuracy, we chose the following architecture for our experiments: two bLSTM layers of 128 and 64 nodes, respectively, with a tanh activation function and sigmoid recurrent activation. The output of the second bLSTM layer, for each alignment, was thus of shape ($n. \text{ sites} \times 64$) and served as input to the two deep networks that output site-specific rates and tree length. For the site-specific rates, we used a neural network with two hidden layers of 64 and 32 nodes and a swish activation function (Ramachandran et al., 2017), followed by an output layer with 1 node and a softplus output function (Szandała, 2021), reflecting a distribution of rates constrained to positive values. The concatenated output

(one value for each site), was then subjected to a rescaling function dividing each value by the mean across all sites. This rescaling ensures that the estimated rates had a mean equal to 1, thus transforming them into relative rates. The output of the second bLSTM layer was also used to infer the log-transformed total tree length, by feeding into site-specific deep networks with two hidden layers of 64 and 1 nodes and swish activation functions. As for the site-specific rates all networks shared the same parameter values. Their output (of size 1 for each site) was then concatenated across sites and fed into a fully connected hidden layer with 8 nodes and swish activation function. The output layer had one node and a linear activation function reflecting the negative-to-positive plausible range of values for the log-transformed tree length.

We trained our model based on 10,000 simulated alignments of 50 species and 1,000 sites, each with a randomly selected mode of rate heterogeneity, of which 80% were used as a training and 20% as validation set. We used the same phyloRNN model trained on an equal mix of simulated datasets for all the comparisons outlined below. Each alignment was transformed into a series of two-dimensional arrays of numbers after one-hot-encoding the sequence of nucleotides. The data fed to the model was thus composed of 1,000 two-dimensional arrays (one per site in the alignment) with each a size defined by the number of species times the one-hot encoding of each nucleotide present at a single site (i.e. $50 \text{ species} \times 4 \text{ states} = 200$). These two-dimensional arrays were then stacked on a third dimension representing the number of instances (i.e. batches used in the training or datasets used for testing) analysed to create the final input for our model (n. instances, n. sites, n. species \times n. bases).

The training and validation losses were calculated as the sum of the mean squared errors (MSE) computed across per-site rates and total tree length. We log-transformed tree lengths to reduce the range of loss values and improve the efficiency of the optimization. We trained the model over multiple epochs with a batch size of 100 and monitoring the validation loss with a patience parameter set to 20 epochs. The model was calibrated in fewer than 100 epochs, after reaching the lowest validation accuracy combining the loss functions assigned to the per-site relative substitution rates and the log-transformed tree length, (Fig. S4). We kept the model parameters inferred from the epoch with lowest validation loss. We implemented our model based on the Functional API of the Tensorflow module (Abadi et al., 2015) and trained it using the RMSprop optimizer with learning rate set to $1e-3$.

Accuracy of rate and tree length estimates

We validated the performance of our model based on a test set of 600 alignments simulated under the different modes of rate heterogeneity described above. We compared the accuracy of the estimated per-site relative rates and number of substitutions based on our phyloRNN model with those obtained through maximum likelihood phylogenetic inference. Specifically, we analyzed each alignment with PhyML v.3.3 (Guindon et al., 2010) running a maximum likelihood optimization. In a typical phylogenetic analysis, model testing is carried out first to select the best fitting substitution model (Abadi et al., 2019). Here, we assumed that the true substitution model (here one among JC, HKY, and GTR; Table S1) was known and we used it in the PhyML optimization to reduce computing time and remove the potential effect of incorrect model testing. We repeated the phylogenetic analyses under two rate heterogeneity models: i) the discrete gamma model (Yang, 1994) –by far the most commonly used in phylogenetic inference– and ii) the more flexible –but less frequently used– free-rates model, which allows for a number of rates to be inferred without making a specific assumption about their distribution (Soubrier et al., 2012). We obtained the marginal per-site rates using the `-print_site_lnl` command in PhyML (column `Posterior mean` in the output file). We quantified the average number of substitutions per site as the sum of all branch lengths from the inferred phylogenetic tree.

After obtaining the relative rates and average number of substitutions per site under both phyloRNN and the two likelihood models (discrete gamma and free-rates), we compared them against the respective true values to quantify their accuracy. We used MSE across all sites and all simulations to quantify the performance of the different models and additionally computed the coefficient of determination R^2 for each alignment to compare the estimated pattern of rate variation across sites against the true rates. To further explore the results, we divided the test set into subsets based on the rate heterogeneity model that was used to generate them (Figures S1, S2) and calculated MSE and R^2 for each subset. We calculated similar summary statistics to evaluate the accuracy of the estimated tree length, but replacing the MSE with the mean absolute percentage error, which we calculated as $|T_{\text{true}} - T_{\text{est}}|/T_{\text{true}}$, where T_{true} is the true tree length and T_{est} is the estimated one.

To assess whether the accuracy of likelihood estimates of rates and tree length was affected by potential errors in the inferred tree topology, we repeated the PhyML analyses of the

test set alignments while constraining the tree topology to be the true one. In these analyses, the maximum likelihood algorithm only optimized the parameters of the substitution model and the branch lengths.

Impact of rate estimated by phyloRNN on tree inference

We evaluated the potential effects of using phyloRNN estimates of rate heterogeneity in phylogenetic inference. Since standard phylogenetic software does not readily allow tree inference with predefined per-site rates, we first used two indirect approaches to approximate the impact of using phyloRNN rates as opposed to the two existing models implemented in PhyML (discrete gamma and free-rates).

First, we compared the likelihood obtained on the true tree using the estimated rates and the true simulated per-site rates. We computed the likelihood while fixing the topology and branch lengths to their true values (i.e., the simulated tree) and using 1) the true simulated per-site rates, 2) the marginal per-site rates estimated by PhyML under the discrete gamma and free-rates models, and 3) the per-site rates estimated with phyloRNN. We recomputed the likelihood of the true tree based on the fixed rates per site and the true parameters for the substitution model (script available at github.com/phyloRNN). For each simulated alignment, we compared the likelihood obtained by the true versus estimated rates by calculating the difference in log-likelihood. We summarized these comparisons by computing the proportion of datasets in which the likelihood of the true tree substantially decreased (i.e. more than 2 log-likelihood unit differences) using estimated versus true rates and interpreted them as a measure of the potential impact of the gamma, free-rates, and phyloRNN models on the accuracy of tree inference.

Second, we compared, for each model used to estimate rates per site, the likelihood of the true tree against the likelihoods of a posterior sample of trees obtained from a Bayesian analysis. We first obtained a posterior sample of 50 trees for each test dataset using MrBayes (Ronquist et al., 2012) under a GTR + gamma model and default prior settings. We then recomputed the likelihood of both the set of sampled trees and the true tree based on the rates inferred under the gamma and free-rates models (as inferred with PhyML) and under the phyloRNN model. Finally, we ranked all trees by their likelihood to assess whether the likelihood of the true tree falls within the range of sampled values in the posterior set of trees.

If the estimated rates adequately reflected the true rate variation in the data, we expected the likelihood of the true tree to rank somewhere within the range of likelihoods of the sampled trees, as this would indicate that the true tree is likely to be included in the posterior distribution obtained under a gamma model. In contrast, if in a substantial proportion of simulations the likelihood of the true tree was lower than that of the sampled trees, then the estimated rates might be inadequate by assigning significantly higher likelihood to solutions that differ from the truth. Finally, if in a substantial proportion of simulations the likelihood of the true tree is higher than that of the sampled trees, then the rates, e.g. estimated through a free-rates or phyloRNN model, would favor sampling the true tree over alternatives sampled under a discrete gamma model.

Since the trees were obtained under a gamma model, we expected the likelihood of the true tree based on rates from the gamma model to be found within the sampled range or lower (if the rates are inadequate). A higher likelihood is still possible, however, if the subset of 50 trees does not capture the actual upper boundary of the posterior distribution of likelihood values. With rates inferred from the free-rates or phyloRNN models, we instead expected a larger proportion of simulations in which the true tree returned a higher likelihood than the sampled ones because of the better fit to the underlying true mode of rate heterogeneity.

Phylogenetic inference with per-site rates

We evaluated the impact of applying per-site rates in phylogenetic inference to assess if phyloRNN-estimated rates could affect the accuracy of the trees estimated in a likelihood framework. We developed an analytical pipeline to first estimate per-site rates using our trained phyloRNN model and then applying them in a Bayesian analysis performed in RevBayes (v. 1.2.1; Höhna et al., 2016). As a substitution model with fixed per-site rate is currently not available in RevBayes, we created per-site data partitions to assign phyloRNN-estimated rates to each site in the alignment (the RevBayes script can be generated through a utility function in phyloRNN). All partitions shared the same GTR substitution matrix and state frequencies, thus differing only in the relative rate applied to each.

Even though the model requires the estimation of fewer parameters compared to a standard model with gamma heterogeneity across sites, the creation of many partitions drastically increased the computing time and we opted to run these simulations on smaller datasets with 20

and 50 tips and 100 sites only. We simulated 20,000 datasets to train two phyloRNN models based on either 20 or 50 tips, with parameter settings as described above and mean branch lengths drawn from an exponential distribution $v \sim \text{Exp}(\lambda)$ with $\log(\lambda) \sim \mathcal{U}(\log(0.01), \log(0.2))$. We increased the lower threshold on the mean branch lengths in these simulations to ensure a sufficient number of substitutions occurred in these smaller datasets.

We then generated two test sets of 200 datasets with again either 20 or 50 tips, respectively, and 100 simulated DNA sites. On each alignment we performed phylogenetic inference using RevBayes, with a GTR substitution model. The analyses assumed either a gamma model of rate heterogeneity with invariant sites, or a model with fixed relative per-site rates, as predicted by our trained models. Note that although phyloRNN models predict both per-site rates and total tree length, we only used the per-site rates in the phylogenetic inference. We ran 10,000 MCMC iterations, which we found to be sufficient to achieve convergence in our datasets, and summarized the resulting posterior sample of trees using the `mapTree` function in RevBayes.

To compare the results of phylogenetic inference based on a standard gamma model with invariant sites against one using phyloRNN-estimated per-site rates, we 1) compared the mean sampled likelihood of the data as an approximation of model fit, 2) calculated the weighted Robisons-Fould distance as implemented in the R package `phanghorn` (Robinson and Foulds, 1981; Schliep, 2011) between the true tree and the estimated tree, and 3) calculated the squared difference between the true and the estimated log-transformed tree lengths.

To improve the scalability of this approach, we implemented a model in which a pre-defined number of rate categories are used to approximate the per-site rates inferred by phyloRNN. We first grouped the per-site rates obtained with phyloRNN in a number of classes equally distributed in log space. We then created one partition for each rate class and assigned all sites within this partition to the average rate for that class. This procedure allowed us to run RevBayes analyses on larger alignments. To test this approach, we simulated 100 datasets of 50 tips and 1,000 sites and obtained per-site rates using the phyloRNN model trained for the simulations described in previous sections. We then estimated phylogenetic trees based on 1) a gamma model with invariant sites and 2) models with phyloRNN rates discretized into 4, 10, 20, or 50 classes. We summarized the results as with the simulations described above and evaluated the effect of using pre-estimated site rates and the impact of increasing the number of classes on the accuracy of the resulting trees.

Analysis of the clownfish genomes

We applied our model to a genomic dataset of 28 species of clownfish representing the first chromosome (Marcionetti and Salamin, 2023). Even though phyloRNN does allow for gaps in an alignment, we decided for simplicity to filter out positions in the alignment containing gaps and ambiguities reducing the initial dataset of over 46 millions nucleotides to a total of 26,294,222 aligned nucleotides. We trained a new model containing 10,000 simulated datasets to match the input of 28 taxa and 1,000 nucleotides. We used it to predict per-site relative substitution rates and total tree length (as a measure of overall divergence across the chromosome) through non-overlapping sliding windows of 1,000 sites. We generated histograms of the rate heterogeneity across a random subset of exons (filtering out those of length smaller than 500 nucleotides) to visually assess whether a gamma distribution adequately approximates the empirical distribution of rates. Finally, we tested whether protein-coding regions showed consistently lower substitution rates compared with neighboring non-coding regions as expected if they are functionally constrained. We estimated the mean substitution rate for each exon (filtering out those of length smaller than 250 nucleotides; results did not change if the limit was set to 100 nucleotides) as well as the mean rates of the directly adjacent non-coding regions selecting 250 nucleotides before and after the start or end of each exon selected. We performed paired t-tests to test whether the rates were significantly different between exons and directly adjacent regions.

We estimated the total number of substitutions across blocks of 10, 100, 500, and 1000 sites along the 26-species phylogenetic tree of clownfish, assuming site-specific estimates obtained with the PHAST software package (v. 1.6; Hubisz et al., 2010, accessed at github.com/cshlsiepellab/phast). We first ran phyloFit (option `-subst-mod REV`) using the clownfish tree topology and the fasta alignment to obtain branch lengths and the nucleotide matrix, and then phyloP (options `-base-by-base -mode CONACC -method GERP`) to obtain site-specific estimates of the number of substitutions. We note, however, that we ran the phyloFit estimation assuming a single category of site-rates (`ncat=1`) as a model with rate heterogeneity (`gamma ncat=4`) applied to the clownfish chromosome 1 dataset did not converge in our experiments.

RESULTS

Performance of the phyloRNN model

We measured MSE and R^2 values for the rate predictions obtained from the trained model on the test set and the rates estimated by PhyML with the gamma and free-rates models. The results showed that our model provided substantially more accurate estimations of the per-site rates compared with maximum likelihood estimates obtained through a gamma model of rate heterogeneity (lower MSE and higher R^2 values; Table 1). The phyloRNN model also outperformed maximum likelihood estimations based on the more flexible free-rates model under most scenarios, especially the more complex heterogeneity modes (Table 1), although the difference was smaller than with the gamma model. After breaking down the results by the simulated mode of rate heterogeneity, we found that the improvement in rate estimation is particularly strong in the case of codon mode of rate heterogeneity (MSE values decreasing by one order of magnitude when using phyloRNN) and in the case of autocorrelated rates (Fig. 2). The phyloRNN estimates appeared to consistently outperform maximum likelihood estimates particularly in their ability to recover multimodal distributions of rates across sites (Figs. 2, S5–S10). Analyses of datasets simulated under mixed rate heterogeneity models yielded substantially less accurate estimations of the per-site rates (Table 1). The lower accuracy was consistent across models indicating that these patterns of rate variation remain difficult to predict under a single model, whether probabilistic or based on DL (Figs. S7–S9). Analyses based on a phyloRNN model re-trained with a larger proportion of data generated under the mixed model (increased from 5% to 20%) yielded only limited improvements in accuracy to estimate the rate and tree length in this subset of the test set. The improvement was however sufficient for phyloRNN estimates to become more accurate than maximum likelihood estimations (Tables S4, S5).

Figure 2: Example of simulated and estimated per-site rates based on phyloRNN. Plots on the left show per-site rates (note that the estimated rates are shifted slightly to the right for clarity). Histograms show the true distribution of rates across 1000 sites in an alignment (the bottom right is in log space for clarity) and the distribution of estimated rates. The simulations show an example of different modes of rate heterogeneity: gamma (a), codon (b), and spike-and-slab autocorrelated (c).

Although the phyloRNN model did not use (or attempt to estimate) a phylogenetic tree, its estimation of the total tree length is unbiased (Fig. 3) and showed comparable accuracy with the estimates obtained from a maximum likelihood analysis based on a gamma model of rate heterogeneity, while the free-rates model generally produced the most accurate estimations. The accuracy of tree length estimation was high in most simulations with mean absolute percentage errors generally below 15% (Table 2). Tree lengths were inferred with higher error for simulations based on mixed modes of rate heterogeneity and this was the case across all three inference methods.

Figure 3: Estimated total tree length (log-transformed) under maximum likelihood models assuming a gamma heterogeneity model (a) and a free-rates model (b), and through phyloRNN (c). The results are shown for a test set of 600 datasets generated under different modes of rate heterogeneity (more details in Table 2).

Fixing the tree topology to the true tree in maximum likelihood analyses had a negligible effect on the estimated per-site rates, with MSE equal to 0.833 for the gamma model and 0.709 for the free-rates model (Table S2). Removing any potential topological errors from the analyses did not substantially change the accuracy in the estimated rates, which remained more accurately predicted by the phyloRNN model. Similarly, fixing the tree topology to the true tree led to minimal changes in the estimated tree length with a mean absolute percentage error of 0.106 under the gamma model and 0.077 under the free-rates model (Table S3). Topological errors did not therefore affect strongly the accuracy of the estimated tree length, which remained comparable between phyloRNN and maximum likelihood phylogenetic inference.

Impact on tree likelihood

In our simulations, the likelihood of the true tree based on the true per-site rates substantially exceeded ($\Delta \log L > 2$) the likelihood of the same tree using a discrete gamma model of rate heterogeneity in 90.2% of the datasets. This means that, in a large fraction of the simulations, a gamma model decreased the likelihood of the correct underlying phylogenetic tree (Table S7). We observed a similar outcome under the free-rates model, where the likelihood of the true tree decreased in 89.6% of the simulations. In contrast, phyloRNN rates resulted in a substantially

lower likelihood for the true tree only in 29.7% of the simulations, suggesting that the use of these rates in phylogenetic inference might result in more accurate estimated trees (Table S7).

A change in absolute likelihood does not necessarily imply that a model is less likely to sample the true tree in phylogenetic inference, because it could simply reflect a homogeneous shift in the likelihood surface. We therefore also evaluated the ranking of the true tree within a posterior sample of trees for each testing dataset described above. We sampled 50 trees from the posterior samples of MrBayes and ranked them in decreasing order based on their likelihood recomputed under predicted rates from the gamma, free-rates and phyloRNN models as above. We then compared the likelihood of the true tree obtained with the same models with the ranked sampled trees. Under the gamma model, the likelihood of the true tree was found within the range of sampled trees in 86.0% of the simulations and ranked first in 1.8% of the cases. In the latter case, the mean difference in log-likelihood between the true tree and the best tree from the posterior samples was small (ranging from -0.130 to -7.757, median of -1.857), which suggests that the true tree would in fact be included in a more extensive posterior sample including more than the 50 trees considered here. However, the true tree ranked last, and thus was outside the sampled range, in 12.2% of the simulations, with a range of log-likelihood differences between the true tree and the worst tree from the posterior samples ranging from 0.146 to 2129.506 (median of 24.137). This indicates that, for most of these cases, the true tree was unlikely to be sampled by the Bayesian algorithm, and thus excluded from the estimated posterior distribution of trees.

In contrast, under the rates estimated with phyloRNN, the true tree ranked last in only 3.7% of the simulations (log-likelihood difference ranging from 0.309 to 3602.207 with a median of 16.141), while it ranked first in 13.5% (log-likelihood difference ranging from -0.006 to -1841.836 with a median of -38.206). This suggests that in a substantial proportion of datasets, the phyloRNN rates would favor the true tree over the trees sampled under a gamma model. For comparison, the free-rates model performed similarly to phyloRNN, with 3.2% of simulations with the true tree ranking last (log-likelihood difference ranging from 0.033 to 2806.083 with a median of 10.348) and 12.2% of the simulations with the true tree ranking first (log-likelihood difference ranging from -0.125 to -1341.208 with a median of -20.146).

The rank of the true tree in all these comparisons did not depend strongly on the mode of rate heterogeneity, the model of substitution or the tree length used in the simulations which

were not significant when analysed with a linear model. However, we did find that instances where the true tree ranked last under a gamma model (therefore being excluded from the sampled posterior distribution of tree) were associated with significantly higher error in both estimated per-site rates and tree length (Fig. S11). This corroborates the idea that an improved estimation of these parameters can lead to a more accurate tree estimation, which explore further in the next section.

Impact on tree estimation

RevBayes analyses comparing a gamma model with phyloRNN-estimated site-specific rates showed that the latter can lead to a significant improvement in the accuracy of phylogenetic inference. The mean sampled log-likelihood was higher using phyloRNN rates compared with rates under a gamma model in all datasets with 20 tips, with a median difference of 22.5 log units (95% interval: 7.2, 88.1) and in almost all the datasets with 50 tips (median difference of 41.1 log units; 95% interval: 7.9, 108.0 log units; Figs. 4a, S12a).

Figure 4: Comparison between phylogenetic inference using a gamma model of across site rate heterogeneity and fixed phyloRNN-estimated per-site rates. Boxplots summarize the results of 200 simulated datasets with 50 tips and 100 nucleotides (the outliers are not shown but are plotted in Fig. S13). All analyses were carried out using RevBayes. Positive values indicate improvements in analyses using phyloRNN rates compared with gamma rates (i.e. higher sampled likelihood (a), lower distance or error between true and estimated trees (b-d)).

The maximum *a posteriori* tree topologies were similarly accurate with median normalized R-F distances between estimated and true trees of 0.27 (20 tips) and 0.31 (50 tips) for both phyloRNN rates and gamma rate heterogeneity (Figs. 4b, S12b). However, when taking into account branch lengths, using the weighted R-F distance, we found that trees inferred using phyloRNN rates were substantially more accurate. The weighted R-F distances between true and estimated trees were on average 48.4% lower in analyses using phyloRNN rates (95% interval: -9.9%, 310.8%) compared with estimates obtained from a gamma model for 20-tip simulations. The weighted R-F distances for 50-tip simulations were on average 40.0% lower using phyloRNN rates (95% interval: -8.4%, 222.9%) compared rates estimated under a gamma model. In both sets of simulations the weighted R-F distances were lower for the inference using phyloRNN rates than the rates under the gamma model in 82% of the datasets (Figs. 4c,

S12c). The difference in the accuracy of branch length estimations was also reflected in the error to estimate the total tree length (mean squared difference between true and estimated log tree length), which was on average more than 5 times lower for the analyses using phyloRNN rates compared to those using rates under the gamma model. Across simulations, the tree length was more accurately estimated using phyloRNN rates in 85.5% of the 20-tip simulations and in 93% of the 50-tip simulations (Figs. 4d, S12d).

Phylogenetic inference carried out in RevBayes under the gamma model of rate heterogeneity involved the estimation of two additional free parameters compared with inference based on phyloRNN-estimated per-site rates, namely the shape of the gamma distribution and the proportion of invariant sites. This difference in parameterization possibly led to the observation that the MCMC sampling was on average more efficient when using phyloRNN rates. Specifically, the effective sample sizes of the posterior computed across the last 500 samples was higher with phyloRNN rates than with rates under the gamma model in 75.5 % of the 20-tip simulations with a median increase of 62.5 effective samples. Similarly, the effective sample sizes were higher with phyloRNN rates than with rates under the gamma model in 77.5% of the 50-tip simulations (median increase of 56.8 effective samples). These results suggested that the use of phyloRNN-estimated per-site rates can reduce the number of MCMC iterations required to reach convergence.

Discrete phyloRNN-rate classes

The analysis of larger datasets (50 taxa, 1,000 sites) using rate classes to approximate phyloRNN rates resulted in higher mean sampled log-likelihood values compared with rates under a gamma model in 97–99% of the simulations, and the mean log-likelihood improvement increased with the number of rate classes (Table S6; Fig. S14a). The accuracy of the estimated topology computed through R-F distances did not change substantially among models. However, the weighted R-F distances between the true and estimated trees were lower in 63-69% of the analyses using phyloRNN rates than with rates under a gamma model, decreasing on average by 23–27% (Table S6; Fig. S14). Similarly, the error in total tree length (mean squared difference between true and estimated log tree length) was lower in 71–73 of the analyses using phyloRNN rates, improving between 3 and 4 fold compared with analyses using rates estimated under a gamma model.

Evolutionary rates and divergence in clownfish

The estimated substitution rates along chromosome 1 in clownfish showed a substantial degree of heterogeneity across sites, with 99% of the values ranging between 0.016 and 0.222, thus encompassing a ~ 14 -fold rate variation (Fig. 5a). The overall distribution of rate heterogeneity across all ~ 26 M sites followed quite closely a gamma distribution (Fig. 5a), indicating that a gamma model should approximate well the true rate variation at a broad genomic scale. However, the distributions of substitution rates across smaller genomic regions showed that across-site rate heterogeneity often diverged substantially from a gamma distribution, showing multimodal patterns and heavy-tailed distributions (Fig. 5b-i).

Figure 5: Estimated rate heterogeneity across sites plotted for all sites in chromosome 1 in the clownfish genome (a) and across a random sample of eight exons (b–i). The red lines show gamma distributions fitted to the data shown in the histogram. The rates match well a gamma distribution of rate heterogeneity across all 26 million sites (a), consistently to the heterogeneity models typically used in likelihood-based phylogenetic analyses. However, at the exon level, the rate distribution often diverges substantially from that of a gamma distribution, displaying multimodal or heavy-tailed distributions. For improved visualization, rates > 3 ($N = 53,061$, or $\sim 0.2\%$) are not shown in panel a).

The degree of clownfish divergence (the total tree length) estimated within blocks of 1,000 sites revealed up to ~ 4 -fold variation in estimated number of substitutions across the chromosome. The distribution of substitution rates per site also highlighted regions in the chromosome spanning several thousands of sites that are more conserved (i.e. low average rates) and others characterized by much higher rates (Fig. 6).

When comparing the substitution rates between coding and non-coding regions, we found substitution rates in exons to be lower than the rates in adjacent regions in the chromosome in 82% of the 854 exons analyzed. The mean rate in exons was on average $\sim 15\%$ lower than in the adjacent non-coding regions (Fig. 6). Paired t-tests showed that this rate difference was overall significant ($p\text{-value} < 9.4e^{-74}$, $T < -20.08$, 95% confidence interval: $[-0.02, -0.01]$), while the left and right adjacent regions did not differ from one another ($p\text{-value} = 0.35$, $T < -0.94$, 95% confidence interval: $[-0.01, 0]$).

The estimated number of substitutions, inferred from phyloRNN-estimated rates, were comparable with those calculated using phyloP (Fig. S15). Our results indicate a stronger correlation between the two types of estimates as we increase the number of sites considered,

Figure 6: Estimated number of substitutions per site inferred through our RNN model across the first 10 million nucleotides of the clownfish genomes (28 species, chromosome 1). a) Substitutions per site as function of chromosomal position, rates are averaged across blocks of 1,000 nucleotides. b) Box-plots of substitutions per site for exonic regions (random sample of 854 exons, minimum of 250 nucleotides) and adjacent non-exonic regions (250 nucleotides before and after the start or end of each exon). Rates in exonic regions are on average 15% lower than in the adjacent regions ($p\text{-value} < 9.4e^{-74}$).

with a correlation coefficient raising from $R^2 = 0.24$, based on blocks of 10 sites, to $R^2 = 0.79$, based on blocks of 1000 sites.

DISCUSSION

A deep learning model to infer molecular evolutionary rates

We presented a framework using stochastic simulations to train a DL model and estimate site-specific rates of evolution and total tree length from an alignment of nucleotide sequences. We specifically designed the architecture of our model to reflect the characteristics of molecular data and assumptions behind the evolution of DNA sequences. Indeed, the bLSTM layers capture the sequential nature of DNA data, while the use of site-specific networks with parameter sharing reflects the fact that each feature in the input layer is an instance of the same nature (i.e. a nucleotide).

Our implementation allowed us to accurately estimate the evolutionary rates for each site of an alignment together with the total evolutionary divergence in the alignment, which, in a phylogenetic context, is quantified by the tree length. This was done using solely the information from the alignment and without the addition of a defined phylogenetic tree in the input to our phyloRNN model. We showed that our model outperforms likelihood estimations based on the standard gamma distribution of rate heterogeneity and that it matched or outperformed, depending on the rate heterogeneity mode used, the estimates from the more parameter-rich and less frequently used free-rates model. The computational efficiency of our approach makes it highly scalable, allowing for the analysis of large-scale genomic data.

Evolutionary inference using phyloRNN

While rate heterogeneity is generally considered as a nuance parameter in phylogenetic inference (Yang, 1994), an accurate estimation of substitution rates per site can be used to identify genomic regions that are under evolutionary constraints or deviate from the average pattern across genomic regions (Mayrose et al., 2005). It can also be used to identify protein coding genes because of the significantly reduced rate of evolution compared with adjacent genomic regions, as we found in the clownfish dataset (Fig. 6). These results could be coupled with other approaches (e.g., Cooper et al., 2005; Siepel et al., 2005) used in genomics to help annotate *de novo* genome assemblies or to identify regions of interest that show unexpected levels of conservation outside of the protein-coding genes.

The analyses of the genomic data for clownfish allowed us to look at the distribution of rates across a large empirical dataset. Although the distribution of rates appears to match a gamma distribution when the full set of ca. 26 million sites are considered, the distributions for individual exons can drastically differ from a gamma distribution (Fig. 5). A similar pattern had been previously shown in the distribution of average rates across genes (Bevan et al., 2007) instead of individual sites like in our study, or in datasets with low level of variation (Jia et al., 2014). To tackle such cases in a likelihood framework, complex mixture of gamma distributions can be used to model this distribution of rate heterogeneity (Mayrose et al., 2005), but at the cost of additional complexity during the optimization process (Bevan et al., 2007). In contrast, our phyloRNN model can easily account for complex distributions through the simulation of various types of rate heterogeneity.

The gamma distribution to model rate heterogeneity was introduced in a landmark paper (Yang, 1994) and led to a substantial improvement in likelihood-based tree inference across most empirical datasets. Although the fit of models of evolution including a gamma distribution is drastically better compared to models without rate heterogeneity, the use of a gamma distribution is not based on a biological assumption. Additionally, our results showed that deviations from the gamma distribution can have a large impact on the basic estimates of rates per site or tree length (Tables 1, 2). Alternative models have been proposed to include more biological realism (Heaps et al., 2020), and allow for more flexible distributions, like the free-rates model, which is implemented in some phylogenetic software (e.g., the PhyML program

used here and IQTREE (Minh et al., 2020)). Other programs use a discrete-rates CAT model, implemented in RAxML (Stamatakis, 2014) and FastTree2 (Price et al., 2010), which is more efficient computationally than fitting a gamma distribution. Finally, a full Bayesian method to simultaneously estimate the substitution model and rate at each site has also been proposed (Wu et al., 2013), but it involves a large number of parameters with unclear effects on tree inference. The model developed here represents an alternative, potentially more efficient, approach, where rates per site are inferred prior to phylogenetic inference and without assuming any specific distribution. This approach reduces the number of free parameters in a phylogenetic inference model and our experiments showed that it can have a beneficial impact on its accuracy.

Impact on phylogenetic inference

Including rate heterogeneity has been shown to have a large impact on phylogenetic inference (Yang, 1994; Sullivan and Swofford, 1997; Abadi et al., 2019). Our simulations further demonstrated that an accurate modeling of the rate variation across site will also affect, sometimes drastically, the estimation of the tree likelihood. The effect is not simply a monotonic increase or decrease of the whole likelihood surface, but affects the ranking of a sample of trees. We showed it indirectly by comparing the likelihood of sampled trees computed by assigning rates per sites estimated using the gamma, free-rates and phyloRNN models. The changes in log-likelihood values demonstrate that an inaccurate estimation of the rates per site can bias tree inference and result in significantly lower likelihood of the true tree relative to alternative hypotheses.

In likelihood-based phylogenetic inference, nucleotide frequencies are routinely set equal to their empirical values calculated from the alignment, rather than estimated in the analysis. This is done to reduce the number of free parameters in likelihood optimization or posterior sampling algorithms. Similarly, per-site rates could be estimated from the alignment using our phyloRNN model before fixing them during the likelihood search. This is what we showed using the flexibility of the RevBayes program to incorporate pre-estimated per-site rates directly in phylogenetic inference. Our results showed that this approach can improve the estimation of the tree, with substantial improvements in the estimated branch lengths. Our results also suggested that this approach might facilitate the convergence of Bayesian phylogenetic inference as it reduces the number of free parameters.

Performance of phyloRNN on big data

Our application of the phyloRNN model to chromosome 1 of 28 clownfish species showed that a trained model can rapidly estimate per-site substitution rates across large genomic datasets with a relatively small computational footprint. For instance, the analysis of the clownfish dataset on a 64-CPU workstation required 122 minutes: 14 minutes used to simulate the training data, 83 minutes to train the model, and 25 minutes to parse the empirical dataset and generate predicted rates across the 26 million sites. For comparison, tests on 100 simulated datasets with 28 species and 1000 sites using PhyML showed that a comparable analysis, with fixed topology and optimization limited to branch lengths and model parameters on the same machine would take about 793 minutes (6.5 times longer than using phyloRNN) under a gamma model and about 1,794 minutes under a free-rates model (almost 15 times longer). Given the short prediction times using phyloRNN once the models were trained, the computational benefit of our deep learning model increased further with increasing size of the datasets. If for instance we used our model to predict rates across another clownfish chromosome of comparable size, this would only add another ≈ 25 minutes of computing time, while it would bring computing time in a maximum likelihood framework above 26 hours with a gamma model and around 60 hours with a free-rates model.

Our implementation also allows the use of GPU for model training, which, in the case of the clownfish dataset using an RTX 4080, reduced the time for training to 33 minutes, i.e. 2.5 times faster than on 64 CPUs. The full analysis including simulations and predictions, thus achieves speedups of 11-fold compared to a gamma model and 26-fold compared to a free-rate model.

Effects of violations of model assumptions

A common critique to supervised DL models over their unsupervised likelihood-based alternatives in regression and other inference tasks, is their unpredictably erroneous behavior when presented with data that differ from the training data (Marcus, 2018). In the case in which the training data were simulated under a generative model, like in our study, differences between training and empirical data could be driven by violations of the assumptions of the generative model in real world evolution. However, violations of the model assumptions have also been

shown to lead to wrong estimations in likelihood-based inference of evolutionary models. For instance, simplistic substitution models assuming equal substitution rates among nucleotides (an assumption clearly violated by the real evolutionary process), have long been known to lead to wrong tree topologies (D’Erchia et al., 1996; Sullivan and Swofford, 1997). More recently, Meyer et al. (2019) found that the presence of co-evolving sites in an alignment, which violates the common assumption of site independence, can bias in unpredictable ways phylogenetic inference, affecting the accuracy of both tree topology and branch lengths. Similar misbehavior in likelihood-based inference has been shown in the context of models of trait evolution (Duchen et al., 2021) and species diversification (Louca and Pennell, 2020). Thus, phylogenetic and macroevolutionary analyses are likely to be generally sensitive to model violations.

Recent research has shown that the current models of nucleotide evolution might be inadequately reproducing realistic nucleotide sequence alignments (Trost et al., 2023), although the effects of this inadequacy on phylogenetic inference remains to be fully explored. While in likelihood-based models the assumptions about how evolutionary mechanisms play out are built directly into the likelihood function itself, in our phyloRNN framework the same assumptions are encoded in the simulation module (Fig. 1). This architectural difference makes it substantially easier to relax these assumptions in a model like phyloRNN that couples stochastic simulations with DL. We have demonstrated this through the implementation and training of a single model able to account for a range of heterogeneity patterns, including auto-correlated rates and codon models, each of which would require a specific parameterization in a likelihood framework. In the phyloRNN framework, the inclusion of additional heterogeneity patterns is straightforward as long as such patterns can be simulated, thus facilitating its extension to more diverse evolutionary scenarios.

Why we still need likelihood-based evolutionary models

While DL is now permeating many research fields in biology (Sapoval et al., 2022), we think that well-principled and fully interpretable likelihood models will continue to play a key role in evolutionary biology, for several reasons. First, likelihood-based methods are (arguably) more suitable for the estimation of complex parameters. In contrast, most of DL models are designed to infer simple output parameters (e.g., continuous values in regression tasks or categorical variables in classification tasks). Their application to more complex parameters

such as the phylogenetic tree topology is instead less straightforward to implement in a standard output layer beyond small scale implementations (e.g., Zou et al., 2019; Sapoval et al., 2022), although recent developments indicate that this will improve in the future (Nesterenko et al., 2022; Smith and Hahn, 2023).

Second, likelihood-based methods provide a more direct and robust assessment of parameter uncertainty, e.g. through bootstrap values, confidence interval estimates or posterior probabilities (e.g., in phylogenetic inference, Felsenstein, 1985; Yang and Rannala, 1997; Huelsenbeck and Ronquist, 2001; Heled and Drummond, 2009; Lemoine et al., 2018; Meyer et al., 2019). Credible intervals can also be obtained through marginal distributions from Bayesian analyses to account for parameter uncertainties. In contrast DL models are generally trained to achieve best predictive accuracy, with limited focus on the quantification of uncertainties. Part of the reason for this stems from the fact that artificial intelligence research, unlike evolutionary biology, has for the most part focused on accuracy scores rather than on the estimation uncertainty (Koch et al., 2021). Class probabilities, as quantified by a softmax output layer have been successfully used as proxies for confidence in evolutionary biology models (Suvorov et al., 2019; Silvestro et al., 2020), even though they have been shown to perform poorly in other classification tasks (Gal and Ghahramani, 2016; Silvestro and Andermann, 2020). Alternatively, Bayesian implementations, Monte Carlo dropout, and model ensembles can be used to approximate confidence intervals around the predictions from DL models, but they are not always scalable for large models or offer somewhat *ad hoc* estimations of uncertainties (Blundell et al., 2015; Gal and Ghahramani, 2016; Polson and Sokolov, 2017; Silvestro and Andermann, 2020).

Third, hypothesis testing is a crucial aspect in evolutionary biology and this is more directly implemented within a probabilistic framework. The statistical comparison between alternative hypotheses typically involves a probabilistic approach, which does not easily have an equivalent in machine learning. Furthermore, delving into the significance of different nodes within a network and comprehending their influence on model performance with a specific dataset assumes an elevated level of complexity. The intricate and nonlinear decision boundaries that are inherent in deep neural networks combined with their extensively parameterized architecture foster an impressive predictive accuracy, but also contribute to the challenge of interpreting them compared to other likelihood-based models.

Semi-supervised learning for phylogenetic inference

The analysis of simulated alignments in RevBayes revealed that the use of phyloRNN per-site rates can substantially improve the accuracy of the estimated phylogenetic trees. An integrated approach combining DL and likelihood-based inference, can be seen as a form of semi-supervised learning (Zhu, 2005), in which supervised and unsupervised parts of the overall model are applied to different sets of parameters. The improvements of this semi-supervised approach to phylogenetic inference were most evident in the estimated branch-lengths and the resulting total tree length. The approach also led to a strong increase in the likelihood of the data (at least based on the posterior samples obtained from MCMC), while reducing the number of free parameters that needed to be estimated by RevBayes, therefore potentially facilitating the posterior sampling via MCMC (as suggested by the higher effective sample sizes achieved in our comparisons).

While the current implementation allows for the use of different rates for each site only for small alignments, we showed that the phyloRNN rates can be discretized in fewer rate categories and still have a beneficial effect on the estimated trees. Future implementations will likely improve the performance of this approach that should also involve fewer evaluations of the likelihood, which is computed once per site compared with four times per site in a gamma model discretized into four categories. Our results indicate that coupling DL with likelihood-based methods offer the opportunity for more accurate, robust and interpretable estimations of macroevolutionary parameters and phylogenetic relationships. Thus, we envision a new generation of semi-supervised phylogenetic models that integrate likelihood-based and DL components to improve the efficiency and scalability of the analyses, while relaxing some of the assumptions currently made for mathematical convenience, paving the way for a better understanding of macro-evolutionary processes.

SUPPLEMENTARY MATERIAL

Supplementary material, including data files and online-only appendices, can be found in the Dryad data repository (provisional DOI: 10.5061/dryad.qz612jmn6.).

CODE AND DATA AVAILABILITY

The phyloRNN model is implemented as an open-source Python library and available at github.com/phyloRNN. The library relies on python modules `dendropy` (Sukumaran and Holder, 2010) and `tensorflow` (Abadi et al., 2015) and offers integrations with `seq-gen` (Rambaut and Grass, 1997) and `PhyML` (Guindon et al., 2010) to simulate and analyze sequence alignments. The library also includes utility functions to parse empirical data and generate `RevBayes` (Höhna et al., 2016) scripts to perform Bayesian phylogenetic inference based on phyloRNN-estimated per-site rates. The scripts showing how to run phyloRNN and replicate the analyses presented here along with pre-trained models are available on `phyloRNN/Scripts_and_data/Scripts`. Codes and data are also available on the Dryad data repository: provisional DOI: 10.5061/dryad.qz612jmn6.

FUNDING

D.S. received funding from the Swiss National Science Foundation (PCEFP3_187012), the Swedish Research Council (VR: 2019-04739), and the Swedish Foundation for Strategic Environmental Research MISTRA within the framework of the research programme BIOPATH (F 2022/1448). N.S. and T.L. were supported by a grant from the Swiss National Science Foundation (310030_185223) to N.S. and funding from the University of Lausanne.

ACKNOWLEDGEMENTS

We thank Anna Marcionetti for help with the clownfish data and Diego A. Hartasánchez for feedback on the manuscript. We thank Laura Mulvey and Michael Landis for help with `RevBayes` and Philippe Baumann for help with the GPU setup.

References

- Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.
- Abadi, S., O. Avram, S. Rosset, T. Pupko, and I. Mayrose. 2020. Modelteller: Model selection for optimal phylogenetic reconstruction using machine learning. *Molecular Biology and Evolution* 37:3338–3352.
- Abadi, S., D. Azouri, T. Pupko, and I. Mayrose. 2019. Model selection may not be a mandatory step for phylogeny reconstruction. *Nature Comm* 10:934.
- Bevan, R. B., D. Bryant, and B. F. Lang. 2007. Accounting for gene rate heterogeneity in phylogenetic inference. *Systematic Biology* 56:194–205.
- Blundell, C., J. Cornebise, K. Kavukcuoglu, and D. Wierstra. 2015. Weight uncertainty in neural network. Pages 1613–1622 in *International conference on machine learning PMLR*.
- Bull, J., M. Badgett, H. A. Wichman, J. P. Huelsenbeck, D. M. Hillis, A. Gulati, C. Ho, and I. Molineux. 1997. Exceptional convergent evolution in a virus. *Genetics* 147:1497–1507.
- Chan, J., V. Perrone, J. Spence, P. Jenkins, S. Mathieson, and Y. Song. 2018. A likelihood-free inference framework for population genetic data using exchangeable neural networks. *Advances in neural information processing systems* 31.
- Cooper, G. M., E. A. Stone, G. Asimenos, E. D. Green, S. Batzoglou, and A. Sidow. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome research* 15:901–913.

- Cooper, R., J. Flannery-Sutherland, and D. Silvestro. 2024. Deepdive: estimating global biodiversity patterns through time using deep learning. *Nature Communications* Pages doi: 10.1038/s41467-024-48434-7.
- D’Erchia, A. M., C. Gissi, G. Pesole, C. Saccone, and U. Arnason. 1996. The guinea-pig is not a rodent. *Nature* 381:597–600.
- Duchen, P., M. L. Alfaro, J. Rolland, N. Salamin, and D. Silvestro. 2021. On the effect of asymmetrical trait inheritance on models of trait evolution. *Systematic Biology* 70:376–388.
- Felsenstein, J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Biology* 22:240–249.
- Felsenstein, J. 1981. Evolutionary trees from dna sequences: A maximum likelihood approach. *J Mol Evol* 17:368–376.
- Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *evolution* 39:783–791.
- Felsenstein, J. 2003. *Inferring phylogenies*. Sinauer Associates.
- Flagel, L., Y. Brandvain, and D. R. Schrider. 2019. The unreasonable effectiveness of convolutional neural networks in population genetic inference. *Molecular biology and evolution* 36:220–238.
- Gal, Y. and Z. Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. Pages 1050–1059 *in* international conference on machine learning PMLR.
- Gernhard, T. 2008. The conditioned reconstructed process. *J Theor Biol* 253:769–778.
- Gers, F., J. Schmidhuber, and F. Cummins. 2000. Learning to forget: Continual prediction with lstm. *Neural computation* 12:2451–2471.
- Graves, A. and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks* 18:602–610.

- Guindon, S., J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of phyml 3.0. *Syst Biol* 59:307–321.
- Harmon, L. J., J. B. Losos, T. J. Davies, R. G. Gillespie, J. L. Gittleman, W. B. Jennings, K. H. Kozak, M. A. Mcpeek, F. Moreno-Roark, T. J. Near, A. Purvis, R. E. Ricklefs, D. Schluter, J. A. S. Ii, O. Seehausen, B. L. Sidlauskas, O. Torres-Carvajal, J. T. Weir, and A. Ø. Mooers. 2010. Early bursts of body size and shape evolution are rare in comparative data. *Evolution* 64:2385–2396.
- Hasegawa, M., H. Kishino, and T.-a. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of Molecular Evolution* 22:160–174.
- Hauffe, T., M. M. Pires, T. B. Quental, T. Wilke, and D. Silvestro. 2022. A quantitative framework to infer the effect of traits, diversity and environment on dispersal and extinction rates from fossils. *Methods in Ecology and Evolution* 13:1201–1213.
- Heaps, S. E., T. M. Nye, R. J. Boys, T. A. Williams, S. Cherlin, and T. M. Embley. 2020. Generalizing rate heterogeneity across sites in statistical phylogenetics. *Statistical Modelling* 20:410–436.
- Heath, T. A., J. P. Hulslenbeck, and T. Stadler. 2014. The fossilized birth-death process for coherent calibration of divergence-time estimates. *Proc Natl Acad Sci USA* 111:2957–2966.
- Heled, J. and A. J. Drummond. 2009. Bayesian inference of species trees from multilocus data. *Molecular biology and evolution* 27:570–580.
- Hillis, D. M., J. J. Bull, M. E. White, M. R. Badgett, and I. J. Molineux. 1992. Experimental phylogenetics: generation of a known phylogeny. *Science* 255:589–592.
- Hochreiter, S. and J. Schmidhuber. 1997. Long short-term memory. *Neural computation* 9:1735–1780.
- Höhna, S., M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist. 2016. Revbayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic biology* 65:726–736.

- Höhna, S., M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Syst Biol* 65:726–736.
- Hubisz, M. J., K. S. Pollard, and A. Siepel. 2010. PHAST and RPHAST: phylogenetic analysis with space/time models. *Briefings in Bioinformatics* 12:41–51.
- Huelsenbeck, J. P. and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755.
- Jia, F., N. Lo, and S. Y. W. Ho. 2014. The impact of modelling rate heterogeneity among sites on phylogenetic estimates of intraspecific evolutionary rates and timescales. *PLoS ONE* 9:e95722.
- Jiang, Y., M. Balaban, Q. Zhu, and S. Mirarab. 2022. Depp: Deep learning enables extending species trees using single genes. *bioRxiv* Page 2021.01.22.427808.
- Jukes, T. H. and C. R. Cantor. 1969. *Mammalian Protein Metabolism. Part IV: Protein Metabolism during Evolution and Development of Mammals* Academic Press, New York.
- Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. 2021. Highly accurate protein structure prediction with alphafold. *Nature* 596:583–589.
- Koch, B., E. Denton, A. Hanna, and J. G. Foster. 2021. Reduced, reused and recycled: The life of a dataset in machine learning research. *in* *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (J. Vanschoren and S. Yeung, eds.) vol. 1 Curran.
- Kulikov, N., F. Derakhshandeh, and C. Mayer. 2023. Machine learning can be as good as maximum likelihood when reconstructing phylogenetic trees and determining the best evolutionary model on four taxon alignments. *bioRxiv* .
- Lambert, S., J. Voznica, and H. Morlon. 2023. Deep learning from phylogenies for diversification analyses. *Systematic Biology* Page syad044.
- Landis, M. J., N. J. Matzke, B. R. Moore, and J. P. Huelsenbeck. 2013. Bayesian analysis of biogeography when the number of areas is large. *Syst Biol* 62:789–804.

- Lemey, P., M. Salemi, and A. Vandamme. 2009. *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. Cambridge University Press.
- Lemoine, F., J.-B. Domelevo Entfellner, E. Wilkinson, D. Correia, M. Dávila Felipe, T. De Oliveira, and O. Gascuel. 2018. Renewing felsenstein's phylogenetic bootstrap in the era of big data. *Nature* 556:452–456.
- Lenski, R. E. 2017. Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. *The ISME Journal* 11:2181–2194.
- Louca, S. and M. W. Pennell. 2020. Extant timetrees are consistent with a myriad of diversification histories. *Nature* 580:502–505.
- Maddison, W. P. and R. G. FitzJohn. 2015. The unsolved challenge to phylogenetic correlation tests for categorical characters. *Syst Biol* 64:127–136.
- Marcionetti, A. and N. Salamin. 2023. Insights into the genomics of clownfish adaptive radiation: the genomic substrate of the diversification. *Genome Biol Evol* .
- Marcus, G. F. 2018. Deep learning: A critical appraisal. *ArXiv abs/1801.00631*.
- Mayrose, I., A. Mitchell, and T. Pupko. 2005. Site-specific evolutionary rate inference: Taking phylogenetic uncertainty into account. *Journal of Molecular Evolution* 60:345–353.
- Meyer, X., L. Dib, D. Silvestro, and N. Salamin. 2019. Simultaneous bayesian inference of phylogeny and molecular coevolution. *Proc Natl Acad Sci USA* 116:5027–5036.
- Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. v. Haeseler, and R. Lanfear. 2020. Iq-tree 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution* 37:1530–1534.
- Nee, S., R. M. May, and P. H. Harvey. 1994. The reconstructed evolutionary process. *Phil Trans R Soc B* 344:305–311.
- Nesterenko, L., B. Boussau, and L. Jacob. 2022. Phyloformer: towards fast and accurate phylogeny estimation with self-attention networks. *bioRxiv Pages* 2022–06.

- Nielsen, R. and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the hiv-1 envelope gene. *Genetics* 148:929–936.
- Nute, M., E. Saleh, and T. Warnow. 2019. Evaluating statistical multiple sequence alignment in comparison to other alignment methods on protein data sets. *Systematic Biology* 68:396–411.
- Polson, N. G. and V. Sokolov. 2017. Deep Learning: A Bayesian Perspective. *Bayesian Analysis* 12:1275 – 1304.
- Price, M. N., P. S. Dehal, and A. P. Arkin. 2010. Fasttree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490.
- Rabosky, D. 2006. Likelihood methods for detecting temporal shifts in diversification rates. *Evolution* 60:1152–1164.
- Ramachandran, P., B. Zoph, and Q. V. Le. 2017. Swish: a self-gated activation function. arXiv preprint arXiv:1710.05941 7:5.
- Rambaut, A. and N. C. Grass. 1997. Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Bioinformatics* 13:235–238.
- Ree, R. H. and I. Sanmartín. 2018. Conceptual and statistical problems with the dec+ j model of founder-event speciation and its comparison with dec via model selection. *Journal of Biogeography* 45:741–749.
- Robinson, D. F. and L. R. Foulds. 1981. Comparison of phylogenetic trees. *Mathematical biosciences* 53:131–147.
- Ronquist, F., M. Teslenko, P. Van Der Mark, D. L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M. A. Suchard, and J. Huelsenbeck. 2012. Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* 61:539–542.
- Salamin, N., T. R. Hodkinson, and V. Savolainen. 2005. Towards building the tree of life: A simulation study for all angiosperm genera. *Systematic Biology* 54:183–196.
- Sapoval, N., A. Aghazadeh, M. G. Nute, D. A. Antunes, A. Balaji, R. Baraniuk, C. J. Barberan, R. Dannenfelser, C. Dun, M. Edrisi, R. A. L. Elworth, B. Kille, A. Kyrrilidis, L. Nakhleh,

- C. R. Wolfe, Z. Yan, V. Yao, and T. J. Treangen. 2022. Current progress and open challenges for applying deep learning across the biosciences. *Nature Communications* 13:1728.
- Schliep, K. P. 2011. phangorn: phylogenetic analysis in r. *Bioinformatics* 27:592–593.
- Siepel, A., G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* 15:1034–1050.
- Silvestro, D. and T. Andermann. 2020. Prior choice affects ability of Bayesian neural networks to identify unknowns. *arXiv Page arXiv:2005.04987*.
- Silvestro, D., S. Castiglione, A. Mondanaro, C. Serio, M. Melchionna, P. Piras, M. Di Febbraro, F. Carotenuto, L. Rook, and P. Raia. 2020. A 450 million years long latitudinal gradient in age-dependent extinction. *Ecology letters* 23:439–446.
- Silvestro, D., N. Salamin, A. Antonelli, and X. Meyer. 2019. Improved estimation of macroevolutionary rates from fossil data using a Bayesian framework. *Paleobiology* 45:546–570.
- Silvestro, D., R. C. M. Warnock, A. Gavryushkina, and T. Stadler. 2018. Closing the gap between palaeontological and neontological speciation and extinction rate estimates. *Nature Comm* 9:1–14.
- Smith, M. L. and M. W. Hahn. 2023. Phylogenetic inference using generative adversarial networks. *Bioinformatics* 39:btad543.
- Soubrier, J., M. Steel, M. S. Lee, C. D. Sarkissian, S. Guindon, S. Y. Ho, and A. Cooper. 2012. The influence of rate heterogeneity among sites on the time dependence of molecular rates. *Mol Biol Evol* 29:3345–3358.
- Stadler, T. 2011. Mammalian phylogeny reveals recent diversification rate shifts. *Proc Natl Acad Sci USA* 108:6187–6192.
- Stamatakis, A. 2014. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.

- Sukumaran, J. and M. T. Holder. 2010. Dendropy: a python library for phylogenetic computing. *Bioinformatics* 26:1569–1571.
- Sullivan, J. and D. L. Swofford. 1997. Are guinea pigs rodents? the importance of adequate models in molecular phylogenetics. *Journal of Mammalian Evolution* 4:77–86.
- Suvorov, A., J. Hochuli, and D. R. Schrider. 2019. Accurate inference of tree topologies from multiple sequence alignments using deep learning. *Systematic Biology* 69:221–233.
- Szandała, T. 2021. Review and Comparison of Commonly Used Activation Functions for Deep Neural Networks Pages 203–224. Springer Singapore, Singapore.
- Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on Mathematics in the Life Sciences* Page 57–86.
- Trost, J., J. Haag, D. Höhler, L. Nesterenko, L. Jacob, A. Stamatakis, and B. Boussau. 2023. Simulations of sequence evolution: how (un)realistic they really are and why. *bioRxiv* .
- Wu, C.-H., M. A. Suchard, and A. J. Drummond. 2013. Bayesian selection of nucleotide substitution models and their site assignments. *Molecular Biology and Evolution* 30:669–688.
- Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol Biol Evol* 10:1396–1401.
- Yang, Z. 1994. Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution* 39:306–314.
- Yang, Z. and B. Rannala. 1997. Bayesian phylogenetic inference using dna sequences: a markov chain monte carlo method. *Molecular biology and evolution* 14:717–724.
- Zaheri, M., L. Dib, and N. Salamin. 2014. A generalized mechanistic codon model. *Molecular Biology and Evolution* 31:2528–2541.
- Zhu, X. 2005. Semi-supervised learning literature survey. Tech. rep. University of Wisconsin - Madison.
- Zou, Z., H. Zhang, Y. Guan, and J. Zhang. 2019. Deep residual neural networks resolve quartet molecular phylogenies. *Molecular Biology and Evolution* 37:1495–1507.

Table 1: Accuracy of site-specific rates estimated under different models calculated across a test set of 600 datasets generated under different modes of rate heterogeneity. The accuracy is quantified as mean squared error and as R^2 (in parenthesis), comparing the true rates with those predicted through two maximum likelihood models (gamma and free-rates) and through phyloRNN. phyloRNN rate estimates consistently outperform those from a gamma model, in most cases matching or slightly exceeding the accuracy of the free-rates model. DL substantially outperforms likelihood methods in simulations with autocorrelated rates or based on codon models.

Heterogeneity model	Gamma model	Free-rates	phyloRNN
All combined	0.833 (0.175)	0.705 (0.206)	0.614 (0.347)
Gamma	1.869 (0.409)	1.311 (0.598)	1.64 (0.446)
Bimodal	0.454 (0.143)	0.28 (0.139)	0.332 (0.063)
Spike-and-slab	0.538 (0.161)	0.32 (0.484)	0.517 (0.08)
Geometric Brownian	1.469 (0.337)	1.408 (0.387)	1.387 (0.268)
Codon	0.379 (0.341)	0.405 (0.316)	0.076 (0.942)
Gamma autocorrelated	1.578 (0.465)	1.613 (0.584)	0.755 (0.798)
Bimodal autocorrelated	0.329 (0.133)	0.25 (0.112)	0.178 (0.294)
Spike-and-slab autocorrelated	0.387 (0.032)	0.323 (0.03)	0.271 (0.109)
Geometric Brownian autocorrelated	0.276 (0.034)	0.27 (0.028)	0.113 (0.457)
Mixed	1.882 (0)	1.656 (0)	1.689 (0)

Table 2: Accuracy of total tree length estimated under different models calculated across a test set of 600 datasets generated under different modes of rate heterogeneity. The accuracy is quantified a mean absolute percentage error and as R^2 (in parentheses), comparing the true rates with those predicted through two maximum likelihood models (gamma and free-rates) and through phyloRNN. The accuracy of phyloRNN tree length estimates is generally similar to that obtained from maximum likelihood methods, notably outperforming the gamma model when the underlying data are simulated with gamma-distributed rate heterogeneity model. Maximum likelihood inference outperforms phyloRNN in simulations based on bimodal, codon, and geometric Brownian autocorrelated models.

Heterogeneity model	Gamma model	Free-rates	phyloRNN
All combined	0.106 (0.933)	0.075 (0.955)	0.152 (0.902)
Gamma	0.147 (0.977)	0.083 (0.987)	0.158 (0.902)
Bimodal	0.105 (0.936)	0.047 (0.999)	0.115 (0.953)
Spike-and-slab	0.066 (0.993)	0.055 (0.998)	0.117 (0.986)
Geometric Brownian	0.118 (0.955)	0.102 (0.929)	0.150 (0.934)
Codon	0.056 (0.999)	0.051 (0.999)	0.103 (0.971)
Gamma autocorrelated	0.156 (0.952)	0.085 (0.921)	0.173 (0.791)
Bimodal autocorrelated	0.091 (0.925)	0.057 (0.996)	0.161 (0.957)
Spike-and-slab autocorrelated	0.057 (0.995)	0.053 (0.976)	0.103 (0.971)
Geometric Brownian autocorrelated	0.055 (0.999)	0.052 (0.999)	0.177 (0.98)
Mixed	0.471 (0.605)	0.459 (0.681)	0.493 (0.60)

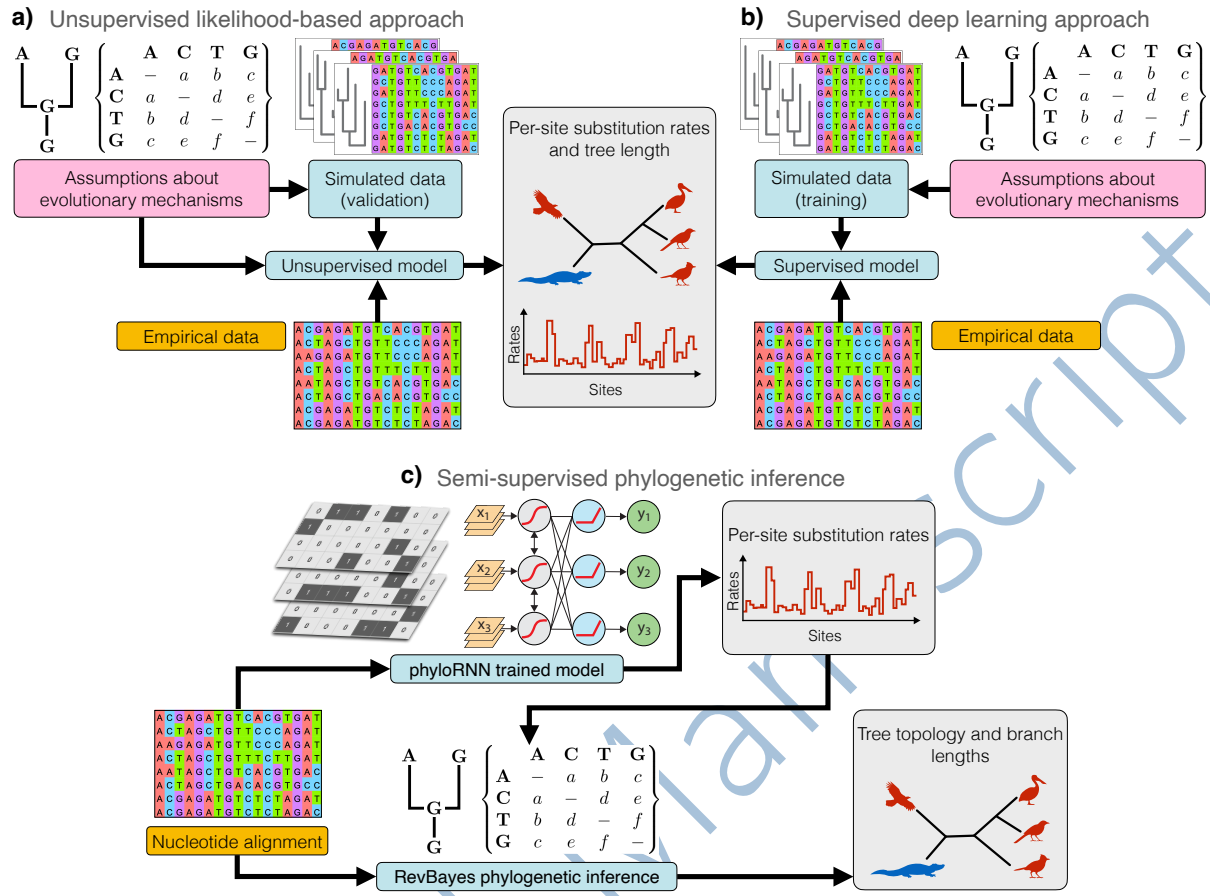


Figure 1

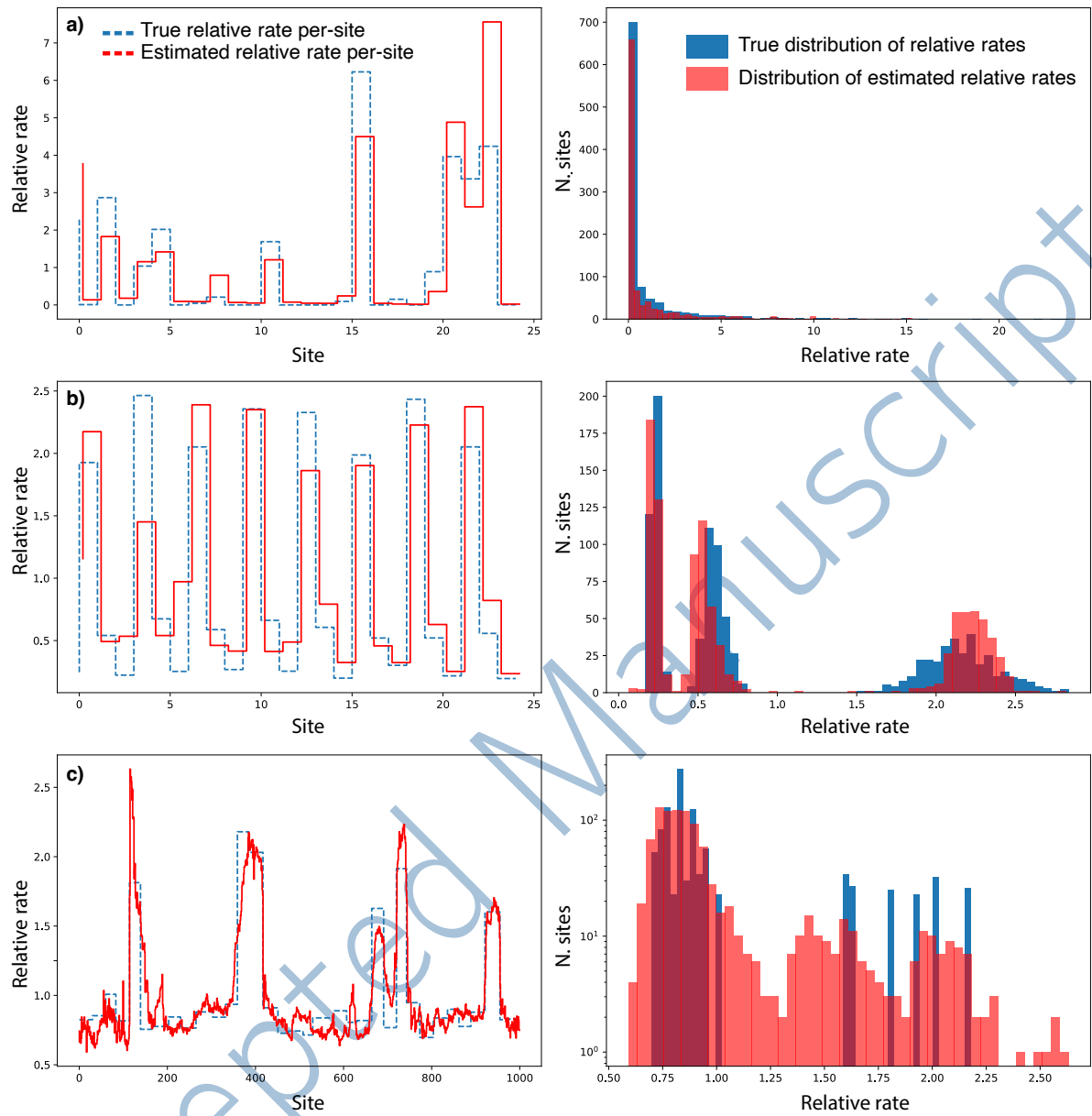


Figure 2

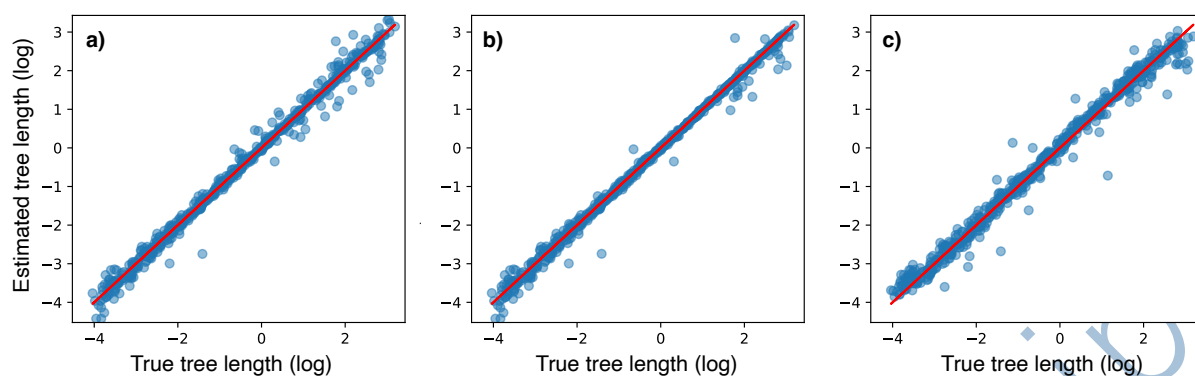


Figure 3

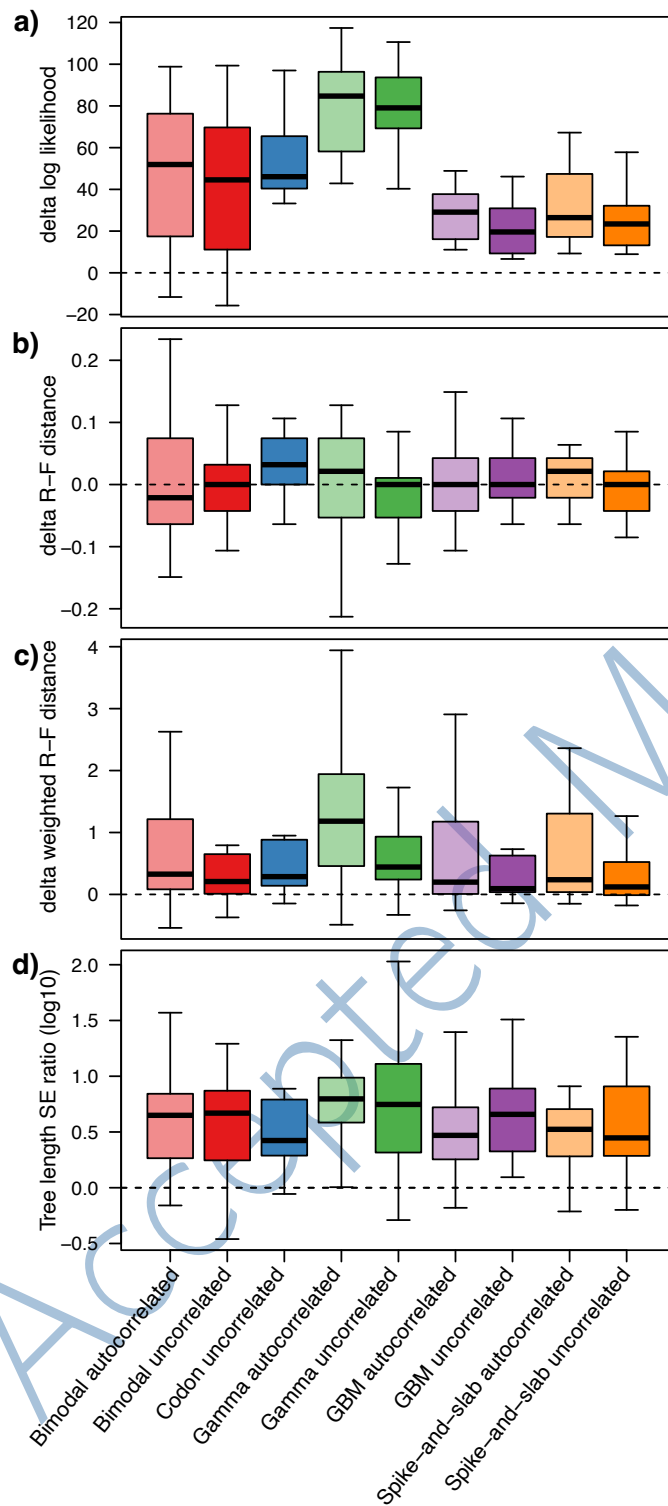


Figure 4

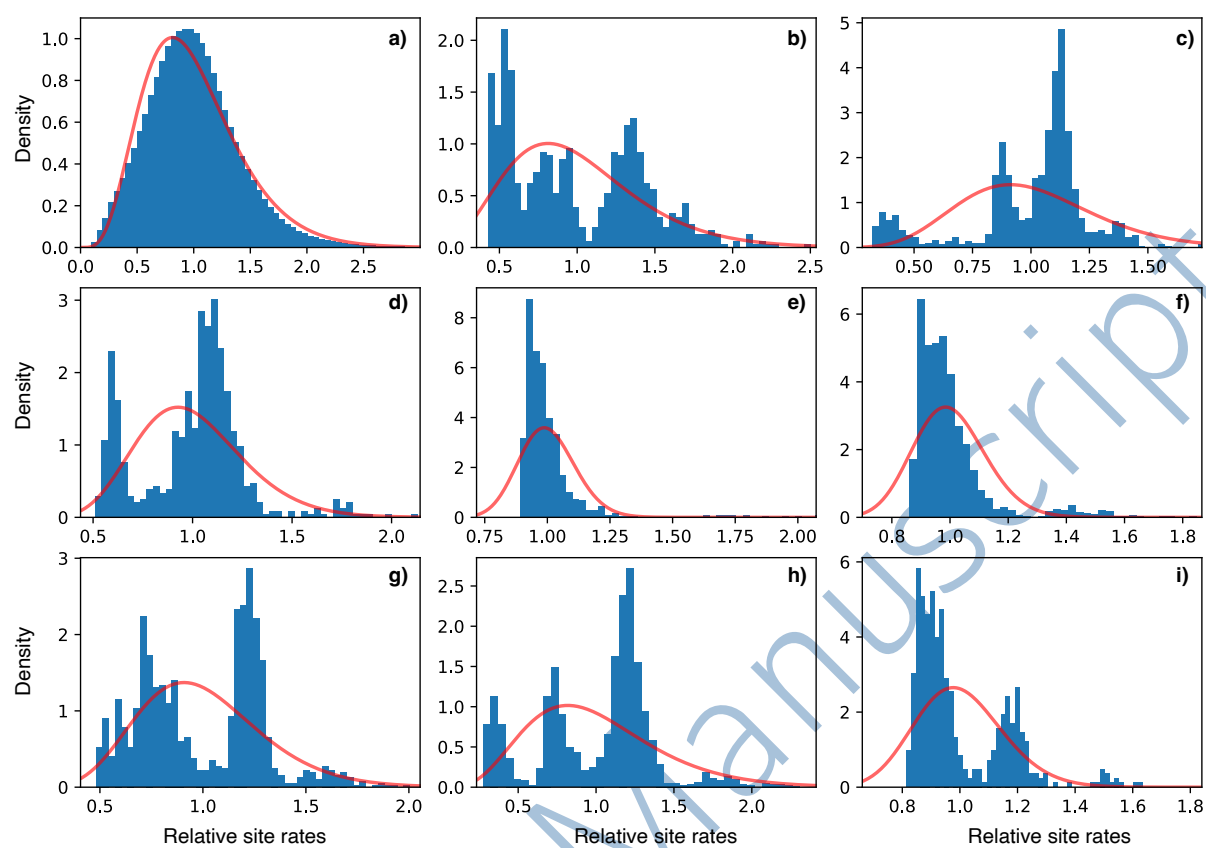


Figure 5

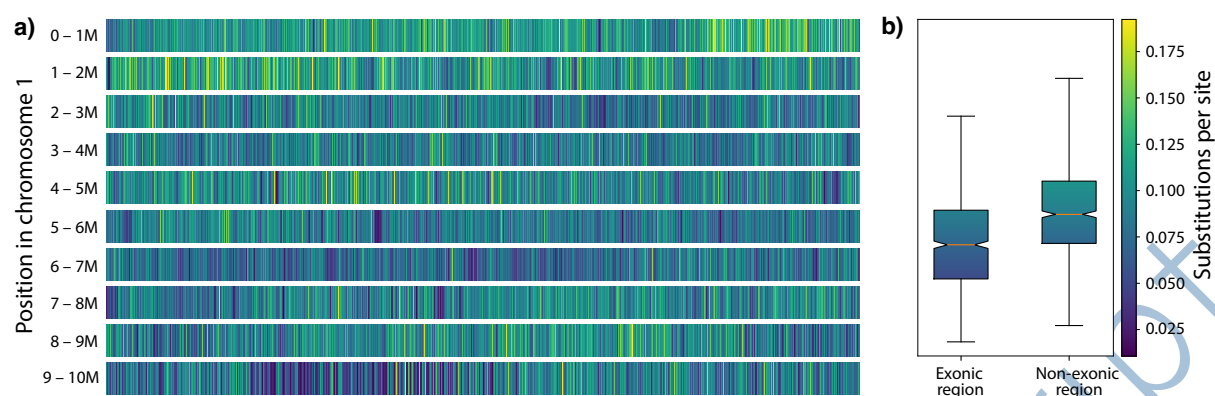


Figure 6